OPERATIONS RESEARCH CENTER

working paper

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

ADAPTIVE EVALUATION METHODOLOGY

PROTOTYPES: EXAMPLES

by

Alan Minkoff

OR 111-81 December 1981

ABSTRACT

Flexibility or adaptivity in public program evaluation can lead to large savings in time and money, with little or no loss in accuracy, if used properly. In this paper, guidelines are suggested for the employment of classical statistics in adaptive evaluation methodology. Through the case setting of a flu clinic, candidate techniques are demonstrated for handling problems in hypothesis testing, estimation, adaptive allocation of information-gathering resources, and before-and-after-type comparisons. In some cases, classical statistics proves quite adaptable to the requirements of the situation, while in others, its introduction is more artificial.

Adaptive Evaluation Methodology Prototypes: Examples

I. Introduction

Adaptive evaluation methodologv is still in its formative stages. The techniques that currently compose its frame have been primarily adapted from other fields. While these techniques may be very successful on their home ground, the peculiarities of public program evaluation might render them ineffective, impractical, or inferior to other available methods. In this short paper, an hypothetical situation will be presented, and several attractive candidates for the new methodology that were nominated previously (see [4]) will be fitted to it. In this manner, we hope to make a start at forging usable adaptive techniques, and to diagnose their strengths and weaknesses.

II. Situation

The case setting for this study will be that of a public clinic provided by a hospital for some service; suppose it is a flu immunization program for the elderly. The clinic provides this service to the community at low or no cost. However, the program's throughput is not what had been forecast. State experts differ on the reasons for this decline. Some say it is poor clinic service, others a lack of awareness of the clinic's existence. Correspondingly, these experts offer different solutions. A team of evaluators is commissioned to investigate the roots of this problem, and to predict the effectiveness of possible solutions.

Some of the questions that may be worth asking in this situation are:

* What do former patients think of the clinic?

* How do those involved with the operation of the clinic regard it?

* How many people use the clinic per day?

* What proportion of the local population is aware of the clinic's existence?

* How do the answers to the last two questions change during and after a local promotional campaign for the clinic?

III. Approaches

Several years ago, we (the evaluators) would have proceeded at once with fixed techniques. That is, we would totally plan each investigation in advance and never waver from the plan during the course of the sampling. Now we endeavor to be adaptive; if we really think we know the answer before time is up, let us try to wrap things up early and move on. Towards this end, we select our initial arsenal from among the techniques mentioned in [4]: The Sequential Probability Ratio Test [SPRT], double sampling, and stratified sampling. For a detailed description of these techniques, see the above-mentioned paper and its references.

Let us review the nature of each technique briefly. The SPRT tests the cumulative body of data collected as each new datum comes in. The decision after each test is to terminate and decide in favor of one hypothesis, or to continue sampling. Double sampling is a looser term, as it can be applied to testing or to estimation. Only one such "terminate or continue" decision is made; it may be regarded as "less adaptive". Stratified sampling, not inherently a sequential technique, apportions sampling weight optimally over sampling groups under certain assumptions. We would like to make use of these ideas, if not the techniques themselves, in our handling of the flu clinic problems.

IV. Examples

Example 1. Patients' Opinion of the Clinic

One possible reason for low clinic usage rates might be that those who have received service were not happy with the treatment they received. We will set out to determine what proportion of those treated felt that service was satisfactory. Consider what one might call a "confirmatory investigation". In this example, clinic proponents suggest that clinic service should be of a quality that would have 80% of those treated feel that they were treated satisfactorily. An unacceptable proportion would be, say, 50%. The evaluation is set up to test the null hypothesis that 80% of those treated were satisfied versus the alternative hypothesis that 50% were satisfied.

SPRT: First let us use the SPRT in this situation. Essentially, we set bounds on the cumulative likelihood function so that surpassing a bound tells us to stop sampling and gives us our decision. Fixed bounds may be directly determined by selecting desired Type I and Type II error rates, α and β, respectively. Wald [5] gives us that bounds $A = \frac{\beta}{1-\alpha}$ and $B = \frac{1-\beta}{\alpha}$ will produce α and β as Type I and Type II error rates,approximately. If we work with the logarithm of the cumulative ratio, then the bounds are transformed to log A and log B.

After the bounds have been set, we may begin taking observations. Suppose an observation consists of having a patient fill out a questionnaire (after receiving service) which includes a question such as, "Overall, did you feel that the service you received at the clinic was satisfactory?" If this is patient i, then $x_i = 1$ if the patient answers "yes", $x_i = 0$ otherwise. After each observation, we calculate

$$\Sigma_{i=1}^{K} \log f(x_i;\theta_1) - \log f(x_i;\theta_0),$$

where $f(x_i;\theta_j)$ is the likelihood of observing x_i if the parameter

$\theta = \theta_j$. For our binomial case, this works out to:

$$\log \frac{\theta_1}{\theta_0} \left(\Sigma_{i=1}^{K} x_i\right) + \log \left(\frac{1-\theta_1}{1-\theta_0}\right)\left[k-\Sigma_{i=1}^{K} x_i\right],$$

where k is the number of patients interviewed thus far, θ_0=.8, and θ_1=.5. If the above expression falls below log A on the kth trial, then we accept the null hypothesis that 80% are satisfied. If it rises above log B, the 50% alternative would be accepted. In either case, we terminate the sampling but if neither holds true on the kth trial. we proceed to the k+1st observation as above.

How do the evaluators select α and β? We might simply opt for the traditional 5% significance levels but this tends to be rather naive. We would probablv do better by considering the relative costs involved: the cost of making a wrong decision one way or the other, the cost of each observation, and budget constraints on both this investigation and the overall evaluation. There are any number of ways to incorporate these considerations into our bounds. Let us try one of the possible tactics.

Suppose it is felt (by the evaluators, the administrators, or the institution initiating the evaluation) that it is somewhat more costly to infer incorrectly that only 50% are satisfied when the actual proportion is 80%, as opposed to stating that 80% are satisfied when only 50% are. For instance, the consequence of accepting the 50% hypothesis may be to throw more money into facilities and personnel training, whereas an 80% conclusion indicates a small promotional campaign be run, whose cost is about half that of the clinic renovation. More money is wasted (or misguided) concluding for 50% erroneously than for 80% erroneously. So let us arrange the bounds to give us a smaller chance of making the first type of error, a Type I error. This means selecting α and β such that α is lower. Guided by costs, we take α=5% and β=10%. The difference in error rates might be made more severe if relative cost of the two types of errors showed a greater imbalance.

The weak point in this method of boundary construction as it stands arises from the possibility that the sampling budget may run out before a boundary is crossed if precautions are not taken to prevent selected error rates from being over-ambitious. Ideally, we would make our choices of α and β also dependent on budget and time constraints, null and alternative hypotheses, and perhaps an hypothesized distribution for the true underlying parameter. Any attempt to interrelate these factors is likely to meet up with incorrigible mathematical expressions in even the simplest cases, though. We would be better off making do with gross approximations or a worst-case analysis, if possible. We might aim to approximate worst-case sample size for a given set of error rates and hypotheses, and see if this is in line with budget constraints. If it exceeds them, error rates can be magnified. A series of simulations might even be performed to gauge an "optimal" set of error rates, though the optimality would carry only as far as the approximations could. This type of procedure relies heavily on computer availability, which is very often a factor in determining how effective a sequential technique may be utilized.

