OPTIMAL BAYESIAN TWO-PHASE DESIGNS FOR SCREENING TESTS

ALAATTIN ERKANLI
ADRIAN ANGOLD
Developmental Epidemiology Program
Duke University Medical Center
REFIK SOYER
Department of Management Science
The George Washington University
and
Institute of Statistics & Decision Sciences
Duke University

DP #94-41

OPTIMAL BAYESIAN TWO-PHASE DESIGNS FOR SCREENING TESTS

Alaattin Erkanli*, Refik Soyer# and Adrian Angold*

Summary

In this paper we present a Bayesian decision theoretic approach by formulating the two-phase design problem as a sequential decision problem. The solutions of such problems is usually difficult to obtain because of their reliance on *preposterior* analysis. In overcoming this problem, we adopt the Monte Carlo based approach of Müller and Parmigiani (1994) and develop optimal Bayesian designs for two-phase screening tests. A rather attractive feature of the Monte Carlo approach is that it facilitates the preposterior analysis by replacing it with a sequence of scatter plot smoothing/regression techniques and the optimization of the corresponding fitted surfaces. The method is illustrated for depression in adolescents using data from past studies.

1. Introduction

The purpose of this article is to develop optimal Bayesian two-phase designs for the estimation of the prevalence of a rare disorder in a given population. In general, the aim is to estimate the prevalence with high precision. In a single-phase design, one can simply randomly sample sufficiently large number of subjects to reach a prescribed precision around the prevalence estimate. However, for an uncommon disorder such a strategy may be very expensive and time consuming. Since research funds are rarely an infinite resource, the cost of the study becomes a real determinant of the degree of precision with which prevalence can be measured.

*Alaattin Erkanli is an Assistant Research Professor of Statistics and Adrian Angold , MRCPsych, is the Director of Center for the Study, Prevention, and Treatment of Disruptive Behavior Disorders, Developmental Epidemiology Program, Department of Psychiatry and Behavioral Sciences, Duke University Medical Center, Box 3454, Durham, NC 27710.
Refik Soyer is an Associate Professor of Management Science, Management Science Department, School of Business and Public Management, George Washington University, Washington, DC 20052. This work was completed while the second author was visiting the Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708.

The problem then is how to get the best precision for a fixed amount of funding. One strategy for uncommon disorders is to adopt a two-phase sampling design. The idea is to screen the population using a moderately effective but relatively cheap instrument at the first phase (Deming, 1977) and then administer a more detailed but more expensive diagnostic test to a subset of the subjects chosen at the first phase. The problem is how to optimize the trade-off between estimation and budget constraints.

Previous work, based on ideas from sampling theory, (Cochran, 1977; Shrout and Newman, 1989) has focused on the development of optimal designs through minimizing the asymptotic variance of the MLE of the prevalence subject to budget constraints. However, in this approach, the optimal design depends on the unknown population parameters that cannot be known with certainty prior to the experiment. It is therefore necessary to plug in some auxiliary estimates, based on pilot studies or expert opinion, in place of the unknown parameters hoping that they are close to their true values, thus leading to a locally-optimal design. Furthermore, the sampling theory approach reduces for what is really a *sequential* two-phase design problem to a fixed design problem which usually leads to sub-optimal designs in non-linear problems (Chaloner and Verdinelli, 1994).

In this paper we present a Bayesian decision theoretic approach that formulates the optimal two-phase design as a sequential decision problem (Berry, 1991). The solution of such problems is usually difficult to obtain because of their reliance on *preposterior* analysis. In overcoming this difficulty, one approach is to use asymptotic approximations to the posterior distributions of the parameters of interest (David, 1970). However, when the sample size under consideration is relatively small, or there are large number of parameters and design variables, the methods based on asymptotic expansions often fail. Herein, as a computational alternative, by using the recent Monte Carlo based approach of Müller and Parmigiani (1994), we develop optimal Bayesian designs for two-phase screening tests. A rather attractive feature of the Monte Carlo approach is

that it facilitates the preposterior analysis by using a sequence of scatter plot smoothing/regression techniques and the optimization of the corresponding fitted surfaces.

A synopsis of this paper is as follows: In Section 2, we introduce the two-phase design problem and the classical approach based on sampling theory methods. In Section 3, we formulate the optimal design problem as a sequential decision problem and introduce the dynamic programming solution. Moreover, we describe the uncertainty about unknown parameters of the experiment in terms of Beta-ordered Dirichlet prior distributions. This choice of priors provides flexibility in reflecting the prior beliefs about the screening efficiency. In Section 4, we discuss approximate Bayesian designs using asymptotic results. Later, in Section 5, we introduce the Monte Carlo approach that enables us to approximate the solution of the dynamic program in the sequential formulation of the two-phase design. In Section 6, we illustrate these approaches by combining information from eleven large-scale studies concerning the depression in young children and adolescents. Section 7, we talk about possible extensions for future research. Conclusions are in Section 8.

2. Two-phase Designs and the Classical Approach

Consider the following two-phase design of experiment: At the first phase, N subjects are selected randomly from a large population and a screening test is administered. Those who get a screening score that is a greater than a threshold value are classified into *screen high* group, $S+$; otherwise they are classified into *screen low* group, $S-$. Let $n(S+)$ and $n(S-)$ be the sizes of these groups, respectively. Let f_1 be a fraction of $n(S+)$ and f_2 a fraction of $n(S-)$, such that $0 \le f_1, f_2 \le 1$. We define $n_1 = f_1 \cdot n(S+)$ and $n_2 = f_2 \cdot n(S-) = f_2 \cdot \{N - n(S+)\}$. The second phase of the design consists of giving a diagnostic test to the n_1 and n_2 subjects. Let $D+$ denote subjects with the diagnosis, while $D-$ denotes those without the diagnosis.

Define $\lambda_1 = P(D+|S+)$, $\lambda_2 = P(D+|S-)$ and $p = P(D+)$; the prevalence of the disorder in $S+$, $S-$, and in the entire population, respectively. Let $\pi = P(S+)$ be the probability of screening high. The prevalence of the disorder can be calculated from the equation $p = \lambda_1\pi + \lambda_2(1-\pi)$. The goal is to design an optimal experiment for the estimation of p by choosing the values of N, f_1 and f_2 in such a way that a criterion function is optimized.

We note that the sampling model at the first phase is given by a binomial distribution $[n(S+)|N, \pi] \sim \text{Bin}(N, \pi)$. Once f_1 and f_2 are specified, we observe two independent binomial outcomes at the second phase, conditioned on screen status: $[x_i|n_i,\lambda_i] \sim \text{Bin}(n_i,\lambda_i)$ for $i = 1, 2$, where (x_1,x_2) are the total number of cases with the diagnosis in $S+$ and $S-$ groups.

