A COMPARISON OF TWO MODEL-DISCRIMINATION CRITERIA

by Duane Meeter, Walter Pirie, and William Blot

October 28, 1968

Technical Report Number 16

NASA Grant Number NGR-10-004-029

FSU Technical Report No. M146

Department of Statistics
The Florida State University
Tallahassee, Florida

A COMPARISON OF TWO MODEL-DISCRIMINATION CRITERIA

by Duane Meeter, Walter Pirie, and William Blot

0. Summary

Within the last few years there has been increased research involving a generalization of sequential analysis problems in which the experimenter is allowed to design his experiment sequentially. Two different approaches to this Sequential Design of Experiments or Model Discrimination problem have been made by Chernoff [4] and Box and Hill [3]. This paper compares the two approaches through examples. Some minor modifications of Chernoff's procedures are shown to lead to improved results.

1. Background

Chernoff's results were obtained by letting the cost of sampling approach zero, in effect allowing large samples. The choice of experiments depends on Kullback information numbers and the assumption that the current estimate for the state of nature (the true model) is the correct one. Chernoff's results were obtained for two possible terminal decisions (actions), and a finite number of states of nature and choices of experiments. They have been extended to an infinite number of states of nature by Albert [1] and to k actions and an infinite choice of experiments by Bessler [2]. We summarize Chernoff's results briefly.

Let $f(y,\theta,e)$ be the probability density of the outcome of experiment e when θ is assumed to be the state of nature. (For different θ the density f may assume different functional forms.) The information about θ_j

in experiment e when θ_i is assumed true is

$$I(\theta_i,\theta_j,e) = \int \log\left[\frac{f(y,\theta_i,e)}{f(y,\theta_j,e)}\right] f(y,\theta_i,e)dy.$$

Let w_i be the set of all θ such that sure knowledge that $\theta \in w_i$ would imply that the experimenter should take action a_i, $i = 1,\ldots,k$, and let $d(\theta)$ be the set of all w_i except the w_i containing θ. Let $\hat{\theta}_n$ be the maximum likelihood estimate of θ after n observations, and let $\tilde{\theta}_n$ be the maximum likelihood estimate restricted to the set $d(\hat{\theta}_n)$. Let

$$z_n(\theta_i,\theta_j) = \log[f(y_n,\theta_i,e^{(n)})/f(y_n,\theta_j,e^{(n)})],$$

where $e^{(n)}$ is the experiment selected for the n^{th} observation, y_n the outcome of the n^{th} experiment. Chernoff's procedure A is as follows:

Stop sampling at the n^{th} observation and select the hypothesis of $\hat{\theta}_n$ if $\sum_{i=1}^{n} z_i(\hat{\theta}_n,\tilde{\theta}_n) > -\log c$, where c is the cost of an observation. Otherwise, let $e^{(n+1)} = e(\hat{\theta}_n)$, where $e(\theta_0)$ is the maximin strategy of the second player (Experimenter) in a game with Nature in which the payoff to the Experimenter when Nature is using strategy ϕ, the true state of Nature is θ_0, and the Experimenter is using strategy e, is

$$I(\theta_0,\phi,e).$$

The value of the game is

$$I(\theta_0) = \inf_{\phi \in d(\theta_0)} I(\theta_0,\phi,e(\theta_0)).$$

We are allowing the possibility that both Nature and the Experimenter may use randomized strategies. Note that the definition of $e^{(n+1)}$ says that

at each stage we will assume that the hypothesis with the highest likelihood is the true state of nature, and use maximin strategy developed under that assumption. The intention is, of course, to design experiments such that after a number of observations have been taken $\hat{\theta}_n = \theta_0$ with high probability.

Designing sequential experiments using information numbers makes intuitive sense. The information $I(\theta_1,\theta_2,e)$ is the expected value of the amount by which the logarithm of the likelihood ratio for testing θ_1 vs. θ_2 will increase if θ_1 is true, and experiment e is used. Large information numbers imply the stopping criterion (-log c) will be achieved sooner. Simply maximizing information numbers may not be wise, however, since Nature can choose a different strategy (from among alternative hypotheses). In order to accept hypothesis θ_1 we must reject all competing hypotheses. An experiment good for rejecting one alternative hypothesis may be poor for another. Theorem 2 among the following results justifies (as least as $c \to 0$) procedure A. Under mild restrictions, some of which can be relaxed (see Bessler [2], e.g.) Chernoff has shown for procedure A:

Lemma 1. Let the stopping rule be disregarded. Let T be the smallest integer such that $\hat{\theta}_n = \theta_0$ for $n \geq T$. Then there exist b_1 and $b_2 > 0$ such that $Pr\{T>n\} \leq b_1 e^{-b_2 n}$.