In this example, suppose that at most 40 observations can be taken. With α=5% and β=10%, we ask, "Is this feasible?" Fortunately, we have an expression for the expected number of observations in the worst case for the parameter θ (from Wald [5]):

$$\frac{-\log \frac{\beta}{1-\alpha} \cdot \log \frac{1-\beta}{\alpha}}{\log ({}^{\theta}1/\theta_0) \cdot \log ({}^{1-\theta}0/1-\theta_1)}$$

For the numbers given, and θ_0=80%, θ_1=50%, this works out to slightly over fifteen. Thus, the 40 possible observations would be sufficient to cover the worst expected number of observations. Whether we feel safe enought that the actual number will not exceed 40 is more difficult to answer. If we are uncomfortable, we may select less stringent error rates.

If the budget should run out before a decision is reached, we must then exercise a decision rule that will judge the final outcome. Such a rule might be: if the final cumulative log-likelihood is positive, select the alternative hypothesis; if negative, take the null as true. Such a rule will alter projected error rates, as they will no longer be α and β. Again, actual evaluation may be complex, so we might settle for diminishing the chance that no boundary will be crossed, and keeping a simple "tiebreaker" rule handy. Truncated SPRT theory does exist, though. Another possibility would be tapered boundaries that meet at the budget limit, allowing no possibility for arbitrary tie-breakers. The mathematical machinery behind such techniques are beyond the scope of this paper, though we might hope to fashion some sort of aesthetic graphical technique.

To get a more concrete view of what this all means, let us produce some fictitious numbers. The series of 1's and 0's that follows indicates the result of a sample of 40 former patients, asked whether they thought the service they received at the clinic was satisfactory (1 if yes):

100111010101110111011111011011111110110110

In testing the hypotheses θ_0=.8 vs θ_1=.5 at significance levels α=5% and β=10%, we would use log bounds -2.25 and 2.89. A marker is shown above at the point where the test would terminate (marker #1). Here, for the first time, one of the boundaries is crossed. It is the lower boundary, for the cumulative log likelihood for 23 ones and 9 zeroes is -2.56. We would then rule in favor of the null hypothesis, that the clinic performed satisfactorily to 80% of the patients seen. If the budget only allowed for 25 sample points, no decision would have been reached. Since the likelihood ratio was negative at that point, we would probably decide in favor of the null hypothesis, but would not be as sure.

Graphically, we might try methods like those shown in Figure 1.

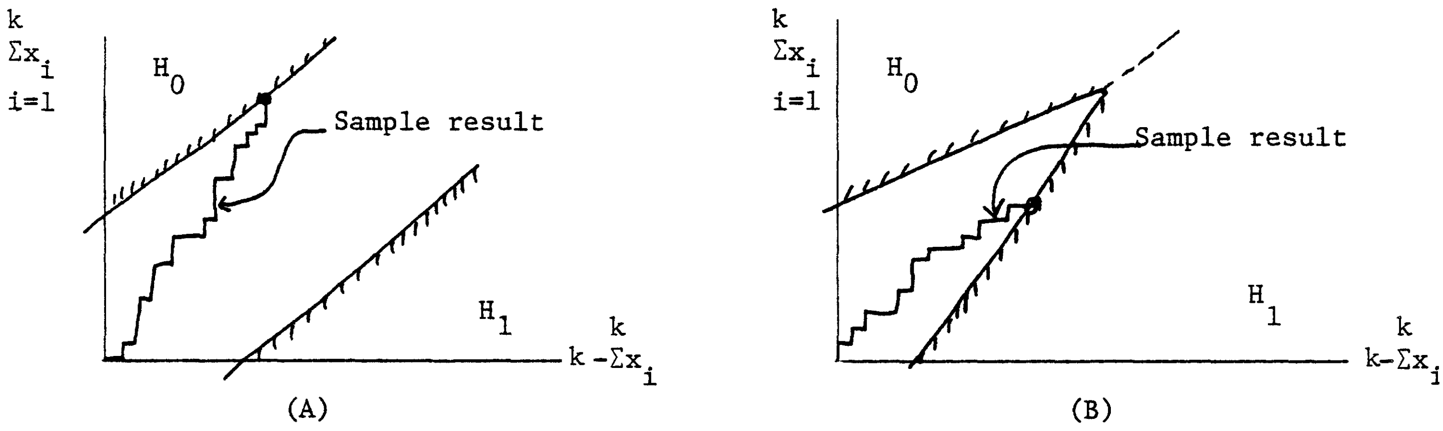


Figure 1: Possible Decision Regions for SPRT

The way these graphs would be utilized is: When a "yes" response comes in, draw a line segment from the last endpoint one unit vertically; when a "No" is received, draw it horizontally. When a boundary is crossed, decide according to which region the path ends in.

The graph in Figure 1A is the SPRT on graph paper. By the parallel boundaries we see that there exist response paths for which the sampling never ends. While these paths occur with probability zero, their existence implies the existence of response patterns with a significant chance of occurring that are associated with drawn-out experiments. The graph in Figure 1B avoids this possibility by tapering the boundaries inward so that one must be crossed by a certain point. This is a more desirable scheme, but it is harder to calculate error rates. One might start with a graph like that in Figure 1A and slowly taper the boundaries in after a certain point, until they meet at the budget constraint. Such a proposal, along with a hypothetical experiment, is shown in Figure 2.

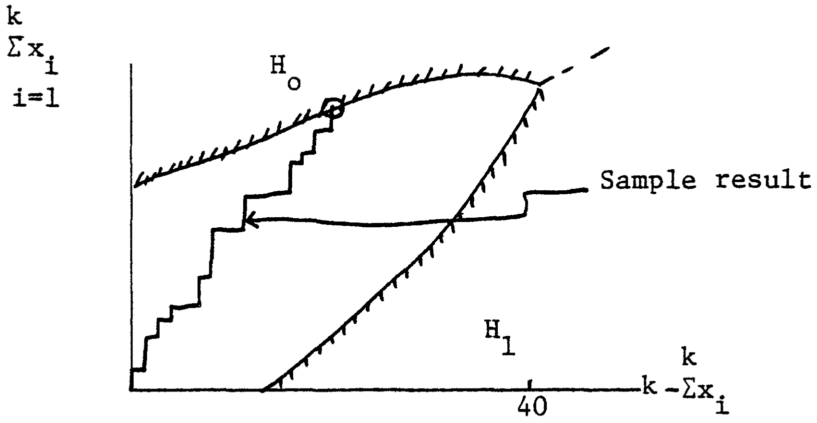


Figure 2: A Proposed New Set of "Tapered Decision" Regions

In the illustrative result of the test shown, the decision is to declare H_o true.

Double Sampling: Double sampling represents the other end of the sequential continuum. Only one sequential-type decision is made. This decision is like an SPRT decision, in that one must decide whether to terminate or to continue. However, if the decision is to continue, one must realize that the next sample to be taken is the final one. Also, one must decide how many points to sample in the second stage. This may be determined before the experiment, when the initial size is also set, or one might await the result of the first sample before fixing the decision, and use that information. It is evident that there are many ways to go about double sampling. The literature on double sampling is quite diffuse, making it difficult to select one technique as "best" for evaluation research. It may well turn out that different situations call for different approaches. Here, let us use a simple-minded technique to demonstrate the main ideas.