The standard approach to two-phase design problems is to minimize the variance of the prevalence estimator given the budget restrictions. This approach has been discussed in detail by Cochran (1977) and more recently by Shrout and Newman (1989) and will be reviewed here, for the sake of completeness. Let $\hat{p}$ be the MLE of p, which can be obtained by substituting the MLEs $\{x_1/n_1, x_2/n_2, n(S+)/N\}$ of λ_1, λ_2 and π in $p = \lambda_1\pi + \lambda_2(1-\pi)$. For fixed N, f_1 and f_2, the asymptotic variance of $\hat{p}$ is given by

$$Var(\hat{p}|\lambda_1,\lambda_2,\pi) = \frac{1}{N}\left\{\frac{\pi\lambda_1(1-\lambda_1)}{f_1} + \frac{(1-\pi)\lambda_2(1-\lambda_2)}{f_2} + \pi(1-\pi)(\lambda_1-\lambda_2)^2\right\} \qquad (2.1)$$

Note that $\hat{p}$ is an unbiased estimator of p and minimizing its (quadratic) risk $E(\delta(x_1,x_2,n(S+)) - p)^2$ among all unbiased estimators δ of p reduces to minimizing the variance of $\hat{p}$, where the expectation is taken with respect to its sampling distribution. The next step is to determine N, f_1 and f_2 so that (2.1) is minimum for a given budget. Let C be the expected budget, and c_S and c_d be the unit costs for screening test and diagnostic tests, respectively:

$$C = N \cdot c_S + c_d E(n_1 + n_2) = N \cdot [c_S + c_d \{\pi f_1 + (1-\pi) f_2\}]. \quad (2.2)$$

To find the "optimal" N, f_1 and f_2, one can solve (2.2) for N as a function of f_1 and f_2 and then minimize (2.1) with respect to f_1 and f_2. The optimal N is in turn found by substituting these values into the equation $N = C/[c_s + c_d\{\pi f_1 + (1-\pi) f_2\}]$, for $0 < f_k < 1$, $k=1,2$. Simple calculations show that the optimal design is given by

$$f_1 = \sqrt{\frac{\lambda_1(1-\lambda_1)}{(\lambda_1-\lambda_2)^2\pi(1-\pi)}} \cdot c, f_2 = \sqrt{\frac{\lambda_2(1-\lambda_2)}{(\lambda_1-\lambda_2)^2\pi(1-\pi)}} \cdot c, \quad (2.3)$$

where $c = c_S / c_d$. If f_k is greater than or equal to 1, then its optimal value is at the boundary 1. The optimal fractions are determined by the ratio of the within-group population variances $\lambda_j(1-\lambda_j), j = 1,2$, to the between-group population variance $(\lambda_1 - \lambda_2)^2 \pi(1-\pi)$. At this point it is clear that there is a problem, because the optimal fractions are functions of the unknown parameters λ_1, λ_2 and π which we are trying to estimate in the first place. The standard practice is to substitute some auxiliary estimates for these parameters, hoping that they are close to the true values of these parameters. However, without any measure of uncertainty, the resulting designs are only locally-optimal. We therefore have an infinity of optimal designs depending on the chosen values of the unknowns.

3. Modeling Uncertainty about the Prevalence

In Bayesian thinking, as all uncertainties must be described probabilistically, in our particular application, the distributions of parameters λ_1, λ_2 and π and p must be specified prior to the experiment. In many real applications, e.g., the assessment of the prevalence of depression in children and adolescents, there is often a good deal of prior information, either based on clinical research such as case-control studies, or obtained from large-scale previous population studies. However, it will rarely be the case that all previous studies agree as to the prevalence-if

they did there would be little point in conducting a new study of the topic. For example, in the case of childhood depression, prevalence estimates from nine previous studies range from 1% to 16%, with most estimates lying between 1% and 5% (Angold et al, 1993). Thus empirical information about the distribution of unknown parameters is available and the Bayesian approach offers a probabilistic means to utilize this information in formulating an optimal design.

A natural restriction on the prevalence given the screening status is to assume that $\lambda_1 \geq \lambda_2$, which reflects the prior belief that the screening must be an efficient procedure to identify more cases with the diagnosis. In fact, it is unlikely that a two-phase study would be contemplated in the absence of data demonstrating that the screen did indeed predict the diagnosis. To reflect this prior belief, we consider a Dirichlet prior distribution (Wilks, 1962) on the functions $1-\lambda_1, \lambda_1-\lambda_2$ and λ_2. Its density is given by

$$f(\lambda_1,\lambda_2) = c(1-\lambda_1)^{\alpha\alpha_1-1}(\lambda_1-\lambda_2)^{\alpha\alpha_2-1}(\lambda_2)^{\alpha\alpha_3-1}, \qquad (3.1)$$

where $\sum_{k=1}^{3}\alpha_k = 1$, $\alpha_k \geq 0, \alpha > 0$, and c is the normalizing constant. This density is defined over the simplex $\{(\lambda_1,\lambda_2)$: $1 \geq \lambda_1 \geq \lambda_2 \geq 0\}$. Thus the desired ordering and probability relationships are preserved without introducing any extraneous restrictions on the random quantities of interest. It is worth noting that if the ordinary Dirichlet distribution were used, the correlation between λ_1 and λ_2 would be negative despite the assumption that $1 \geq \lambda_1 \geq \lambda_2 \geq 0$. The correlation under ordered Dirichlet prior described in (3.1) is, on the other hand, given by $\rho = [\alpha_1\alpha_3/\{(\alpha_2+\alpha_3)\cdot(\alpha_1+\alpha_2)\}]^{1/2}$, which is always positive.

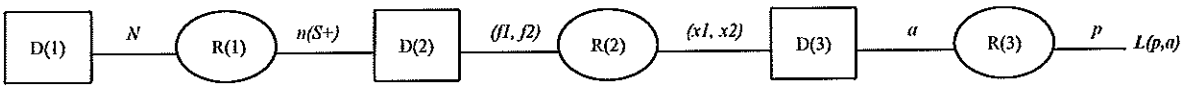
The ordered Dirichlet prior was used by Mazzuchi and Soyer (1993) in relation to product reliability testing, and by Gelfand and Kuo (1991) for bioassay problems. It easy to see that the

marginal distributions of λ_1 and λ_2 are given by beta densities with hyperparameters $[\alpha(1-\alpha_1),\alpha\alpha_1]$ and $[\alpha\alpha_3,\alpha(1-\alpha_3)]$, respectively. It can also be shown that the conditional distributions $[\lambda_i|\lambda_j, i \neq j]$ are given by truncated beta densities. The hyperparameters are determined by the analyst prior to the study. The prior of π is chosen as a beta distribution with hyperparameters α_0 and β_0. Note that this prior is a member of the Beta-Binomial conjugate family and thus provides a simple way of describing the prior information about the screening status. The parameters λ_1 and λ_2 are assumed statistically independent of π. The choices of the hyperparameters of these distributions reflect the strength of beliefs or information about the location and the spread of λ_1, λ_2, π, $\lambda_1 - \lambda_2$, and hence of p.