Lemma 2. The expected sample size satisfies, as $c \to 0$,

$$E(N) \leq -[1 + o(1)]\log c/I(\theta_0).$$

Lemma 3. The probability of error (i.e. of accepting the hypothesis that $\theta_0 \in d(\theta_0)$) is $\alpha = O(c)$.

Theorem 1. The risk function $R(\theta)$ satisfies $R(\theta) \leq -[1 + o(1)]c \log c/I(\theta)$.

Theorem 2. Any procedure for which $I(\theta) > 0$ and $R(\theta) = O(-c \log c)$ for all θ satisfies

$$R(\theta) \geq -[1 + o(1)]c \log c/I(\theta) \quad \text{for all } \theta.$$

Chernoff has termed procedure A asymptotically optimal in the sense that if another procedure has risk substantially smaller than procedure A for any θ then its risk will be of a greater order of magnitude for some other value of θ, this argument applying as $c \to o$. However in view of the statement of the Theorem a better term might be asymptotically admissible.

The approach of Box and Hill [4] begins, on the other hand, with k hypotheses or models symbolized by $\theta_1,\ldots,\theta_k$ and the concept of entropy, measured by $-\sum_{i=1}^{k} p_i \log p_i$, where p_i is the probability that hypothesis θ_i is true. Maximum entropy occurs when $p_i = 1/k$, $i = 1,\ldots,k$; the information about the hypotheses is at a minimum. Minimum entropy (greatest information) occurs as the probability of one of the hypotheses approaches one. Before the $(n+1)^{st}$ observation is taken the entropy is $-\sum_{i=1}^{k} p_{ni} \log p_{ni}$, where p_{ni} is the posterior probability that θ_i is true after n observations have been made. Box and Hill seek to maximize the difference

$$R = \text{entropy at input} - \text{expected entropy at output}$$

where input and output refer to before and after the taking of observation n+1. Instead of working with R which involves a difficult integral, they use an upper bound D for R which reduces to

$$D = \sum_{i=1}^{k} \sum_{j>i}^{k} p_{ni}p_{nj}\left[\int f_i \log(f_i/f_j)dy + \int f_j \log(f_j/f_i)dy\right]$$

where f_i stands for $f(y,\theta_i,e)$, the density of y under hypothesis θ_i and experiment e. We immediately recognize the expression in square brackets as $I(\theta_i,\theta_j,e) + I(\theta_j,\theta_i,e)$ which is Kullback's [6] measure of divergence. Thus D is a weighted measure of divergence for discriminating between all possible pairings of the n hypotheses, the weights being the products of the posterior probabilities of the hypotheses after n observations. This criterion can be expressed in another way. If we knew that hypothesis θ_i was true, and wanted to maximize information about θ_j, $j\neq i$, in the absence of any knowledge except the magnitude of the p_{nj} about which alternative θ_j would be most difficult to discriminate against we might try to maximize

$$\sum_{j\neq i} p_{nj} I(\theta_i,\theta_j,e).$$

However since we assumed that θ_i was true when in fact its posterior probability is p_{ni}, it is natural to multiply the above expression by p_{ni} and sum over $i = 1,\ldots,k$, yielding D.

We note three things about this criterion. One is that it seems strange to maximize an upper bound to the expected change in entropy rather than a lower bound. (The bound is based on the inequality $\sum_j p_{nj} \log f_j \leq \log(\sum_j p_{nj} f_j)$ so that it will tend to sharpen as one hypothesis attains high probability.) The other is that this criterion may tend to pick out experiments that yield high information even though they correspond to hypotheses with low probabilities. That is, it may happen that the behavior of D may be dominated by $p_{ni}p_{ni'}[I(\theta_i,\theta_{i'},e) + I(\theta_{i'},\theta_i,e)]$ where $p_{ni}p_{ni'}$ is relatively small and the expression in square brackets is relatively large. Thus we may be led to experiments which yield large amounts of information about hypotheses which are already close to being

ruled out by the previous data. Finally, if k=2, $D = p_{n1}p_{n2}[I(\theta_1,\theta_2,e) + I(\theta_2,\theta_1,e)]$, so that maximizing D results in the same choice of experiments whether p_{n1} or p_{n2} is near one, whereas maximizing $I(\theta_1,\theta_2,e)$ may yield quite different choices of experiments than maximizing $I(\theta_2,\theta_1,e)$.

On the other hand, the Chernoff procedure uses a maximin strategy appropriate if the maximum likelihood estimate is the state of nature. This may lead to "initial bungling", since, in Chernoff's words, "At first it is desirable to apply experiments which are informative for a broad range of parameter values. Maximizing the Kullback-Liebler information number may give experiments which are efficient only when θ is close to the estimated value." Another question is, how small does c have for the asymptotic properties to assert themselves? At this point, it seems appropriate to examine these two procedures by means of examples. The first two examples are from Bessler [2].