Before the experiment that generated the yes-no series above, we evaluators examine the alternative hypothesis. After n observations, we can reasonably say: if θ_0 were correct, the standard deviation for the average of n observations of x (the yes-no variable) would be $\sqrt{\frac{0.8x0.2}{n}}$; if θ_1 were correct, it would be $\sqrt{\frac{0.5x0.5}{n}}$. If this were a single-sample experiment, we might want to choose a value c such that if $\sum_{i=1}^{n} x_i > c$, we would accept H_0 and otherwise, accept H_1 . If c is positioned so that α = 5% and β = 10% then (assuming we can use the Ganssian distribution to approximate) we should have:

$$\frac{c-.8}{\sqrt{\frac{.8.2}{n}}} = \Phi^{-1}\ (.05)\simeq -1.645 \text{ and } \frac{c-.5}{\sqrt{\frac{.5.5}{n}}} = \Phi^{-1}(.90)\simeq 1.28,$$

where $\Phi^{-1}(x)$ is that point of the standard normal distribution at which the area to the left under the representative curve is x. We can solve these for n & c, yielding c =.65, n = 19. To implement a double sampling scheme, we divide n into a 3/5 part and a 2/5 part so that we have n_1 = 11 and n_2 = 8 (this has certain nice minimax properties; see Hald [3]). These shall be our first and second sample sizes, respectively.

Now we need stopping and decision criteria for the decision point after the first batch of observations. This means selecting C_0 and C_1 such that: if $\sum_{i=1}^{11} x_i > C_0$, we stop and accept H_0; if $\sum_{i=1}^{11} x_i < C_1$, we stop and accept H_i; and if neither holds, we sample eight more. To maintain α=5% and β=10%, we need

$$\frac{(C_0/11)-.5}{\sqrt{\frac{0.5x0.5}{11}}} = 1.28 \text{ or } C_\theta = .69x11=7.6$$

and

$$\frac{(C_1/11)-.8}{\sqrt{\frac{0.8x0.2}{11}}} = -1.645 \Rightarrow C_1 = .60x11 = 6.6$$

Thus, if after sampling eleven points, we find $\Sigma_{i=1}^{K} x_i > 7.6$ or that we have eight or more "yes" responses, stop and conclude H_0. If we find we have only six or fewer, we accept H_i. If we have seven, then we sample eight more, and our final criterion value is c x n = 0.65 x 19 = 12.35. We decide for H_0 if we have thirteen or more "yes" responses, for H_i otherwise.

Using the same series of responses as before, we see that the first eleven points produce six ones. This is a "terminate and conclude H_1" result. Had we sampled eight more, we would have come up with twelve "yes" responses, and should conclude H_1 again. Note the differences between this and the SPRT results. On the same set of data, we took fewer observations with double sampling, and made the opposite conclusion.

This difference can be attributed in part to the data pattern, as the later responses are predominantly "yes". Also, the two tests are very different in nature.

The second sample size in the double sample need not have been fixed before the experiment. Another option within the double sample is to use the information received in the first sample, primarily that relating to estimated variance. That is not of too much service here, for variance is directly related to the parameter θ and the sample size in the binomial distribution. Predicted variance would not change unless we changed an hypothesis. First sample size would be much more valuable when the variable under investigation is Gaussian, where hypotheses about the means have little bearing on the variances. Assuming that a prediction of variance was needed to determine the first sample size, an estimated variance not in line with this prediction should direct a better second sample size than can be decided beforehand. One other aside: we might switch to t-distributions if we feel that the data are not inherently Gaussian and that sample sizes are too small for Gaussian approximations. Computation with the t-values can be more difficult, though, because they often depend on n.

The foregoing example is also illustrative of the benefits of adaptivity. Our initial double sampling calculation told us that we should sample 19 points if we wanted to conduct a single-sample test. Yet in performing double sampling, we realized we could quit after 11 observations. This represents a savings of 8 x(cost per observation) in sampling costs. Adaptivity allowed us to devote some of the resources' which we might have spent on sampling under a fixed plan, to investigation of other questions. Or, we might content ourselves with a reduced cost per fixed amount of information collected. On the other hand, the SPRT employed here overruns the projected single sample size. This demonstrates some of the risk involved in using an open-ended test of its nature. We do not mean to imply that the SPRT is inferior to the double sample. We hope to undertake a project in the near future which will run simulations to compare sample sizes and resource savings under double sampling and various forms of the SPRT.

Example 2: Several Groups' Opinions of the Clinic

We now expand the previous example to the task of evaluating the performance of the clinic as seen by a variety of groups. Suppose the evaluators wish to assess the proportion of people satisfied with the clinic's performance simultaneously from the categories clinic doctor, clinic nurse, clinic administrator, past patient, and local resident (who has never used the clinic). We shall put ourselves under a budget constraint for this survey. Assume no "program" such as a promotional campaign is in action. Let us see how stratified sampling may be used to perform this evaluation adaptively.

Stratified Sampling: The guiding principle to stratified sampling is to sample within each stratum, or group, a number that is directly proportional to the size of the stratum and to its true standard deviation, and inversely proportional to the square root of the cost per observation of sample points in that group. Symbolically,

$$n_h \propto \frac{N_h\ S_h}{\sqrt{C_h}} \quad .$$

The n_h's we actually use will depend either on the total number to be sampled, or the total budget alloted. This rule is designed to measure some parameter in aggregate by looking at homogeneous groups. It is optimal in the sense that it minimizes the variance of the overall estimate for a given total observation cost, or minimizes the cost for given variance. This is not strictly our situation, in that some of these groups may be quite heterogeneous, and we might be interested in the individual group estimates. Nevertheless, if it is taken as an heuristic, rather than as an optimal rule, we might find that it is a sensible approach.

Let us consider some specific numbers. Denoting by d, n, a, p, and l the categories doctor, nurse, administrator, patient, and local resident, respectively, take these as values for population size and cost per observation:

$N_d = 5$	$C_d = 5$
$N_n = 15$	$C_n = 1$
$N_a = 10$	$C_a = 3$
$N_p = 300$	$C_p = 3$
$N_l = 20{,}000$	$C_l = 2$

The missing element here is the standard deviation.

There are two ways to handle the problem of not knowing standard deviations. The first sampling period may be devoted to estimating these quantities, perhaps by sampling an equal number of points from each group, or expert opinions may be solicited to provide working guesses. In cases such as this one (binomial), an opinion on the proportion satisfied is itself an opinion regarding the standard deviation (see Example 1).

We shall take the first option. It is decided to sample three doctors, five nurses, three administrators, and ten each of former patients and local residents during the first sampling period. Where do we come up with these numbers? Let us say that relative costs and numbers per group, plus the objective of acquiring standard deviation estimates, were the main factors. Using the same response code as in Example 1, the results of this first sample are:

Category	Result	Mean	S_{h_1} (Standard deviation)	
d	111	1.00	.00	
n	10010	0.40	.22	Cost: 79
a	010	0.33	.27	
p	1001110101	0.60	.15	
l	1111011011	0.80	.13	

We immediately detect that the results from the interview of doctors show perfect homogeneity: all are satisfied. This is an estimate of zero for the standard deviation, which sounds fairly certain. Yet we should have no reason to believe that the other two doctors are satisfied. So, for the sake of completeness of information, we decide to include the other two doctors in the next sample.

We see another problem brewing. Sampling proportionately to <u>size</u> of group would prompt a groundswell of local resident interviews, while few of the people in the other groups would be surveyed. For this reason, we ought to consider scrapping pure sample size as a criterion and favor a measure to describe how important it is to know each group's opinion. We probably do not care too much what the local residents think, since they have no first-hand experience with the clinic. We feel it is mildly desireable to be aware of their impressions. The logical fix-up would seem to replace "number in group" with "information weight of group", and substitute this weight W_h for

N_h.

This seems all well and good. We can define, for instance, 10,000 local residents to be one sampling unit, comparable to five former patients or one doctor. We must caution ourselves that if we change the sampling unit, so too must we change the sampling cost per unit. Given the weight, W_h, and the number in group h, N_h, we must also calculate the number of sampling points per unit weight, $UN_h = N_h/W_h$, and the cost per unit weight, $UC_n=C_h \times UN_h$. With these figures in hand, we may go about allocating the sampling budget along stratified sampling guidelines.