4. The Bayesian Formulation of the Two-Phase Design Problem

The classical approach to the two-phase design problem treats it as a fixed design problem, since the formulas in (2.3) do not depend on the outcome of the first phase of the study but on the value of the unknown parameters. In reality, however, the two-phase design is a sequential process. At the first phase, the screening sample size N is determined, the screening test is given, and depending on the outcomes $n(S+)$ and N-$n(S+)$, the sample sizes n_1 and n_2 are determined. Then, at the second phase, the diagnostic test is given and the outcomes x_1 and x_2 are observed. Finally, inference is made about the prevalence using a loss function. The decision tree in Figure 1 describes the two-phase design process as a sequential decision problem.

Figure 1. Decision Tree Representation of the Two-Phase Design Problem



In Figure 1, the decision node D(1) represents the specification of N prior to the first phase of the experiment. Random node R(1) represents the outcome $n(S+)$ of the first phase. Decision node

D(2) refers to the choice of f_1 and f_2 after $n(S+)$ has been observed and R(2) denotes the binomial outcomes x_1 and x_2 of the second phase. The selection of the Bayes rule a given the outcomes from the two phases is represented by decision node D(3). Finally, the random node R(3) represents the unknown prevalence p, and the $L(p,a)$ denotes the associated loss.

The solution of the design problem is obtained in the conventional manner by using dynamic programming. In other words, the solution can be obtained by "folding back" the decision tree through taking expectations at random nodes and minimizing the expected loss at the decision nodes of Figure 1 (Lindley, 1985). The following steps describe this procedure.

At **R(3):**

Evaluate $E_{p|x_1,x_2,n(S+)}[L(p,a)|x_1,x_2,n(S+)]$.

At **D(3):**

Compute $a^* = \text{argmin}\ E_{p|x_1,x_2,n(S+)}[L(p,a)|x_1,x_2,n(S+)]$ and evaluate

$\sigma^2(x_1,x_2,n(S+),N,f) = E_{p|x_1,x_2,n(S+)}[L(p,a^*)|x_1,x_2,n(S+)]$.

At **R(2):**

Evaluate $\sigma^2(n(S+),N,f) = E_{x_1,x_2|n(S+)}[\sigma^2(x_1,x_2,n(S+),N,f)|n(S+)]$,

where $f = \{f_1, f_2\}$.

At**D(2):**

Compute f^*=argmin$\{\sigma^2(n(S+),N,f)\}$ and evaluate $\sigma^{2*}(N,n(S+)) = \sigma^2(n(S+),N,f^*)$.

At **R(1):**

Evaluate $\sigma^{2*}(N) = E_{n(S+)}[\sigma^{2*}(n(S+),N,f^*)]$.

At **D(1):**

Compute N^* = argmin$\{\sigma^{2*}(N)\}$, evaluate $\sigma^{2*}(N^*)$and terminate with the optimal N^*.

Here, the operators $E_{p|x_1,x_2,n(S+)}[\cdot]$, $E_{x_1,x_2|n(S+)}[\cdot]$ and $E_{n(S+)}[\cdot]$ denote expectations with respect to the posterior distribution of p given the data $\{x_1,x_2,n(S+)\}$, the conditional prior predictive

distribution of x_1, x_2 given $n(S+)$, and the prior predictive distribution of $n(S+)$, respectively. We note that the solution of the optimal design problem is not trivial as it involves implicit computations of expectations and minimizations at each step. A computational strategy based on Monte Carlo simulations and smoothing techniques is discussed in Section 5.

5. Bayesian Solutions to Two-Phase Design Problem

5.1. A Naive Bayesian Approach

Under the probabilistic structure introduced in Sections 3 and 4, the solution to the sequential design problem can not be obtained in closed forms. One solution to this problem is to ignore its inherent sequential nature and treat it as a fixed design problem. Then, assuming the quadratic loss $L(p,a) = (p-a)^2$, the optimal design problem reduces to choosing the decision variables by minimizing the risk $E_x\{Var(p|x)\}$, where $x = (x_1, x_2, n(S+))$ and $Var(p|x)$ is the posterior variance of p, which involves weighted sums of Beta functions with arguments being nonlinear functions of the design variables. Thus, the minimization of the risk function with respect to the design variables is computationally infeasible even for moderate number of variables, see for example, Mazzuchi and Soyer (1993).

It is well known that, under regularity conditions, one can approximate the posterior distribution sufficiently closely by a normal distribution when the sample size is large enough. If the prior distribution is locally uniform around the MLE's of the parameters, such an approximation to the posterior distribution of the prevalence p can be obtained by using a Taylor expansion (i.e., the delta method) of p in the neighborhood of the MLE's of λ_1, λ_2, and π so that approximately, $p|x \sim N(\hat{p}, \sigma^2(x))$, where $\hat{p}$ is the MLE of p and $\sigma^2(x)$ is the asymptotic variance of p, evaluated at the MLE' s λ_1, λ_2, and π. Under the quadratic loss, the Bayes rule is the posterior mean $a^*(x) = \hat{p}$, and its associated posterior risk is the posterior variance $\sigma^2(x)$.

Thus, $E_x[Var(p|x)] \approx E_x\{\sigma^2(x)\}$, where the expectation is taken with respect to the marginal (prior predictive) distribution of x. After some algebra, we arrive at the expression

$$E_x\{\sigma^2(x)\} = E_{\pi,\lambda_1,\lambda_2}\left[E_{x|\pi,\lambda_1,\lambda_2}\{\sigma^2(x)|\ \pi,\lambda_1,\lambda_2\}\right] = E_{\pi,\lambda_1,\lambda_2}\left[Var(\hat{p}|\pi,\lambda_1,\lambda_2)\{1+O(N^{-1})\}\right],$$

where $Var(\hat{p}|\lambda_1,\lambda_2,\pi)$ is the asymptotic variance of the sampling distribution of $\hat{p}$ given by the equation (2.1) of Section 2. Thus, the naive Bayesian optimal design, to the errors of order N^{-1}, is determined by the optimization $\min_{(N,f_1,f_2)\in\mathcal{C}} E\{Var(\hat{p}|\lambda_1,\lambda_2,\pi)\}$, where $\mathcal{C}$ is the constraint set determined by the design restrictions, and $E(\cdot)$ denotes the expectation with respect to the prior distribution of λ_1, λ_2 and π. Thus, the problem becomes,

$$\min_{(N,f_1,f_2)\in\mathcal{C}} \frac{1}{N}\Big[\ \frac{E(\pi)\mathrm{E}\{\lambda_1(1-\lambda_1)\}}{f_1} + \frac{\{1-E(\pi)\}E\{\lambda_2(1-\lambda_2)\}}{f_2} + E\{\pi(1-\pi)\}E(\lambda_1-\lambda_2)^2\ \Big] \quad (4.1)$$