2. Examples

Example 1. Choosing the unusual coin out of a set of k coins.

Let the probability that say coin number 1 yields heads be γp, whereas for the other k-1 coins it is p, where γ and p are known and are not equal to 1. Let θ_i, i = 1,...,k, be the hypothesis that coin i is the "odd" coin. The k possible experiments are e_i, i = 1,...,k, to take an observation from coin i, and the k actions are to accept the hypothesis $\theta=\theta_i$. Letting $I(\theta_i,\theta_j,e_\ell) = I_{ij}(e_\ell)$ denote the information obtained by e_ℓ about θ_j when θ_i is assumed true,

$$I_{ij}(e_\ell) = a = \ln\{\gamma^{\gamma p}[(1-\gamma p)/(1-p)]^{1-\gamma p}\} \quad \text{if } \ell=i$$
$$= b = \ln\{\gamma^{-p}[(1-p)/(1-\gamma p)]^{1-p}\} \quad \text{if } \ell=j$$
$$= 0 \quad \text{otherwise.}$$

Without loss of generality, it can be assumed that after n observations have been taken the estimate $\hat{\theta}_n$ of the hypotheses θ_i is equal to θ_1. The same argument holds with a relabeling of e's and θ's if $\hat{\theta} \neq \theta_1$. The payoff matrix is shown below.

		Strategy of Player II (Experimenter)							
		e_1	e_2	e_3	.	.	.	.	e_k
Strategy of Player I (Nature)	θ_2	a	b	0	.	.	.	.	0
	θ_3	a	0	b	0	.	.	.	0
	θ_4	a	0	0	b				.
	.	.	.	.		.			.
	.	.	.	.			.		.
	.	.	.	.				.	0
	θ_k	a	0	0	.	.	.	0	b

If Nature chooses strategy θ_i, i = 2,...,k with probability 1/k-1 then the Experimenter's winnings are limited to

$$\max (a, b/k-1).$$

The Experimenter's maximin strategy depends on a, b, and k.

Case I. If $a \geq b/k-1$, then the Experimenter's maximin strategy is to choose e_1 with probability 1 and the value of the game is

$$I(\theta_1) = a.$$

Case II. If $a < b/k-1$, the Experimenter's maximin strategy is to choose e_i, $i = 2,\ldots,k$ with probability $1/k-1$ and the value of the game is

$$I(\theta_1) = b/k-1.$$

For this example it is also possible to predict what the Box-Hill procedure will do. For experiment e_ℓ

$$I_{ij}(e_\ell) + I_{ji}(e_\ell) = a+b \quad \text{if } \ell = i \text{ or } j$$
$$= 0 \quad \text{otherwise,}$$

so that

$$D(e_\ell) = (a+b)p_{n\ell} \sum_{j\neq\ell} p_{nj}$$
$$= (a+b)p_{n\ell}(1-p_{n\ell}).$$

Maximizing D is equivalent to minimizing $|p_{n\ell} - 1/2|$ which, if $k>2$, is equivalent to the rule "choose the coin with the largest likelihood (or posterior probability)"*. If $k=2$ then $D(e_1) = D(e_2)$ so no decision is possible. In this case, we made an arbitrary decision to randomize between e_1 and e_2. This procedure yields $(a+b)/2$ units of information whereas the Chernoff procedure should yield $\max(a,b)$ units of information. For $k>2$, if $a \geq b/k-1$ the Box-Hill and Chernoff procedures are identical whereas if $a < b/k-1$, Nature can plan any strategy and the Experimenter will be limited to a payoff of a units. Following Bessler, we might say therefore that the (asymptotic) efficiency of the Box-Hill procedure relative to Chernoff's is

*In each of the three examples in this paper, we have assumed equal prior probabilities $1/k$ for each state of nature, θ_i, $i = 1,\ldots,k$, where required.

$$E = 1 \qquad \text{if } k>2, \quad a \geq b/k-1$$

$$= a/(b/k-1) \qquad \text{if } k>2, \quad a < b/k-1$$

$$= (a+b)/2 \max(a,b) \qquad \text{if } k=2,$$

in the sense that (referring to Lemma 2 and Theorem 1) we might expect that, for sufficiently small cost c of a single observation, the above ratio would approach the ratio of risks or of average sample numbers for the two procedures. Values of a, b, and k such that the Box-Hill procedure would be expected to be at a disadvantage occur only when γ is not too close to 1 and k is small. For example, see Table 1. By symmetry, the efficiencies hold also for $p' = 1-p$, $\gamma' p' = 1-\gamma p$.