Let us continue our example of the clinic. For the second sampling period, we decide to sample the two remaining doctors, and spread $80 worth of sampling among the four other groups. Through casual interviews with people who want and people who know, we weight the four remaining groups as shown in column W_h, according to how valuable the opinion of each group is, with former patients the standard at 100:

Group	W_h	UN_h	UC_h	$RSSW_h$	$PSSW_h$
n	100	0.15	0.15	56.8	.74
d	40	0.25	0.75	14.4	.19
p	100	3	9	5.0	.07
1	20	100	200	0.18	.0024

UN_h and UC_h are also listed above. In the second to last column are the raw stratified sampling weights determined as:

$$RSSW_h = \frac{W_h \times S_{h_1}}{\sqrt{UC_h}} .$$

The final column shows the proportion each $RSSW_h$ forms of the total. We may now proceed to the next step.

Suppose we use our $80 to sample the value SW worth of sampling weight.

Then we will sample .74 SW nurse weight units, .19 SW administrator weight units, etc. This leads to the cost equation:

$$[(.74SW) \times .15] + [(.19SW) \times .75] + [(0.75SW) \times 9] + [(.0024\ SW) \times 200] = 80,$$

which we can solve for SW. Here, SW turns out to be around 59. We apply this to the $PSSW_h$ column, multiplying the two to determine total weight sampled from each group, and convert this to number of people:

Group	W_h	n_h	$n_h \times C_h$
n	43.67	6.55→7	7
d	11.21	2.80→3	9
p	4.13	12.39→12	36
l	0.14	14	28

We will spend $80, which is exactly what we wished to.

After this sample, we may compute new standard deviations $\{S_{h_2}\}$ based on the current knowledge profile and apply this technique again. However, there will be few unsampled people left in the nurse and administrator categories, so we might alter our plan. We may feel certain enough with what we already know about these two groups, or we might sample the remaining personnel in these groups and only apply the technique to the patient and resident categories. This depends on what our goals are, how much time we have, and how large our overall sampling budget is.

Example 3: Number of People Who Use the Clinic Per Day

If we consider the arrivals of prospective patients at the clinic to be a Poisson process (verified perhaps through a goodness-of-fit test), then the parameter of interest in an investigation would likely be the rate parameter λ, the mean number of arrivals per day. Suppose that the evaluators decide to employ this criterion as a measure of clinic use. Suppose, also, that the government will soon be starting a promotional campaign that is intended to make the community better aware of the existence of the clinic and its capabilities. The evaluators want to measure the change in λ after the promotional campaign has started.

To measure a change, we must have a measurement of the state of things as they were, and an idea of what has happened since. Toward this goal, we initially set up our evaluation under a two-stage plan. First, we take samples of clinic usage during the baseline period, which means prior to the promotional campaign; then, after the campaign has been run for some length of time, clinic usage will again be sampled, to establish a mean arrival rate in the experimental period.

Above all, we want to be adaptive; we only want to sample as much as we need to become reasonably sure of the underlying baseline rate, and then move on to the experiment. We believe this will prove to be more cost-efficient than if fixed sample sizes are arranged beforehand. Here, resources might be diverted to the experimental phase if it is felt this would acquire more valuable information.

If we take the long-range goal of state government to be that of bringing clinic usage up to projected levels, we must ascertain that this is not the case already. We could go to the clinic files and calculate an estimate of λ_b from records of the last couple of weeks, but that would not be instructive. So, let us propose that the records had been kept in a nearby town which has just been ravaged by earthquake, flood and famine. There is now no way of saying anything about λ_b without someone staying at the clinic and counting patient arrivals. We do not want to do this for very long, either.

Being adaptive folks, we decide to set up a SPRT pitting the null hypothesis $\lambda_b = 4$ patients/day, which represents a significant below-projected level, against the alternative hypothesis $\lambda_b = 9$ patients/day, which had originally been projected. The costs of wrong decision are somewhat opposite to what they were in Example 1, for we are looking at the other possible root of the problem. For variety's sake, let us set $\alpha = 5\%$ and $\beta = 2\%$. These give us bounds on the log likelihood ratio of -3.86 and 2.98. The log likelihood ratio after k trials can be expressed as

$$\log \left[\left(\frac{\lambda_1^{\Sigma_K x_i} \quad e^{-k\lambda_1}}{x_1! \ldots \, x_k!} \right)\right] / \left(\frac{\lambda_0^{\Sigma_K x_i} \quad e^{-k\lambda_0}}{x_1! \ldots \, x_k!} \right)]$$

$$= \sum_K x_i \log (\lambda_1/\lambda_0) - k(\lambda_0 - \lambda_1).$$

If we had observed for ten days, we would have observed the following number of patients each day:

2 10 3 4 5 3 7 5 4 8

Using the SPRT, we would have concluded our investigation in four days, where the log ratio is $-4.59 < -3.86$. We accept that λ_b is currently at the rate below the projected, desirable rate, at 4 rather than 9 patients/day.

Now we would like to run the promotional campaign and, perhaps after two weeks, resume looking at patient arrivals per day to measure λ_e. Before we do this, we ought to backtrack a little. So far, we have come away from the baseline period with the educated opinion that $\lambda_b = 4$ rather than 9. If we wanted to eventually construct a confidence interval for the difference between λ_b and λ_e (maybe as a measure of the campaign's effect), we would need an estimate of λ_b rather than a hypothesis about it, and an associated standard error of the estimate. If we merely wanted to state whether the promotional campaign brings about a significantly greater rate of patient visits, we also ought to work with an estimate of λ_b. While the previous SPRT confirmed our suspicions about usage rates, a 90% confidence interval for λ_b based on the four days observed would stretch from 0.52 to 8.98! We definitely need more observation time at the baseline stage.

Let us then organize our investigation into three stages. The first stage shall be a "needs assessment" determination; i.e., we check whether we need to introduce a change or not. If not, we might terminate this portion of the evaluation. The second stage will be devoted to estimating the baseline value for the parameter(s) of interest. We may use data collected during the first stage, if we consider it still valid, but we probably need more data than it alone provides us. The third stage consists of initiating the experiment and estimating or testing the parameter's value during this period. At the end of this stage, we draw our conclusions and may make recommendations.

Such a game plan as described above does not apply uniquely to adaptive evaluation. To make it adaptive, we want flexible time boundaries between stages. This involves the generation of a set of criteria for switching stages, and for drawing intra-stage and inter-stage conclusions. There could be many ways to do this. In keeping with the spirit of this paper, we will look at a couple of strategies for dealing with our flu clinic example. The two strategies differ primarily in what they do at the third stage. One prepares a confidence interval for the difference between baseline and experimental parameter values; the other tests for whether the experiment is an improvement (higher parameter value) over the normal state of affairs. This difference implies different goals during the second stage, too, so we will treat each strategy separately (assuming we have completed stage one and are continuing).