It can be seen easily that when the prior distribution is degenerate at a point $(\lambda_{10},\lambda_{20},\pi_0)$ in the parameter space, then (4.1) reduces to the locally optimal design described in Section 2. As an example, we consider the cost constraint given by (2.2), which can be represented in terms of the prior distribution of π as $C = N[c_S + c\cdot\{E(\pi)f_1 + (1-E(\pi))f_2\}]$. Using the ordered Dirichlet and beta priors, it can be seen that optimal fractions are obtained by replacing the parameter values by their prior expectations in the formulas described in (2.3), where

$$E(\pi) = \alpha_0/(\alpha_0+\beta_0),$$

$$E(\lambda_1-\lambda_2)^2 = \alpha_2(\alpha\alpha_2+1)/(\alpha+1),$$

$$E(\lambda_1(1-\lambda_1)) = \alpha\cdot\alpha_1(1-\alpha_1)/(\alpha+1),$$

$$E(\lambda_2(1-\lambda_2)) = \alpha\cdot\alpha_2(1-\alpha_2)/(\alpha+1),$$

and

$$E\{\pi(1-\pi)\} = \alpha_0\beta_0/[(\alpha_0+\beta_0)(\alpha_0+\beta_0+1)].$$

In some cases, all of the subjects identified as belonging to the high screening group in the first stage may be included in the second stage i.e., $f_1 = 1$. The naive Bayesian design in this case is given by

$$f_2 = \sqrt{\frac{E[\lambda_2(1-\lambda_2)[E(\pi)+c]}{E(\pi)\mathrm{E}[\lambda_1(1-\lambda_1)]+E[\pi(1-\pi)]E(\lambda_1-\lambda_2)^2}}.$$

5.2. A Monte Carlo Solution to Sequential Design

Müller and Parmigiani (1994) approach the sequential design problems using Monte Carlo methods, where the expectation steps in the dynamic program are replaced by scatter plot smoothers based on Monte Carlo samples of $\{\pi^{(r)},\lambda_1^{(r)},\lambda_2^{(r)},x_1^{(r)},x_2^{(r)},n^{(r)}(S+)\}$. The minimization steps are replaced by the minimization of the fitted smoother. To apply their ideas to the two-phase design problem, let $d_1 = N$, $d_2 = \{f_1, f_2\}$ denote the design variables and define $y_1 = n(S+)$, $y_2 = (x_1, x_2)$, and $\theta = (\pi, \lambda_1, \lambda_2)$. The following algorithm describes implementation of the Monte Carlo approach of Müller and Parmigiani to our case.

Step 1. Select designs $d_i = (d_{i1}, d_{i2})$, $i = 1,..., M$;

Step 2. Draw M points θ_i, y_{i1}, y_{i2} from $f(y_1, y_2 | d_i, \theta) \cdot f(\theta)$. Compute $l_{i2} = L(p_i, a_i^*)$ and record the Monte Carlo sample points $(d_{i1}, d_{i2}, y_{i1}, l_{i2})$. Note that under the quadratic loss, $a_i^* = E(p | y_{i1}, y_{i2})$, the posterior mean of p. Steps 1 and 2 together are analogous to the operations performed at the nodes R(3) and D(3) of the decision tree described in Figure 1.

Step 3. Fit a surface $\pounds_2(d_1, d_2, y_1)$ to the points $(d_{i1}, d_{i2}, y_{i1}, l_{i2})$. This step is analogous to the expectation performed at the node R(2).

Step 4. Find the minimum over d_2 of $\pounds_2(d_1, d_2, y_1)$. Replace the Monte Carlo points by the pairs (d_{i1}, l_{i1}), where $l_{i1} = max_{d_2} \pounds_2(d_{i1}, d_2, y_{i1})$. Here the minimal value of $d_2 = \{f_1, f_2\}$ is the solution to the second stage design problem and it depends on $d_1 = N$ and $y_1 = n(S+)$. This step is analogous to the minimization performed at the node D(2).

Step 5. Fit a curve $\pounds_1(d_1)$ to the points (d_{i1}, l_{i1}).This step is analogous to the expectation performed at the node R(1).

Step 6. Find the minimum over d_1 of $\pounds_1(d_1)$. Here the minimal value of $d_1 = N$ is the solution to the first stage design problem .This step is analogous to the minimization performed at the node D(1).

In Step 1, we chose the design sequence $\{d_i\}$ deterministically from a set of feasible candidates. For example, they can be chosen in the vicinity of previously found optimal designs or based on prior experiments. In Step 2, for each design point d_i, we solve a Bayesian decision problem based on the simulated data $\{\theta_i, y_{i1}, y_{i2}\}$ from $f(y_1, y_2 | d_i, \theta) \cdot f(\theta)$. We then calculate $l_{i2} = L(p_i, a_i^*)$ for each experiment and collect all the sample points (d_{i1}, d_{i2}, y_{i1}, l_{i2}). The Bayes rule a_i^* can be obtained by using a Gibbs sampler (Gelfand and Smith, 1990) based on an auxiliary Monte Carlo simulation of size, for example K, from the conditional posterior distributions of λ_1, λ_2 and π. The disadvantage of Gibbs sampling is the increase in total computation time. Alternatively, since the sample size in these studies is typically large, analytical approximations can also be used, e.g., the MLE of p_i. In fact, in the example below the MLE was used throughout the calculations as an approximation to the Bayes rule a^*.

Here, we think of the collection (d_{i1}, d_{i2}, y_{i1}, l_{i2}) as the data of a regression problem with l_{i2} as the dependent variable and d_{i1}, d_{i2}, y_{i1} as the predictors, or the independent variables. In Step 3, we fit a (nonparametric) regression model to the scatter (d_{i1}, d_{i2}, y_{i1}, l_{i2}). The fitted regression surface $\pounds_2(d_1, d_2, y_1)$ is a scatter plot smoothing approximation to the conditional expectation surface $E_{x_1,x_2,\theta|n(S+),N,f_1,f_2}\{L(p,a)|n(S+),N,f_1,f_2\}$. In Step 4, we minimize the fitted surface $\pounds_2$ with respect to $d_2 = \{f_1, f_2\}$. This may require numerical techniques since most nonparametric regression surface estimators are not available in closed forms. The resulting optimal value of d_2 is a random function of $n(S+)$, and a deterministic function of N given $n(S+)$. We substitute this value in $\pounds_2$ to obtain $l_{i1} = max_{d_2} \pounds_2(d_{i1}, d_2, y_{i1})$. Thus, we have now the *new*

regression data $(d_{i1}, l_{i1}) = (N_i, l_{i1})$ and we perform a one-dimensional regression curve fitting. The value N^* of N that minimizes the fitted curve $£_1(N)$ is the optimal N.