Table 1. Predicted Efficiency of Box-Hill Procedure for Case II.

	k=2						k=3				
γ	.9	.9	.5	.5	.1	.1	γ	.05	.05	.01	.01
p	.5	.1	.5	.1	.5	.1	p	.5	.1	.5	.1
E	.998	.98	.95	.90	.80	.75	E	.99	.81	.68	.54

In Case II, it is possible to modify Chernoff's procedure yielding an improved procedure asymptotically equivalent to A which we will call procedure $\tilde{A}$: choose that coin corresponding to $\tilde{\theta}$, the maximum likelihood hypothesis among the alternatives to $\hat{\theta}$. Clearly, for large samples, $\tilde{\theta}$ will be θ_i, $i = 2,\ldots,k$ with approximately equal frequency so that its asymptotic properties should be the same as that of A. If b>a, this is essentially the same procedure described in Chernoff [4] as: Player 2 (Experimenter) selects e for observation n+1 as though Nature is going to use that strategy which repeated n times would have been most effective against the combination of the past choices of the Experimenter.

Table 2. Results of a Simulation for Case II, Example 1.

γ=.1 p=.3 k=2 I=.462 -log c/I=9.9

Procedure	$\bar{N}$	Standard Deviation	Range of N	Empirical Efficiency	Predicted Efficiency
A	13.8 ± 1.0	9.6	2-51	-	1.0
Box-Hill	14.5 ± 1.0	9.7	2-57	.95	.77

γ=.5 p=.5 k=2 I=.144 -log c/I=31.9

Procedure	$\bar{N}$	Standard Deviation	Range of N	Empirical Efficiency	Predicted Efficiency
A	34.7 ± 2.3	22.8	10-111	-	1.0
Box-Hill	35.6 ± 2.2	22.5	10-120	.97	.95

γ=.05 p=.05 k=3 I=.052 -log c/I=88.9

Procedure	$\bar{N}$	Standard Deviation	Range of N	Empirical Efficiency	Predicted Efficiency
A	201 ± 13.8	138.2	28-614	-	1.0
Box-Hill	140 ± 5.0	50.3	95-322	1.44	.80
$\tilde{A}$	115 ± 6.9	68.8	18-386	1.75	1.0

Procedure $\tilde{A}$ would not be expected to perform well in Case I, since (asymptotically) selecting θ_i, $i = 2,\ldots,k$ with equal frequency would yield $b/k-1$ units of information against Nature's best strategy, while selecting the coin corresponding to $\hat{\theta}$ should yield at least a units of information against any strategy of Nature. In Table 2 we have listed some results of simulations of procedures A, $\tilde{A}$, and that of Box and Hill for combinations of γ, p and k leading to Case II. In each case, the criterion for termination was $\sum_{i=1}^{n} z_i(\hat{\theta}_n,\tilde{\theta}_n) \geq -\log c$, with $c = 1/99$, i.e., the likelihood of $\hat{\theta}_n$ relative to $\tilde{\theta}_n$ had to exceed 99/1. The quantity -log c/I is the approximation to the average sample number for procedure A given by Lemma 2. One hundred simulation runs were made for each row of the table; beside each $\bar{N}$ is an estimate of its standard error. Interest-

ingly, in none of the seven hundred runs represented in this table was the true hypothesis rejected.

We immediately notice several things about this table. First, the Box-Hill procedure is probably not as inefficient as Table 1 would suggest and in fact is quite superior to A for the k=3 example. Procedure $\tilde{A}$ on the other hand is an even greater improvement and the ratio of its empirical Average Sample Number (ASN) to that of the Box-Hill procedure (115/140 = .82) is very close to the efficiency predicted for Box-Hill relative to Chernoff's procedure A. There is some suggestion that the approximation -log c/I to the ASN is not as good for larger values of k. To explore this further, another simulation was made, this time for Case I ($a > b/k-1$), with $\gamma=.9$, $p=.5$, so that $a=.005008$, $b=.0050025$. Fifty runs were made for k=3, 6, and 12. The Box-Hill and Chernoff procedures give identical results for Case I, but we compared their procedures instead to the "no design" procedure which merely takes k observations, one on each coin, between each likelihood ratio test. (What was actually done in the simulation was to randomly choose a coin for each observation and test after each observation, which should give fairly similar results.) It is easy to see that Nature can hold this strategy to a gain of (a+b)/k units of information per observation while the maximin strategy can yield a units of information. The results are shown in Table 3.

It is evident that the approximation to the ASN breaks down with increasing k. This is because an increasingly greater part of the sample is taken up with trying each of the coins a sufficient number of times until the odd coin establishes a considerable lead. Once this happens the remainder of the observations are taken on a single coin.

Table 3. Results of a Simulation for Case I, Example 1.