<u>Estimate/Estimate</u>: The objective is to measure a difference in underlying parameters. The stated form of this difference will be a point estimate and associated confidence interval (CI). Were it possible to perform baseline and and experiment in parallel, we could use one of these formulas for determining the confidence interval with confidence level (CL) of 1-α:

If we can assume $\sigma_b^2 = \sigma_e^2 = \sigma^2$, then from [2] we have

$$(1)\ CI = (\overline{x}_b - \overline{x}_e) \pm t_{n_b + n_e - 2;\ 1 - \alpha/2}\ Sw\sqrt{\frac{1}{n_b} + \frac{1}{n_e}}$$

for n_b points sampled from baseline and n_e from experiments with associated sample means $\overline{x}_b$ and $\overline{x}_e$, s_w^2 the pooled estimate of σ^2, and $t_{n_b + n_e - 2;\ 1 - \alpha/2}$

the 1-α/2 percent point of a t-distribution with $n_b + n_e - 2$ degrees of freedom. Otherwise, we must resort to

$$(2) \quad CI = (\bar{x}_b - \bar{x}_e) \pm t_{m;\ 1-\alpha/2}\sqrt{\frac{s_b^2}{n_b} + \frac{s_e^2}{n_e}}$$

where m, the modified degrees of freedom, is found by

$$(3) \qquad \frac{1}{m} = \frac{C^2}{n_b-1} + \frac{1-C^2}{n_e -1} \quad ,$$

C comes from

$$(4) \qquad C = \frac{S_b^2/n_b}{S_b^2/_{n_b} + S_e^2/_{n_e}} \quad ,$$

and S_b^2 and S_e^2 are the estimated variances by group.

Let the task be to shrink the width of the CI within a certain criterion at some pre-specified CL. The rationale for this might be, we make out pretty well if we are off by no more than a certain amount in our estimate. Also, if we can produce a guess at the average loss suffered through being outside our CI, as opposed to inside it, we can better select a CL. We might take a "Tarzan" approach to this problem, swinging from assumption to assumption grabbing as much data as we can hold along the way. We shall estimate by double sampling, that being the only single-group estimation method we are looking at in this paper.

The following proposed scheme illustrates one way of approaching the defined task. The operation sequence can get convoluted at certain points, so we advise the reader to follow the action on the flowchart depicted in Figure 3. The main points to keep in mind are that:

*We would like to use Equation (1) to transform proposed sample sizes into projected interval widths, and thereby select the lowest sample size that will do the job;

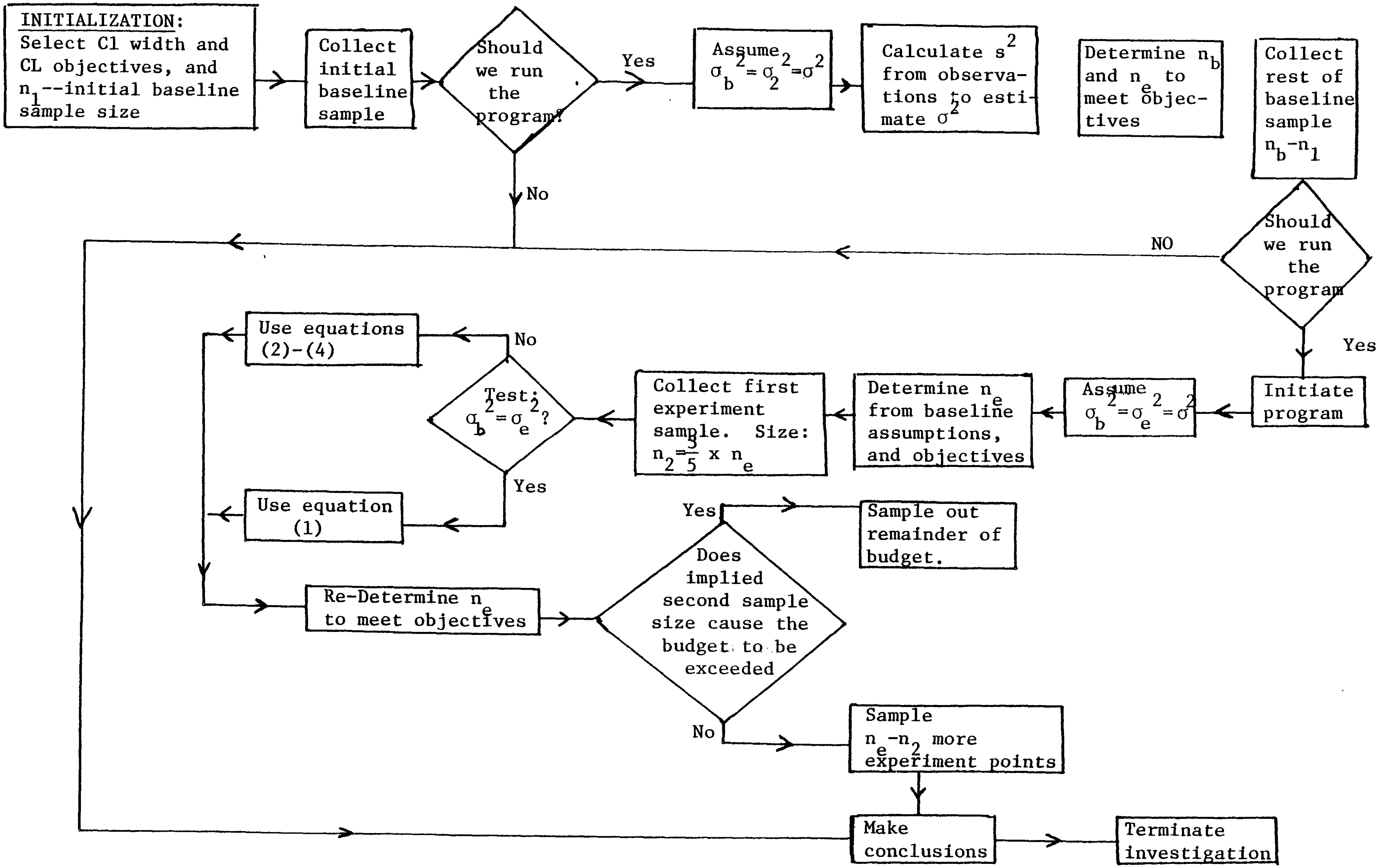

INITIALIZATION:
Select Cl width and CL objectives, and n_1--initial baseline sample size
Collect initial baseline sample
Should we run the program?
Yes
No
Assume $\sigma_b^2 = \sigma_2^2 = \sigma^2$
Calculate s^2 from observations to estimate σ^2
Determine n_b and n_e to meet objectives
Collect rest of baseline sample $n_b - n_1$
Should we run the program
NO
Yes
Initiate program
Assume $\sigma_b^2 = \sigma_e^2 = \sigma^2$
Determine n_e from baseline assumptions, and objectives
Collect first experiment sample. Size: $n_2 = \frac{3}{5} \times n_e$
Test: $\sigma_b^2 = \sigma_e^2$?
No
Yes
Use equations (2)-(4)
Use equation (1)
Re-Determine n_e to meet objectives
Does implied second sample size cause the budget to be exceeded
Yes
No
Sample out remainder of budget.
Sample $n_e - n_2$ more experiment points
Make conclusions
Terminate investigation

Figure 3: Flowchart for proposed Estimate/Estimate Procedure

*To use (1), we need to assume that the baseline and experiment period variances are the same, for if not, we have to employ the more complicated Equation (2)-(4).

*We check our assumptions statistically, using an F-test to test $\sigma_b^2 = \sigma_e^2$.

Assume $\sigma_b^2 = \sigma_e^2 = \sigma^2$ initially, and gather some sample large enough to estimate σ^2, but not too large. Substitute this estimate in for s_w^2 in (1), and attempt to maximize the CL, staying within the budget constraints and the interval criterion, by fiddling with n_b and n_e in (1). (Alternatively, one could set a CL and minimize the width of the CI.). Let the second sample size in the baseline segment be n_b- (first sample size), and finish baseline sampling. (We may, at this point, alter our first stab at n_e, based on the new data.) We enter the experimental period at this point, so initiate the experiment. Then take, say, (3/5) x n_e (prompted by double sampling) sample points for the first segment in the experimental period. Test whether $\sigma_b^2 = \sigma_e^2$ is valid, and if not, "swing" to equations (2) - (4). Given s_b^2, s_e^2 as it stands now, and n_b, maximize the CL (or attempt to hit the CL and CI) through appropriate choice of n_e. The second sample size is n_e- (first exp. sample size). Then look at s_b^2 and s_e^2, decide whether it is reasonable to assume $\sigma_b^2 = \sigma_e^2$, and use the appropriate expression for determination of the CI, swinging back to equation (1) if the coast is clear.