Steps 3 and 5 involve a four dimensional and a univariate smoothing. In the example below, we used the local regression model *Loess* (Chambers and Hastie, 1992) to fit a smoothing surface in Step 3. In Step 5, any reasonable one-dimensional smoothing method will be sufficient, e.g. smoothing splines. Once the optimal N^* has been found, we perform the experiment and observe a value of $n^*(S+)$ from the screening test at the first phase of the sequential process. Now, the optimal fractions $(\hat{f}_1, \hat{f}_2)$ have to be determined to step up to the next phase. If $\hat{f}_1 = \hat{f}_1(N^*, n^*(S+))$ and $\hat{f}_2 = \hat{f}_2(N^*, n^*(S+)$ are available in closed forms, then finding the optimal fractions is just a simple matter of calculation. If, on the other hand, $(\hat{f}_1, \hat{f}_2)$ are not available in closed functional forms they should be determined numerically. To do so, we can plot the fitted regression surface $£_2(d_1, d_2, y_1)$ evaluated at N^* and $n^*(S+)$, as a function of $d_2 = \{f_1, f_2\}$ and numerically find the optima. In practice, given N^*, several possible scenarios can be considered by generating a collection of values of $n^*(S+)$ from the Beta-Binomial distribution with parameters (N^*, α_0, β_0).

6. Example: Prevalence of Depression in Children and Adolescents.

The Great Smoky Mauntains Study (GSMS) was designed to assess met and unmet need for mental health services for children and adolescents. In order to maximize the number of cases of psychiatric disorder recruited into the study, a brief screening questionnaire was administered to a random sample of approximately 4,500 parents. Pilot testing had determined a cutoff score on the questionnaire that identified a quantile of the population at increased risk for mental health problems and mental health service use. In the main study, all of those who screened above this cutpoint were recruited into an interview phase for diagnosis. In addition, 10% of those screened below the cutpoint were also interviewed. It was clear at the time that these proportions were not

optimal as far as producing prevalence estimates were concerned, but other considerations rendered this design desirable.

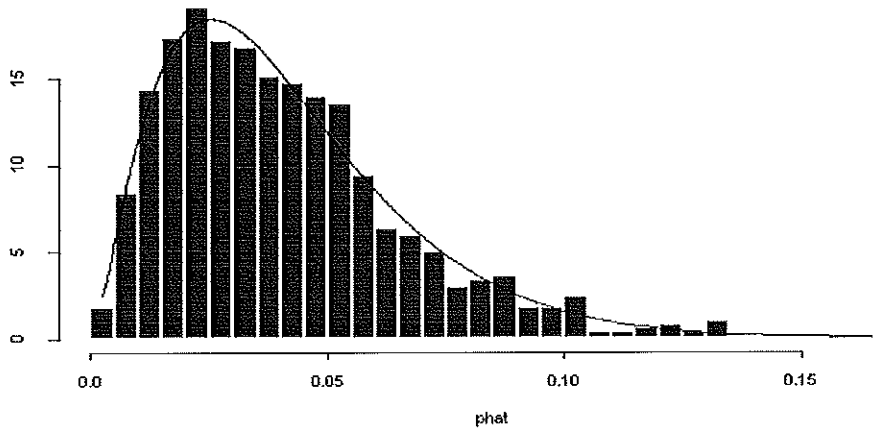
However, for the purpose of designing a new study where the optimal estimation of the prevalence was a more central consideration, we needed to determine first whether a two-phase strategy would be the right approach and if so what proportions should be selected from the screen high and screen low groups. For the purpose for this example we shall consider an optimal design for the prevalence of depression for a future study. In assessing the prior distribution of the prevalence of depression, we used the results of ten studies conducted in the past in several locations in the United States and overseas (Adrian et al., 1993), in addition to the results from GSMS. The estimates of the prevalence of depression in these eleven studies changed from 1.8% to 8% with an overall mean of approximately 4% and a standard deviation of approximately 2.2%. The GSMS estimate based on the $N = 4500$ screened subjects was 1.45%.

The prior for π was chosen to be a beta distribution with parameters $(\alpha_o, \beta_o) = (4.44, 13.32)$, which give a prior mean and standard deviation of 0.25 and 0.1, respectively. This choice was not arbitrary. Only four out of eleven studies were conducted using two-phase design. The proportion in $S+$ was estimated as 25% in GSMS and varied slightly in the other three studies. The prior standard deviation of 0.1 was chosen to provide sufficient prior uncertainty around 25%. The GSMS data produced a standard error of approximately 0.006 for the MLE of π based on the screening sample size of $N = 4500$ subjects.

For the prior of (λ_1, λ_2), the hyperparameters were chosen as $(\alpha_1, \alpha_2, \alpha_3) = (0.929, 0.039, 0.032)$ and $\alpha = 54$. These values give the prior means of (λ_1, λ_2) as 7.1% and 3.2% (the GSMS estimates were 3.3% and 0.78%), the depressive disorder in the screen high and low groups. The corresponding prior standard deviations of (λ_1, λ_2) are 3.5% and 2.3%. (Note that the choice $(\alpha_1, \alpha_2, \alpha_3) = (1/\alpha, 1/\alpha, 1/\alpha)$ and $\alpha = 3$ give the uniform distribution on the simplex

$\{(\lambda_1,\lambda_2) : 1 \geq \lambda_1 \geq \lambda_2 \geq 0\}$ which in turn gives the means (0.66, 0.33) and the same standard deviation of 0.23 for λ_1 and λ_2). These choices correspond to the belief that, on the average, the proportion of subjects with the diagnoses in the high risk group S+ are about twice the proportion of subjects in the low risk group S-. The prior of p is obtained by using a simple Monte Carlo sampler which simulates successively from the conditional density $[\lambda_1|\lambda_2]$~Beta($\alpha\alpha_2$,$\alpha\alpha_1$)~$I([\lambda_2, 1])$ and from the marginal density $[\lambda_2]$~Beta($\alpha\alpha_3$,$\alpha(\alpha_1+\alpha_2)$), where $I([a, b])$ denotes the indicator function of the interval [a, b]. Thus, to simulate the prior distribution of the prevalence, we simulate a Monte Carlo sample of size M from the prior of π, and a Monte Carlo sample of size M from the joint prior of $\{\lambda_1,\lambda_2\}$ and then use the histogram of $p = \lambda_1\pi + \lambda_2(1-\pi)$, as displayed in Figure 2 below. Although the exact distribution of the prevalence is not known, the solid line nevertheless shows the matching Beta density with the mean and the standard deviation 4% and 2.2%, respectively. Thus, the prior distribution of the prevalence is well approximated by a Beta distribution.