γ=.9 p=.5 I=.005008 c=1/99 -log c/I=918

	Procedure	$\bar{N}$	Standard Deviation	Range of N	Empirical Efficiency	Predicted Efficiency
k=3	A	1,518±135	953	234-5,086	-	1.
	"No Design"	2,695±125	881	1,462-5,416	.563	.67
k=6	A	2,713±180	1,273	677-5,953	-	1.
	"No Design"	8,896±304	2,150	5,463-14,725	.305	.333
k=12	A	6,060±508	3,589	1,360-19,649	-	1.
	"No Design"	24,206±771	5,451	16,112-38,303	.25	.17

The predicted efficiencies are fairly accurate, and do not say much in favor of non-design sequential experiments when a choice of experiments is available. As in the previous simulation, none of the 300 runs resulted in accepting the wrong hypothesis. We now take up an example involving normal populations.

Example 2. Identifying three normal populations with known means and common known variance.

Suppose that π_1, π_2, and π_3 are three normal populations with known means (μ_1,μ_2,μ_3) respectively and known common variance σ^2. Suppose $\mu_1 > \mu_2 > \mu_3$. The following six permutations of (μ_1,μ_2,μ_3) represent the six possible hypotheses that the experimenter can accept:

$$\theta_1 = (\mu_1,\mu_2,\mu_3)$$
$$\theta_2 = (\mu_1,\mu_3,\mu_2)$$
$$\theta_3 = (\mu_2,\mu_1,\mu_3)$$
$$\theta_4 = (\mu_2,\mu_3,\mu_1)$$
$$\theta_5 = (\mu_3,\mu_1,\mu_2)$$
$$\theta_6 = (\mu_3,\mu_2,\mu_1).$$

The three possible experiments e_i are to select an observation from population π_i, $i = 1,2,3$. The maximin strategies computed by Bessler depend on $q=b/a$, where

$$a = \frac{\mu_1-\mu_2}{\sigma}$$

$$b = \frac{\mu_2-\mu_3}{\sigma} .$$

The information in observations from normal populations about hypotheses say H' and H'' differing only in the specification of the mean is

$$I = \frac{(\mu'-\mu'')^2}{2\sigma^2} ,$$

where μ' and μ'' are the means specified in the two hypotheses. The payoff matrix for the Experimenter's game with Nature is given below, assuming without loss of generality that θ_1 is the hypothesis of $\hat{\theta}_n$. (If $\hat{\theta}_n = \theta_j$, $j\neq 1$, then we just interchange the strategies for playing e_1, e_2, and e_3 to correspond to the interchanges of (μ_1,μ_2,μ_3) in going from θ_1 to θ_j.)

Experimenter's Strategy

Nature's Strategy	e_1	e_2	e_3
θ_2	0	$b^2/2$	$b^2/2$
θ_3	$a^2/2$	$a^2/2$	0
θ_4	$a^2/2$	$b^2/2$	$(a+b)^2/2$
θ_5	$(a+b)^2/2$	$a^2/2$	$b^2/2$
θ_6	$(a+b)^2/2$	0	$(a+b)^2/2$

Let $\lambda = (\lambda_1,\lambda_2,\lambda_3)$ be the probabilities for a randomized strategy (for the Experimenter) over e_1, e_2 and e_3. There are three cases to the maximin strategy. Let $u = (1+q)^2$, $s = 1/(1+q^2)$

Case I. $u/(u+1) < q^2 < u/(u-1)$ $(.883 < q < 1.132)$

$$\lambda = [1 - 2su/(u-s+1),\ u - (1-s)/(u-s+1),\ 1 - 2u(1-s)/(u-s+1)]$$

$$I = \frac{a^2}{2}\left[\frac{2}{1 + 1/q^2 + 1/(1+q)^2}\right] .$$

Case II. $q^2 \geq u/(u-1)$ $(q \geq 1.132)$

$$\lambda = (\lambda_1, 1-\lambda_1, 0)$$

where

$$1/(1+q)^2 \leq \lambda_1 \leq 1 - (1/q)^2;$$

$$I = a^2/2.$$

Case III. $q^2 \leq u/(u+1)$ $(q \leq .883)$

$$\lambda = (0, 1-\lambda_3, \lambda_3)$$

where

$$q^2/(1+q)^2 \leq \lambda_3 \leq 1-q^2.$$

Table 4 shows the results of some simulation experiments for Case II. Two hundred runs were made for each set of experimental conditions. Using the criterion of a 99/1 likelihood ratio for the hypotheses with the largest likelihoods, the true hypothesis was rejected 34 times in the 3034 runs. (Runs accepting the wrong hypothesis were not included in the figures in the table.) Procedure A was used in two ways. Procedure A_1 used $\lambda_1 = 1 - 1/q^2$, the upper bound. This can be seen to be definitely inferior to the Box-Hill procedure for the two cases tried, $q = \sqrt{2}$ and $q=4$. Now Nature's minimax strategy for Case II is the pure strategy θ_3. However, occasionally Nature ignores her own best interests and plays θ_2 or θ_6. Strategies θ_4 and θ_5 are not often used since they are both dominated by θ_3. Procedure A_2 selects λ_1 to equalize the information

Table 4. Results of a Simulation for Case II, Example 2.