Let us "test drive" this technique with some numbers back at the flu clinic. We have just confirmed that λ_b is more likely 4 than 9, so we will run the promos. We can use the data from the first stage to estimate s_w^2, treating things as if they were Gaussian, or, initially, we can use our $\overline{x}_b$ estimate, as we are working with the Poisson distribution. This would give us $s_w = 2.18$. Suppose we want to be no more than one unit off in either direction in our estimate of the difference, and that we want to hit an 80% CL. The width of the CI is

$$2 \cdot t_{n_b + n_e - 2; .10} \quad s_w \sqrt{\frac{1}{n_b} + \frac{1}{n_e}}$$

which we want to be less than or equal to 2. Also assume our budget and/or time constraints limit us to 40 observations overall. If we shoot for equal sample sizes, to simplify their determination, we find that $n_b = n_e = 16$ about does what we want.

Our determination of $n_b = 16$ directs us to take 16 - 4 = 12 more observations during the baseline. Take them to be:

5 3 7 5 4 8 7 5 8 6 4 1

We conclude the baseline with $\bar{x}_b = 5.13$ and $s_b = 2.42$. Taking $s_w = S_b$ recalculating n_e to get a CI of length 2 and a CL of 80%, we discover that n_e should be around 26. This would have us exceed 40, so we use $n_e = 24$, and the first experimental sample comes out to (3/5) x 24 $\approx$ 14.

The first batch of the patients-per-day figures to come in during the experiment are:

14 5 9 9 11 8 3 5 12 7 8 9 6 6 .

Thus far, $S_e^2 = 8.92$. The F-test for equality of variances (two-sided) arises from testing the ratio $\frac{S_b^2}{S_e^2}$ against

$F_{n_b-1, n_e-1; 1-\alpha/2}$ and $F_{n_b-1, n_e-1; \alpha/2}$. $\frac{S_b^2}{S_e^2} = 0.657 > 0.487$

$= F_{15, 13; .90}$, so we can't reject the hypothesis that $\sigma_b^2 = \sigma_e^2$. This means we can pool s_b^2 and s_e^2 to get s_w^2. At this point, $s_w^2 =$ $\frac{(n_b-1)s_b^2 + (n_e-1)S_e^2}{n_b + n_e - 2} = 7.28$, $s_w = 2.70$. n_e ought to be even higher than was determined before, since the S_w we are using has increased. Since we were already straining our budget, we will simply collect the last 10 days worth of observations, reaching the end of the budget strings.

Our final batch of values looks like this:

$$8\ 13\ 3\ 8\ 7\ 7\ 9\ 5\ 7\ 9;\ \bar{x}_e = 7.83;\ S_e^2 = 7.80$$

S_e^2 is even closer to S_b^2, so we are going to again accept that $\sigma_b^2 = \sigma_e^2 = \sigma^2$. Our final estimate of σ^2 is $S_w^2 = 7.03$, which gives us a final Cl for $\lambda_e - \lambda_b$ of

$$(\bar{x}_e - \bar{x}_b) \pm t_{38,\ .10} \times S_w\sqrt{\frac{1}{n_b} + \frac{1}{n_e}} = (1.58,\ 3.82),$$

with a CL of 80%. The length of the interval is only slightly greater than two. With a slightly larger initial sample and looser budget, we might have been able to hit the objectives more readily.

Estimate/Test: Here we would like to double sample during the baseline to get a CI for λ_b. The most straightforward way to pre-select its width and CL follows the approach used in "Estimate/Estimate": establish a significant difference or maximum-tolerable error, double that for CI goal width, then select a CL in line with budget constraints to minimize expected loss. This is easy to say, probably difficult to put into practice, but we hope to illustrate the basic ideas in an example to follow.

After we conclude the baseline, we should have a CI plus the corresponding point estimate. Now, with the experiment running, we want to establish whether there has been some improvement in the state of affairs, or no change. Change must be measured against the baseline rate. Theoretically, the parameter's value could lie anywhere within the CI, although if we had to give one number, we would likely give the point estimate, which lies at the center of the CI. Let improvement indicate $\lambda_e > \lambda_b$ (we could follow similar reasoning for the case where the parameter should decrease under improvement). Denote the upper and lower limits of the CL by λ_h and λ_1, respectively. Let the CL be $1-\alpha$. Then we might say that $\lambda_h \geq \lambda_b$ with probability $1-\alpha$, because with probability 1-α, λ_b lies between λ_1 and λ_h. We can even add $\alpha/2$ to this probability for the result $\lambda_b<\lambda_1$ if we distribute the remaining α worth of probability uniformly over $\lambda<\lambda_1$ and $\lambda>\lambda_h$, bringing our confidence in $\lambda_h \geq \lambda_h$ to 1-α/2.

Now suppose we test H_o: $\lambda_e=\lambda_h+\delta$ vs $H_1:\lambda_e=\lambda_b$ during the experimental phase, where δ is an arbitrarily positive number. If we decide for the null hypothesis, then we conclude that λ_e is significantly greater than λ_b, we can make δ large enough so that a difference $\lambda_h+\delta-\lambda_e$ indicates significant improvement. If we reject the null hypothesis, we are left to conclude that $\lambda_e=\lambda_b$--no improvement. If we had set the test up to give us Type I and Type II error rates of γ and β, respectively, then we might employ a probability tree to give us overall predicted rates of these four outcomes:

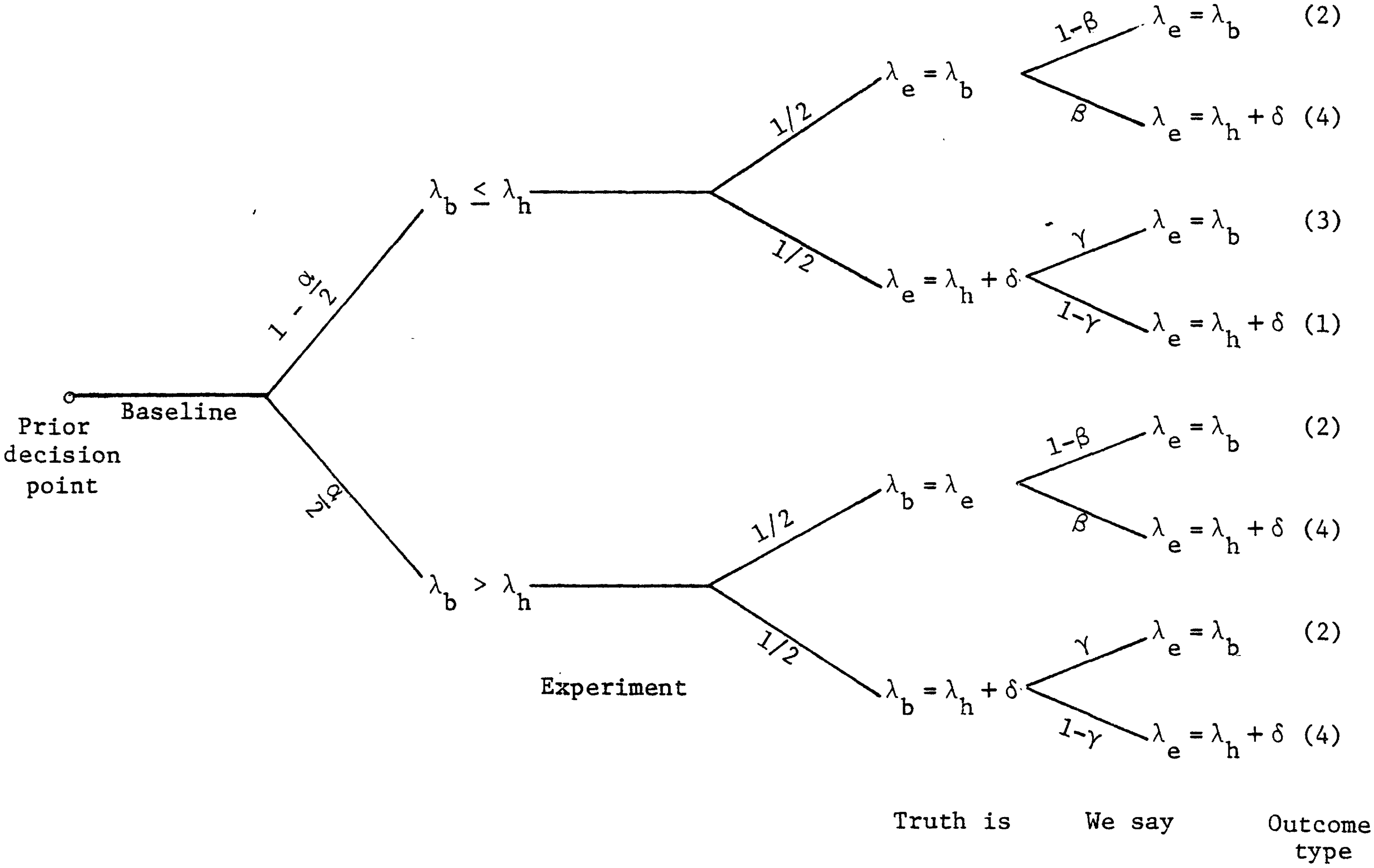
(1) Correct, an improvement occurred

(2) correct, no improvement occurred;

(3) wrong, there really is an improvement;

(4) wrong, there really is no improvement.

We would like to input a priori hunches P_0 and P_1 concerning whether $\lambda_e > \lambda_b$ or $\lambda_e = \lambda_h + \delta$, respectively. This complicates the issue, so to bow out gracefully, we will assume that $P_0 = P_1 = 1/2$, and only consider ranges for $\lambda_h + \delta - \lambda_b$ that would not give us any a priori reason to believe in one particular hypothesis more than the other (P_0 is actually some complicated function of α). Then the tree looks like:


Prior decision point
Baseline
$1 - \alpha/2$
$\alpha/2$
$\lambda_b \leq \lambda_h$
$\lambda_b > \lambda_h$
Experiment
1/2
1/2
1/2
1/2
$\lambda_e = \lambda_b$
$\lambda_e = \lambda_h + \delta$
$\lambda_b = \lambda_e$
$\lambda_b = \lambda_h + \delta$
$1-\beta$
β
γ
$1-\gamma$
$1-\beta$
β
γ
$1-\gamma$
$\lambda_e = \lambda_b$ (2)
$\lambda_e = \lambda_h + \delta$ (4)
$\lambda_e = \lambda_b$ (3)
$\lambda_e = \lambda_h + \delta$ (1)
$\lambda_e = \lambda_b$ (2)
$\lambda_e = \lambda_h + \delta$ (4)
$\lambda_e = \lambda_b$ (2)
$\lambda_e = \lambda_h + \delta$ (4)
Truth is
We say
Outcome type

If we associate a loss L_i with outcome (i), we might try to minimize:

$$L_1\left[\frac{1}{2}(1-\frac{\alpha}{2})(1-\gamma)\right] + L_2\left[\frac{1}{2}(1-\beta) + \frac{\alpha\gamma}{4}\right] + L_3\left[\frac{1}{2}(1-\frac{\alpha}{2})\gamma\right] + L_4\left[\frac{1}{2}\beta + \frac{1}{2}\left[\frac{\alpha}{2}\right](1-\gamma)\right]$$

by appropriate selection of γ and β. We can add a term for expected cost of observations given γ and β, or simply check that a chosen γ and β is feasible as far as sampling goes.

Owing to the complex nature of the preceding minimization task, we present an alternative. Outcomes (1) and (2) signify "correct" statements. We can redefine our confidence in terms of the predicted probability of being correct. Since we went into the investigation seeking a $1-\alpha$ CL in our original CI, let us shoot for overall $1-\alpha$ confidence. Setting P{outcome (1)} + P{outcome (2)} = $1 - \alpha$, we wind up with the following relationship:

$$\alpha\gamma + \frac{3}{2}\alpha - \gamma = \beta.$$

If we ignore the small $\alpha\gamma$ term and figure on choosing $\gamma = \beta$, we are thus guided to use $\beta = \gamma = \frac{3}{4}\alpha$. One could alternatively select γ and β to fit the above relationship exactly, perhaps using costs of wrong decisions.

At this point we might feel ready to begin. (The following decision process is summarized in the flow chart in Figure 4.) Suppose we proceed with baseline measurement , gathering the first two batches of data, and come up with an estimate of variance $S_b^{\ 2}$. We would want to sample n_b overall in the baseline stage, such that

$$\text{CI width} = 2 \cdot t_{n_b-1;\ 1-\alpha/2} \cdot S_b\sqrt{\frac{1}{n_b}} = 2d = 2 \text{ x (max.tolerable error).}$$

We can tinker with n_b to find an appropriate baseline sample size. We can also use S_b to estimate, assuming variance changes little moving to the experimental