Figure 2. Prior distribution of p: the grand prevalence of depressive disorders.



6.1. Optimal Designs based on the naive Bayesian approximation

For the GSMS study the cost ratio was determined to be around c = 1:18 i.e., the diagnostic test is 18 times more expensive than the screening test. The Bayesian design produces the optimal fractions $f_1 = f_2$ = 100%. With f_1 = 1, the Bayesian optimal design gives f_2 = 67%, when the cost ratio is equal to zero, and increases proportionally to the square root of c. For example, when c = 1:18; f_2 = 74%, and when c = 1:2; f_2 = 100%. In both designs, the optimal screening sample size N can be found once the total expected budget C is determined.

6.2. Optimal designs based on the Monte Carlo approach

Here, we illustrate the Monte Carlo approach using the GSMS prior. For the sake of illustration, we again choose the quadratic loss function although more complicated loss functions can be used. To find the optimal design, we judiciously picked M = 1000 design points $\{d_{i1}, d_{i2}\}$, and simulated M = 1000 random vectors $\{x_{i1}, x_{i2}, n_i(S+), p_i\}$, for i = 1,..., M, from the prior of p and from the joint distribution $[x_1|n_1,\lambda_1]\cdot[x_2|n_2,\lambda_2]\cdot[n(S+)|N,\pi]$, with the components given by $[x_1|n_1,\lambda_1]\sim\text{Bin}(n_1,\lambda_1)$, $[x_2|n_2,\lambda_2]\sim\text{Bin}(n_2,\lambda_2)$ and $[n(S+)|N,\pi]\sim\text{Bin}(N,\pi)$.

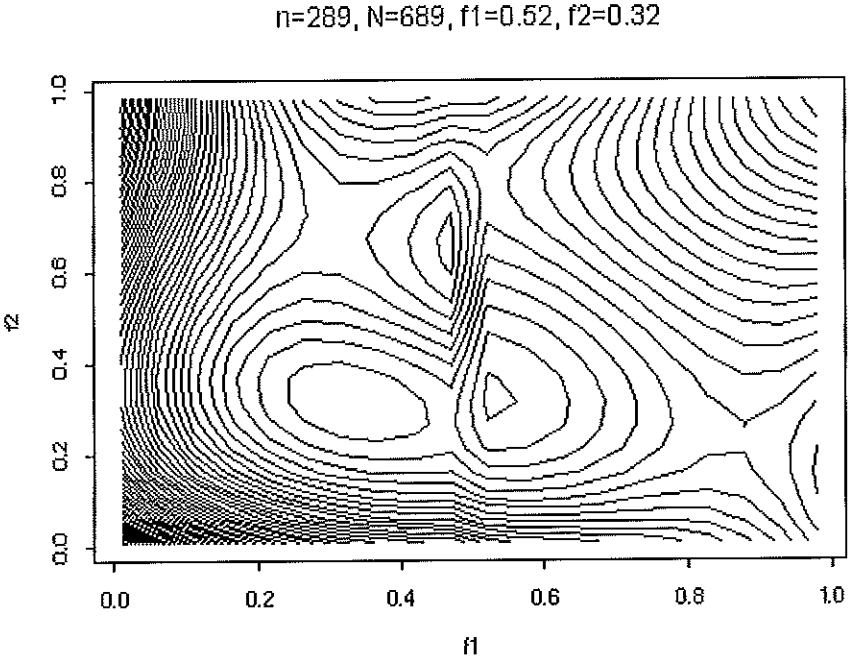
For each $\{x_{i1}, x_{i2}, n_i(S+), p_i\}$, we then approximate the Bayes rule $a_i^* = E(p|x_{i1}, x_{i2}, n_i(S+))$, the posterior mean of p, by its MLE and evaluate the loss function $(p_i - a_i^*)^2$. The grid (and the simulation) size M =1000 is chosen for illustrative purposes only. In more complicated problems, we might choose larger M to obtain better approximations to the optimal design and the procedure can be repeated iteratively to check the convergence of the design variables.

The nonparametric estimates of the expected loss are displayed in Figures 3.a - 3.b and 3.c for the two-dimensional problem, and in Figures 4.a- 4.b for the one dimensional problem (i.e., f_1 = 1). As an illustration, suppose N = 689 and $n(S+)$ = 289, observed at the first phase. Then, the optimal fractions are found to be around f_1 = 52 and f_2 = 0.32 . These values are obtained by minimizing (by using the grid search) the fitted regression surface £2 with respect to f_1 and

f_2, when $(N, n(S+)) = (689, 862)$. As another example, suppose $n(S+) = 914$ and $N = 3774$. Then $f_1 = 0.32$ and $f_2 = 0.36$. In general, a grid search can be performed to create a table of the optimal fractions for different realizations of the outcomes obtained in the first phase of the experiment.

As seen in Figure 3. b. , the optimal screening size is found as $N^* = 4000$, which is the maximal value of the range of N chosen for illustration. Indeed, one would expect the optimal screening sample size for the unconstrained problem to be very large. To accommodate the budget constraint into the Monte Carlo approach, we can generate a subset of the design by imposing the equation (2.2) directly. For example, when $C/c_D = 100$ (the ratio of the total cost to the cost of the diagnostic test) and $c = 1{:}18$, the optimal N is found to be around 759. Again, once the $n(S+)$ is observed after , the optimal second stage fractions can be determined by minimizing the corresponding two-dimensional loss surface.

Figure 3.a. Nonparametric regression approximation $\pounds_2$ to expected loss $\sigma^2(n(S+),N,f)$ as a function of f_1 and f_2 for two sets of values of the pairs $(n(S+),N)$.