$b=\sqrt{2}/2$ $a=1/2$ $q=\sqrt{2}$ $I=.125$ $c=1/99$ $-\log c/I=36.8$

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A_1	2	60.2 ± 1.8	25.8	22-170	(.50,.50,0)	(.48,.44,.08)
A_2	2	49.3 ± 1.3	19.0	22-131	(.26,.74,0)	(.31,.59,.10)
Box-Hill	3	59.9 ± 1.6	22.5	23-149	-	(.27,.59,.14)

$b=\sqrt{2}/4$ $a=1/4$ $q=\sqrt{2}$ $I=.03125$ $c=1/99$ $-\log c/I=147$

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A_1	2	229 ± 6.7	94.6	89-669	(.50,.50,0)	(.48,.44,.08)
A_2	3	190 ± 6.4	90.9	81-705	(.26,.74,0)	(.32,.59,.09)
Box-Hill	2	191 ± 6.8	96.7	85-674	-	(.28,.62,.10)

$b=\sqrt{2}/8$ $a=1/8$ $q=\sqrt{2}$ $I=.0078125$ $c=1/99$ $-\log c/I=588$

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A_1	3	853 ± 24.5	346	373-2298	(.50,.50,0)	(.49,.45,.06)
A_2	2	794 ± 25.0	353	357-2001	(.26,.74,0)	(.32,.60,.08)
Box-Hill	3	746 ± 23.5	333	342-2219	-	(.26,.61,.13)

$b=1$ $a=1/4$ $q=4$ $I=.03125$ $c=1/99$ $-\log c/I=147$

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A_1	1	188 ± 6.5	92	53-617	(15/16,1/16,0)	(.77,.21,.02)
A_2	4	153 ± 6.6	93	41-525	(.39,.61,0)	(.43,.55,.02)
Box-Hill	3	156 ± 7.1	100	43-606	-	(.42,.54,.04)

$b=1/2$ $a=1/8$ $q=4$ $I=.0078125$ $c=1/99$ $-\log c/I=588$

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A_1	4	727 ± 25.8	365	147-2246	(15/16,1/16,0)	(.80,.19,.01)
A_2	0	569 ± 20.5	290	142-1565	(.39,.61,0	(.42,.56,.02)
Box-Hill	1	615 ± 25.5	361	161-1839	-	(.41,.55,.04)

coming from e_1 when Nature uses θ_6 with the information coming from e_2 when Nature uses θ_2, namely $\lambda_1 = 1/(1+(a+b)^2/b^2)$. It can be seen from the table that A_2 is as good as the Box-Hill procedure, while still being maximin. The "empirical mixed strategy" listed in the table is the proportion of times each of the three populations was selected over 200 simulations. The actual frequencies varied from run to run, particularly for the Box-Hill method, depending on which hypotheses had the highest probabilities.

A final run was made simulating the Case I situation, with q=1. The results are shown in Table 5.

Table 5. Results of a Simulation for Case I, Example 2.

b=1/2 a=1/2 q=1 I=1/9 c=1/99 -log c/I=41.4

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A	3	64.9 ± 1.8	25.6	29-161	(.11,.78,.11)	(.21,.61,.18)
Box-Hill	1	63.7 ± 1.7	23.6	29-163	-	(.20,.61,.19)

b=1/4 a=1/4 q=1 I=1/36 c=1/99 -log c/I=165

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A	1	266 ± 8.1	114	100-778	(.11,.78,.11)	(.21,.58,.21)
Box-Hill	2	230 ± 5.5	77	117-507	-	(.18,.62,.20)

b=1/8 a=1/8 q=1 I=1/144 c=1/99 -log c/I=662

Procedure	Errors	$\bar{N}$	Std. Dev.	Range of N	Theoretical Mixed Strategy $(\lambda_1,\lambda_2,\lambda_3)$	Empirical Mixed Strategy $(\hat{\lambda}_1,\hat{\lambda}_2,\hat{\lambda}_3)$
A	5	1084 ± 31.5	445	434-2916	(.11,.78,.11)	(.22,.57,.21)
Box-Hill	0	918 ± 23.9	338	473-2004	-	(.21,.62,.17)

For Case I, q=1, the maximin strategy is $\lambda = (1/9,7/9,1/9)$. The true hypothesis was rejected in 12 of 1212 simulations in this table. The Box-Hill procedure is clearly superior. Note that the asymptotic optimality of Chernoff's procedure A is proved for sufficiently small c, not for large n, so that we cannot necessarily expect procedure A to exhibit its optimality even for these large samples without decreasing c from the 1/99 used in the examples presented.