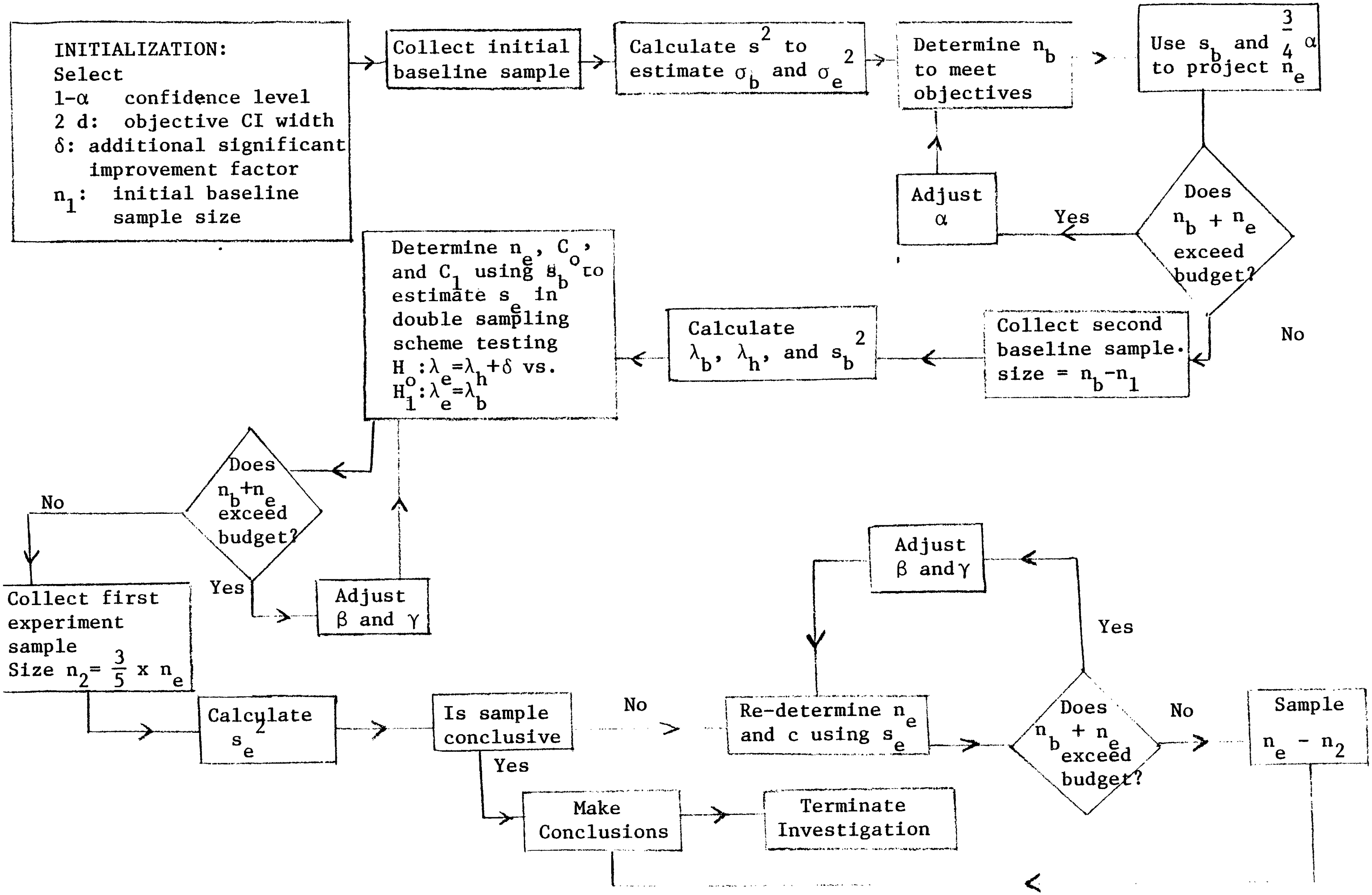

INITIALIZATION:
Select
1-α confidence level
2 d: objective CI width
δ: additional significant improvement factor
n_1: initial baseline sample size
Collect initial baseline sample
Calculate s^2 to estimate σ_b^2 and σ_e^2
Determine n_b to meet objectives
Use s_b and $\frac{3}{4}\alpha$ to project n_e
Does $n_b + n_e$ exceed budget?
Yes
Adjust α
No
Collect second baseline sample. size = n_b-n_1
Calculate λ_b, λ_h, and s_b^2
Determine n_e, C_o, and C_1 using s_b to estimate s_e in double sampling scheme testing
H_o:λ_e=λ_h+δ vs.
H_1:λ_e=λ_b
Does n_b+n_e exceed budget?
No
Yes
Adjust β and γ
Collect first experiment sample Size n_2= $\frac{3}{5}$ x n_e
Calculate s_e^2
Is sample conclusive
No
Yes
Make Conclusions
Terminate Investigation
Re-determine n_e and c using s_e
Does n_b + n_e exceed budget?
Yes
Adjust β and γ
No
Sample n_e - n_2

Figure 4: Flowchart for Proposed Estimate/Test Procedure

phase, the overall experimental sample size. Since we have assumed $\beta = \gamma$, then the criterion value $\underline{c}$ will lie midway between λ_b and $\lambda_h + \delta$, which ought to be $d + \delta$ apart. Therefore, using the Gaussian to approximate:

$$\Phi\left(\frac{-\frac{(d+\delta)}{2}\sqrt{n_e}}{S_b}\right) \approx \frac{3\alpha}{4} .$$

We can solve this for n_e. If the overall evaluation sample size or sampling cost exceeds our initial allotment , we know we must make adjustments, raising α and thereby γ and β. If there is some slack, we may ask if we desire more accuracy. Perhaps we may just use some of the slack in the experimental phase, honing γ or β (or both). Or we may conserve valuable information-gathering resources, perhaps to use in other investigations.

Whatever our adjustments, if any, we take a second baseline sample to bring our overall baseline sample size to n_b. We have λ_b and λ_h out of this, as well as a new S_b^2, which we use to guide our double sample hypothesis test in the experimental period. The first experimental sample might tell us to use a different standard error in our determination of n_e. Finally, we take our second sample and form conclusions.

We are going to play with the numbers one final time. Using the same arrival series as was generated before, we take 7 observations to estimate S_b initially. For

2 10 3 4 5 3 7

We get $S_b = 2.79$. This tells us that, in order to get an 80% CL on a CI for λ_b of length 2, we need about 14 observations. We also guess at n_e to gauge the feasibility of our objectives. Suppose $\delta = 1$ and $d = 1$. Then we find that n_e should be about nine. We can even make this ten or fifteen to compensate for not using the t-distribution, and we still have predicted $(n_b + n_e) < 30$. Assuming the same 40 observation limit, we have plenty of room to work in, so we take the next seven observations to conclude the baseline.

$$5 \quad 4 \quad 8 \quad 7 \quad 5 \quad 8 \quad 6,$$

and calculate the results $\bar{x}_b = 5.50$, $s_b = 2.28$, CI = (4.48, 6.32).

Note that we have determined a CI whose width falls inside our desired goal. We are appreciative, but we now must decide where we want to put our null hypothesis for λ_e. Let us stick it where $\lambda_h = 6.32$ becomes the criterion $\underline{c}$ of the experimental period; this occurs where $\lambda_e = 7.14$ (we by no means imply that this is the "only" or "best" place to put it). Now, using t-values, we would surmise that we need a sample size of $n_e = 15$, for

$$t_{14;0.15} \cdot 2.28\sqrt{\frac{1}{15}} \approx - 0.82 = 6.32 - 7.14.$$

(3/5) x $n_e = 10$; let us take ten observations for our first experimental batch.

This is no sooner said than done using the previous numbers, so we would see:

$$14 \quad 5 \quad 9 \quad 9 \quad 11 \quad 8 \quad 3 \quad 5 \quad 12 \quad 7; \; s_e = 3.43.$$

s_e is somewhat different than s_b here. This is a sign that we should re-calculate n_e based on this s_e, for which we would require $n_e = 37$. This is way over budget, so it looks as if we should either just sample out the rest of our budget, or reconsider our desired error rates. We might have foreseen this complication if we took the "Poisson-ness" of these observations into account, meaning that variance increases as mean does. We will not push this example any further, though.

V. Summary

We have looked at several possible routes toward a sensible approach to adaptive evaluation of public programs using classical statistics. The SPRT, double sampling, and stratified sampling techniques were singled out for particular exploration. Although no samples involving estimation of Gaussian parameters were shown, Gaussian approximations were employed to a degree that should give sufficient insight into how they should be handled. And with that, we can see how this small group of techniques might be used to handle a wide variety of problems. The final verdict on these techniques lies a good way off; more empirical and analytical investigation is necessary to determine the worth of the contents of this paper.

References

[1] Cochran, W. G. (1977), Sampling Techniques, John Wiley and Sons, New York.

[2] Guttman, I., Wilks, S.S. and Hunter, J.S. (1965), Introductory Engineering Statistics, John Wiley and Sons, New York.

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

[4] Minkoff, A.S. (1981), "Preliminary Survey of Classical Statistical Techniques for Incorporation into Adaptive Evaluation Methodology," Working Paper, Operations Research Center, MIT.

[5] Wald, A. (1947), Sequential Analysis, John Wiley & Sons, New York.