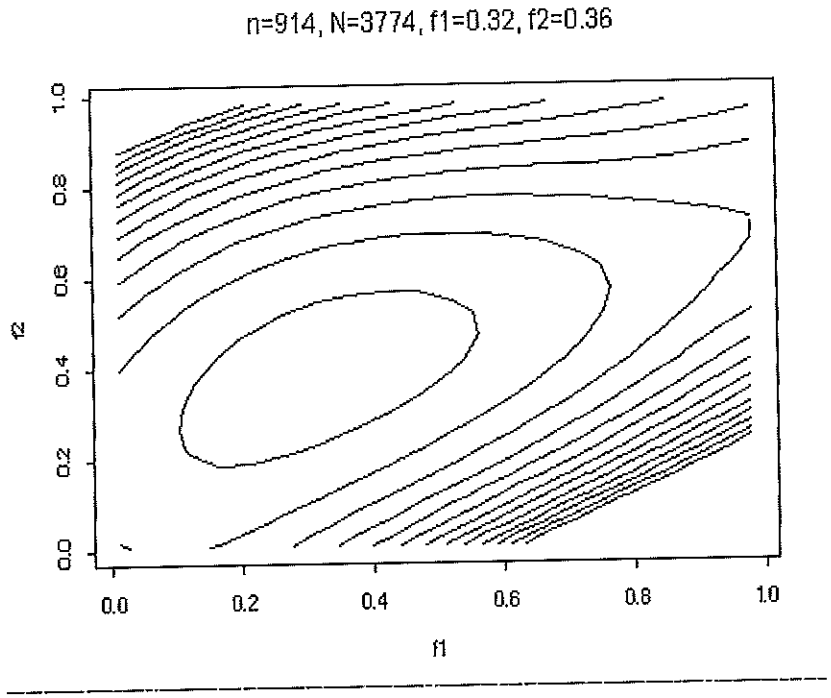


Figure 3. b. Nonparametric regression approximation $£_1$ to expected loss $\sigma^{2*}(N)$ as a function of N. The optimal screening sample size is $N^* = 4000$: unconstrained problem.

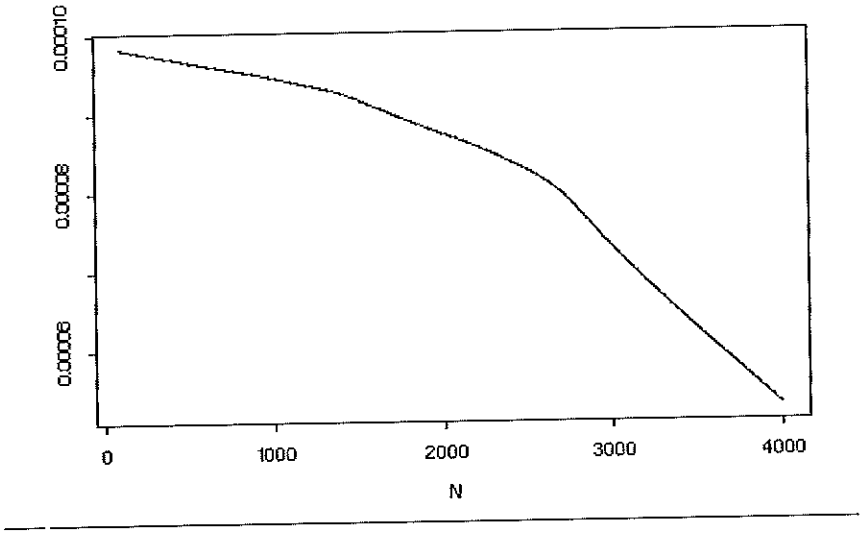


Figure 3.c. Nonparametric regression approximation $\pounds_1$ to expected loss $\sigma^{2*}(N)$ as a function of N: constrained problem when $C/c_D = 100$ and $c = 1/18$. The optimal screening sample size is $N^* = 759$.

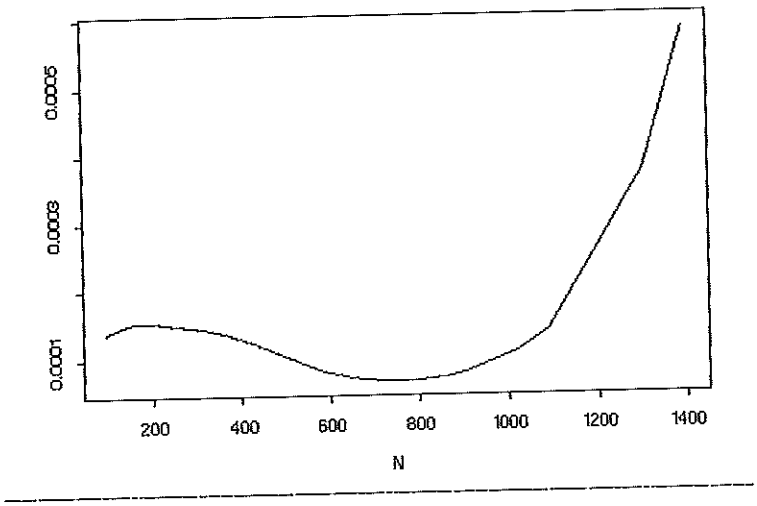


Figure 4.a. Nonparametric regression approximation $\pounds_2$ to expected loss $\sigma^2(n(S+),N,f)$ as a function of f_2 when $f_1 = 1$ for several values of the pairs $(n(S+),N)$. The minimum of each curve corresponds to the solution of the second-stage.

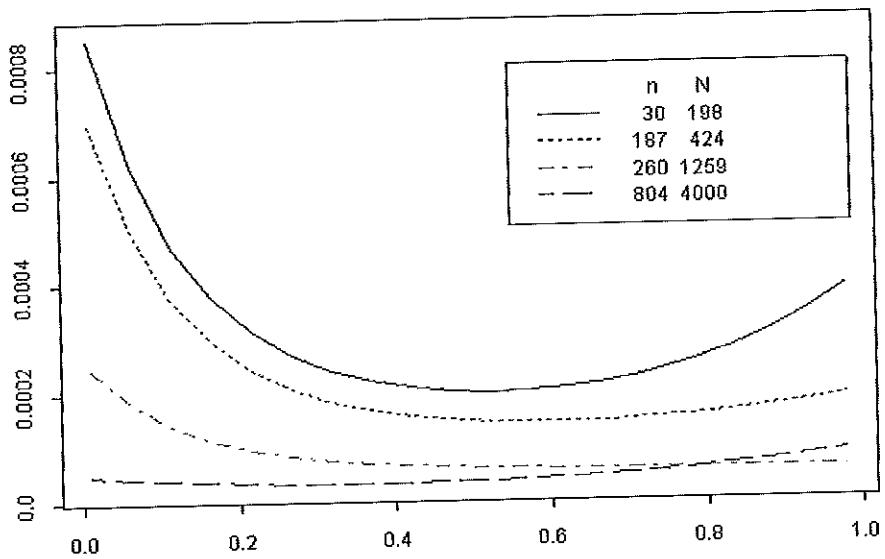


Figure 4. b. Nonparametric regression approximation $\pounds_1$ to expected loss $\sigma^{2*}(N)$ as a function of N. The optimal screening sample size is $N^* = 4000$: unconstrained.