Example 3. Distinguishing an exponential from polynomial models.

Consider the following regression hypotheses:

$$\theta_1: \quad E(y) = \beta_{11}x$$
$$\theta_2: \quad E(y) = \beta_{21} + \beta_{22}x$$
$$\theta_3: \quad E(y) = \beta_{31} + \beta_{32}x + \beta_{33}x^2$$
$$\theta_4: \quad E(y) = \beta_{41}x + \beta_{42}x^2$$
$$\theta_5: \quad E(y) = \beta_{51} \exp(\beta_{52}x)$$

Here our hypotheses are not points but actually sets in a multidimensional parameter space. A more precise description would state all of the hypotheses as special cases of the model $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 \exp(\beta_4 x)$. We add the assumption that the observations y are normally distributed about E(y) for the true model with mean zero and variance σ^2. Box and Hill have derived an expression for the probability density of a future observation y_{n+1} given n previous observations and locally uniform prior distributions for the unknown parameters, and assuming that non-linear models are approximately linear near the maximum likelihood estimates of their parameters. The density of a future observation y_{n+1} under hypothesis θ_i is normal with mean $\hat{y}_n^{(i)}$ and variance $\sigma^2 + \sigma_i^2$, where $\hat{y}_n^{(i)}$ is the predicted

value of y under hypothesis θ_i and $\sigma_i^2 = \mathrm{Var}(\hat{y}_n^{(i)})$. From this were calculated the information numbers and the criterion D of Box and Hill. The information numbers are

$$I(\theta_i,\theta_j,e) = \log[(\sigma^2+\sigma_j^2)/(\sigma^2+\sigma_i^2)] + [\sigma^2 + \sigma_i^2 + (\hat{y}_n^{(i)}-\hat{y}_n^{(j)})^2]/2(\sigma^2+\sigma_j^2)^{-1/2}.$$

The choice of the experiment was made by choosing a level for the independent variable from the set 0,1/4,2/4,...,15/4,16/4 at each stage; this affects the information numbers through the predicted values $\hat{y}^{(i)}$ and $\hat{y}^{(j)}$ and their variances. To use procedure A, information numbers were calculated assuming that the hypothesis currently with highest likelihood is the true state of nature, and assembling a 4×17 payoff matrix. Since it was deemed impossible to compute the Experimenter's maximin strategy analytically, linear programming methods were used at each stage to solve the matrix game and compute the strategy. Parameters in each of the models were re-estimated by least squares after each stage. The data were generated by adding normal (0,1) pseudorandom numbers to the model of hypothesis θ_5, the exponential model, with $\beta_{51} = 1$ and $\beta_{52} = .75$. In the first set of experiments reported below in Table 6 we took a preliminary sample of five observations at x = 0, 1, 2, 3, 4 to estimate the parameters in the model. One hundred simulations were run for each row of the table. Since there were a relatively large number of errors, the results for runs which rejected the true hypothesis were analyzed separately. One modification of both procedures that was tried was to reject any hypothesis or model as soon as its likelihood relative to that of the hypothesis of $\hat{\theta}$ was less than 10^{-4}. Experiments which used this modification are indicated in the third column of the table.

Table 6. Discrimination of an Exponential with Five Initial Observations

Procedure	Errors	Drop poor models?	c	Runs accepting θ_5			Runs rejecting θ_5		
				$\bar{N}$	Std. Dev.	Range of N	$\bar{N}$	Std. Dev.	Range of N
A	24	Yes	1/99	5.96±.48	4.15	2-20	4.92± .51	2.50	2-11
Box-Hill	15	Yes	1/99	5.00±.29	2.71	2-15	4.93± .59	2.28	2-9
A	29	No	1/99	6.51±.36	3.04	1-15	5.45± .48	2.60	2-11
Box-Hill	15	No	1/99	4.74±.24	2.25	2-13	4.40± .40	1.55	2-7
A	17	Yes	1/9999	11.29±.67	6.12	3-41	9.53± .90	3.73	5-18
Box-Hill	9	Yes	1/9999	7.96±.44	4.24	3-26	8.33± .85	2.55	4-13
A	19	No	1/9999	9.93±.59	5.33	3-30	7.84±1.05	4.60	2-22
Box-Hill	15	No	1/9999	7.11±.38	3.47	3-24	6.20± .66	2.54	3-11

The most striking feature of this table is that the proportion of errors or rejections of the exponential model is, unlike the first two examples, not approximately equal to the likelihood ratio criterion c. (In an earlier unreported calculation, with hypothesis θ_5 omitted and data generated from hypothesis θ_3, neither procedure made any errors in 100 simulations.) Thus the ability to distinguish an exponential from quadratic polynomials provides a severe test for these two procedures. A second experiment was run in which the initial sample was 20 observations, obtained by replicating the design x = 0, 1, 2, 3, 4 four times.