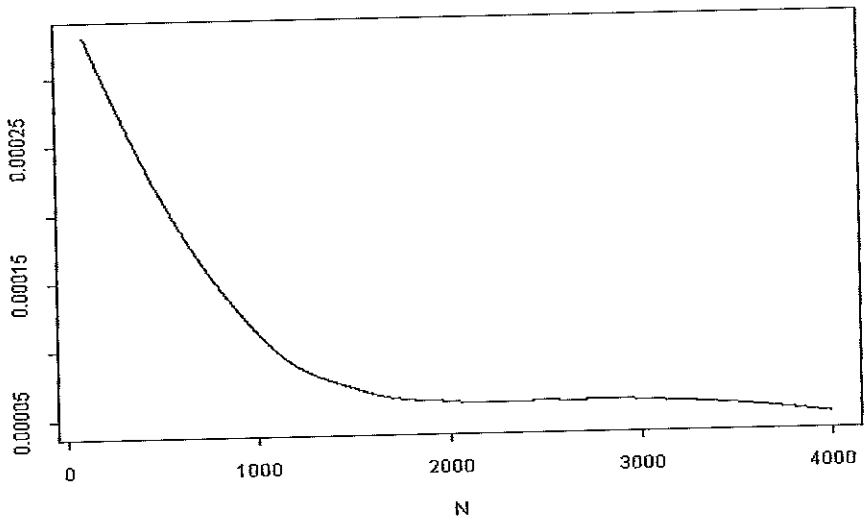
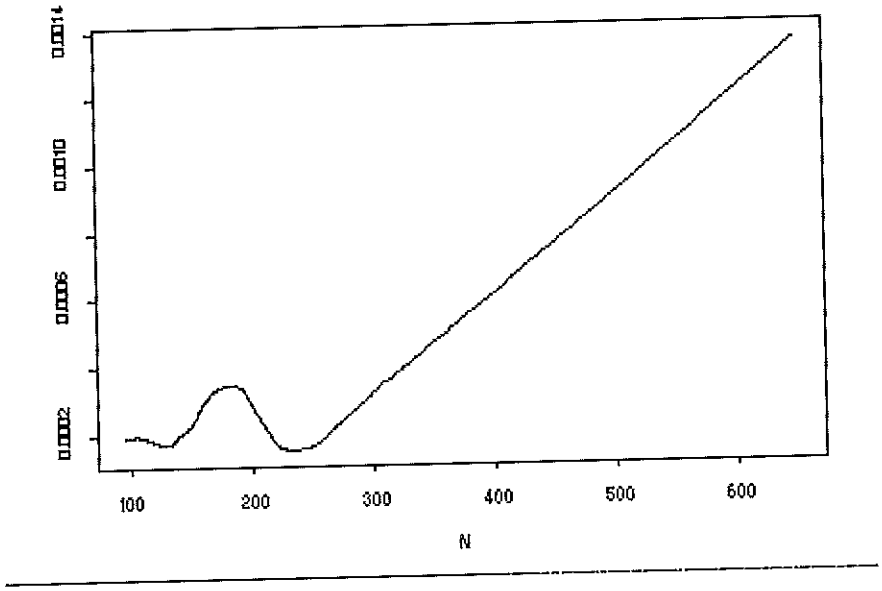


Figure 4.c. Nonparametric regression approximation $\pounds_1$ to expected loss $\sigma^{2*}(N)$ as a function of N : constrained problem when $C/c_D = 100$ and $c = 1/18$. The optimal screening sample size is $N^* = 241$.



7. Other Loss Functions

So far, we have only mentioned the optimal designs for the prevalence estimation. When there are other quantities of interest, the loss functions should be chosen appropriately to reflect the cost and the benefits expected from the experiment. For example, some times the interest is on the screening test itself and the experiment may be conducted to compare its efficiency to the existing diagnostic test. A suitable loss function for this experiment may be defined in terms of a function of the distance $\lambda_1 - \lambda_2$. Another interesting design problem in the prevalence studies is the maximization of the number of cases with the positive diagnosis while maintaining statistical efficiency. This type of mixture loss problems are discussed in Verdinelli and Kadane (1992), where both inference on the parameters and the maximization of the output of the experiment were of interest. If more complex loss functions are used the expected loss could also be difficult to evaluate. However, Monte Carlo/Smoothing methods such as the one described in Section 5 can be adopted to evaluate and optimize the criterion .

8.Conclusions

The design variables are defined as the sample sizes to be chosen at the first and the second phases of the experiment. In this particular application, the outcomes at both phases are binary, however, our approach can be generalized easily to situations where the outcomes are multinomial or continuous. Also, extensions to multi-phase experiments are feasible.

Acknowledgments

The authors would like to thank to Leon Lowenstein Foundation for their support, and to Peter Müller of ISDS for various discussions concerning the Monte Carlo approach.

REFERENCES

ANGOLD, A., COX, A., PRENDERGAST, M., RUTTER, M. and SIMONOFF, E. (1993). The Child and Adolescent Psychiatric Assessment (CAPA): Child Interview, Version 4.0. Developmental Epidemiology Program, Duke University Medical Center.

ANGOLD, A. and COSTELLO, E. J. (1993). Depressive Comorbidity in Children and Adolescents: Empirical, Theoretical, and Methodological Issues. *Am. J. Psychiatry*, **12,** 1779-1791.

BERRY, D.A. (1991). Experimental design for drug development: A Bayesian approach. *Jour. of Biopharm. Stat.*, **1,** 81-101.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models. *Ann. of Stat.*, **17,** 453-555.

CHALONER, K. and VERDINELLI, I. (1994). Bayesian experimental design: a review. Unpublished manuscript. University of Minnesota.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd edition. New York, Wiley.

DAWID, A. P. (1970). On the limiting normality of posterior distributions. *Proc. of Cambridge Phil. Soc.*, **67,** 625-633.

DEMING, W. E. (1977). An assay on screening, or on two-phase sampling. *Int. Stat. Rev.* **45, 29-37.**

GELFAND, A. E. and KUO, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika*, **78,** 657-666.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Jour. of Ame. Stat. Assoc.*, **85,** 972-985.

LINDLEY, D. V. (1985). *Making Decisions*, 2nd edition. New York, Wiley.

MAZZUCHI, T. A. and SOYER, R. (1993). A Bayes method for assessing product reliability during development testing. *IEEE. Trans. on Rel.*, 42, 503-510.

MÜLLER, P., ERKANLI, A. and WEST, M. (1995). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* (to appear).

SHROUT, P. E. and NEWMAN, S. C. (1989). Design of two-phase prevalence surveys of rare disorders. *Biometrics*, **45**, 549-555.

VERDINELLI, I. and KADANE, J.B. (1992). Bayesian designs for maximizing information and outcomes. *Jour. of Ame. Stat. Assoc.*, **87**, 510-515.

WILKS, S. S. (1962). Mathematical Statistics. New York, Wiley.