These two tables yield the following conclusions about this example:

a. The Box-Hill procedure performs consistently better than Chernoff's procedure A in error rate, ASN, and range of N.

b. As expected, increasing the size of the initial "non-designed" sample decreases the error rate (but prohibiting termination of the experiment before twenty observations had been taken would also accomplish this).

Table 7. Discrimination of an Exponential with Twenty Initial Observations

Procedure	Errors	Drop poor models?	c	Runs accepting θ_5			Runs rejecting θ_5		
				$\bar{N}$	Std. Dev.	Range of N	$\bar{N}$	Std. Dev.	Range of N
A	5	Yes	1/99	1.58±.28	2.72	0-21	2.60±.51	1.14	1-4
Box-Hill	3	Yes	1/99	1.61±.21	2.07	0-10	2.67±.33	.58	2-3
A	7	No	1/99	1.76±.22	2.09	0-14	2.14±.40	1.07	1-4
Box-Hill	5	No	1/99	1.67±.19	1.88	0-9	3.4 ±.81	1.81	1-6
A	3	Yes	1/9999	3.80±.40	3.90	1-23	6.0±1.53	2.64	3-8
Box-Hill	0	Yes	1/9999	2.77±.22	2.22	0-12	-	-	-
A	8	No	1/9999	3.92±.43	4.14	1-21	4.25±.56	1.58	2-6
Box-Hill	2	No	1/9999	2.90±.22	2.13	0-13	6.0±2.0	2.83	4-8

c. Dropping out poor models has a beneficial effect on the ASN for runs accepting θ_5, while it seems to increase the maximum of N for these runs. It has a beneficial effect on the error rate.

d. With five initial observations, runs rejecting θ_5 have a higher ASN; with twenty initial observations, the situation is reversed.

3. Conclusion

Two procedures for sequentially designing experiments to select the correct model or state of nature have been compared. Chernoff's procedure A is asymptotically optimal for experiments with sufficiently small costs of experimentation, but as we have seen this assumption may not be satisfied in practical problems involving even large samples. In some cases it is possible to modify procedure A to achieve much more efficient "large c" performance without affecting the maximin character of the strategy.

However unless the maximin procedure can be computed analytically for a given problem, the maximin procedure involves difficult and time-consuming computations.

The Box-Hill procedure, on the other hand, performs well on these examples probably because it avoids over-committing to a single strategy until one hypothesis is clearly favored. However, it has no known optimal properties and can perform poorly as in Case II of Example 1. Its good performance on other problems suggests that future progress in this area depends on the development of procedures that use initial samples to discover the most promising hypotheses and then make the transition to some type of maximin procedure. This idea is like that of Kiefer and Sachs [6], who prove asymptotic optimality for design procedures which take an initial sample to discover the state of nature and then design one final large sequential experiment. The size of the initial sample is specified to tend to infinity in such a way that its proportion of the total sample tends to zero. However little is known about how to design this preliminary sample efficiently. Results have been obtained for simple special cases, but not without difficulty. Also, Lorden [7] has obtained bounds on the increase relative to Bayes tests in risk (averaged over the prior probabilities of the hypotheses) of some asymptotically optimal sequential tests so that one can compute the loss in efficiency involved in designing experiments optimal for $c \to 0$, when in fact c is not too small.

References

[1] Albert, A. E. "The sequential design of experiments for infinitely many states of nature," Ann. Math. Statist., 32, 774-799 (1961).

[2] Bessler, S. "Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments. Part I - Theory. Part II - Applications," Technical Report Nos. 55 and 56, Applied Mathematics and Statistics Laboratories, Stanford University (1960).

[3] Box, G. E. P. and Hill, W. J. "Discrimination among mechanistic models," Technometrics 9, 57-71 (1967).

[4] Chernoff, H. "Sequential design of experiments," Ann. Math. Statist., 30, 755-770 (1959).

[5] Kullback, S. Information Theory and Statistics, John Wiley & Sons, New York (1959).

[6] Kiefer, J. and Sachs, J. "Asymptotically optimum sequential inference and design," Ann. Math. Statist., 34, 705-750 (1963).

[7] Lorden, G. "Integrated risk of asymptotically Bayes sequential tests," Ann. Math. Statist., 38, 1399-1422 (1967).