AD-752 077

EXPERIMENTS FOR NON-LINEAR FUNCTIONS

W. G. Cochran

Harvard University

Prepared for:

Office of Naval Research

3 November 1972

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY

EXPERIMENTS FOR NON-LINEAR FUNCTIONS

by

W. G. Cochran

Technical Report No. 39

November 3, 1972

Contract N00014-67A-0298-0017, NR-042-097

This paper is the result of research supported by the Office of Naval Research. Reproduction in whole or in part is permitted by the U.S. Government. Distribution of this document is unlimited.

40

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

Harvard University
Department of Statistics
Cambridge, Massachusetts

2a. REPORT SECURITY CLASSIFICATION

2b. GROUP

3. REPORT TITLE

Experiments for Non-Linear Functions

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report;

5. AUTHOR(S) (Last name, first name, initial)

W. G. Cochran

6. REPORT DATE

November 3, 1972

7a. TOTAL NO OF PAGES

~~37~~ 40

7b. NO OF REFS

43

8a. CONTRACT OR GRANT NO.

N00014-67A-0298-0017

b. PROJECT NO.

NR-042-097

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

Technical Report No. 39

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Logistics & Mathematical Statistics Branch
Office of Naval Research
Dept. of the Navy, Washington, D.C.

13. ABSTRACT

This paper reviews the work that has been done on the planning of experiments with response functions non-linear in at least one of the parameters. Apart from older work on the design of dilution series experiments and of quantal bioassays, this field is relatively recent, the mathematical and computing aspects being more complex than for linear responses. In particular, an efficient design requires good advance estimates of the unknown parameters. Research has concentrated on (i) optimum estimation of the parameters, under an asymptotic criterion, for both sequential and non-sequential approaches, (ii) tests of the adequacy of the model, discrimination between models and model-building. There is need for more work on the small-sample behavior of the designs and on compromise designs that perform well under typical complications that arise in practice-missing observations, need for robustness against poor initial estimates, and need for estimates of the experimental error variance.

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Non-linear experiments						
Experimental design						
Sequential design						
Discrimination among models						
Multivariable responses						
The Kullback-Liebler information						
Shannon entropy						
Model building						
Dilution series experiments						
Quantal bioassay						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report titlein all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
______________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
______________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
______________________________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.

1. Introduction

I would like to discuss some of the work that has been done in designing experiments involving response functions non-linear in at least one of the parameters. Formally, this excludes the large volume of work on the planning of factorial experiments and on the estimation of multiple regressions, including polynomial response functions, although there are many similarities in both the methods of attack and the results obtained in the linear and non-linear situations.

This is not an area to which Fisher devoted a great deal of attention. But the first design problem for which he published a solution was non-linear. This was in his 1922 paper on the mathematical foundations of theoretical statistics, before he had published anything on either the analysis of variance or on randomization and the design of agricultural experiments. The problem is the estimation of the density of small organisms in a liquid by means of a series of dilutions. This problem forms a convenient introduction. It is a one-parameter problem, yet illustrates some of the basic features of non-linear problems.

In this review I shall try to concentrate on the issues that obviously present themselves, the methods of attack adopted, the progress made thus far, and some problems still awaiting, so far as I know, published research. The area is an exciting one. On the technical side, a high degree of both mathematical and computing skill is required in the more complex problems. On the practical side, there is the important question: is the research producing the kinds of results that assist the investigator in what he regards as his main problems? Equally important and by no means easy, are we able to explain the methods in terms that the experimenter can understand and use?

2. Dilution series experiments

A volume V of a liquid contains N tiny organisms, thoroughly mixed and with no tendency to clumping or mutual rejection. A small volume x is taken out. The probability that this volume contains no organisms is

$$P = \left(1 - \frac{x}{V}\right)^N \doteqdot e^{-Nx/V} = e^{-\theta x} \quad .$$

Here θ, the density per unit volume, is the parameter to be estimated, while x corresponds to the level of a factor which can be chosen by the experimenter. In practice a standard volume is taken out by pipette, a desired x being obtained by diluting the original volume with pure water. The lab test can detect only whether the sample is sterile (contains no organisms) or fertile (contains one or more organisms).

If n samples are drawn for given x, the probability that s are sterile is the binomial

$$\frac{n!}{s!(n-s)!} P^s Q^{n-s}$$

The criterion which Fisher selected can be described in two equivalent ways. One is that he minimized the large-sample formula for the coefficient of variation o~ the maximum likelihood (ML) estimate of θ. Fisher himself described it as maximizing the sample information about log $\theta = \theta^2 I(\theta)$, where

$$I(\log \theta) = n(\theta x)^2/(e^{\theta x}-1) \quad . \qquad (2.1)$$

He regarded this criterion as the natural one in small as well as large samples, since he used the phrase "without any large-sample approximation" in referring to it.

To maximize I(log θ) in (2.1), the quantity θx should be set at 1.59, giving P=0.20. To find x such that $x\theta = 1.59$, we need to know θ. This is a standard feature that distinguishes non-linear from linear problems. In a non-linear problem, the statistician can say to the experimenter: "You tell me the value of θ and I promise to design the best experiment for estimating θ". If the experimenter replies, "Who needs you?", this is natural but not helpful.

What can be done in practice? Three possibilities suggest themselves. With a good initial estimate θ_o of θ, the experimenter can use Fisher's solution, setting $x = 1.59/\theta_o$, and assuming that he has a good if not an optimum experiment. In Fisher's problem the value of θ is usually known poorly -- perhaps within limits θ_L, θ_H whose ratio is 100 or 1,000 to 1. The natural first question here is: can the experiment be done sequentially? The first experiment has $x = 1.59/\theta_o$, where θ_o is perhaps a poor first guess. The second experiment has $x = 1.59/\hat{\theta}_1$, where $\hat{\theta}_1$ is the M.L. estimate of θ from the first experiment, and so on, creeping up on the best θx.

So far as I know, dilution series experiments are routinely done non-sequentially in a single operation. If θ is thought to lie between θ_L and θ_H, Fisher's approach was not to optimize anything, but to try to guarantee a specified expected value of I(log θ). In a series of two-fold dilutions, for example, the percentage of the total information supplied by different dilutions is shown in Table 1.

Table 1. I(log θ) in percents at different levels of θx

θx	≥8	4	2	1	1/2	1/4	1/8	≤1/16
I(%)	0.9	12.6	26.4	24.5	16.2	9.3	4.9	5.2

The five dilutions from θx=4 to θx=1/4 provide 89% of the total information. To ensure that these dilutions are covered, we want $x_{min}\theta_H \leq 1/4$ and $x_{max}\theta_L \geq 4$. This gives $x_{max}/x_{min} \geq 16\theta_H/\theta_L$. With $\theta_H/\theta_L=100$, twelve two-fold dilutions suffice to cover this range, and 15 when $\theta_H/\theta_L = 1{,}000$.

The Rothamsted laboratory which brought the problem to Fisher did 38 dilution series daily, and he observed that daily calculation of the 38 M.L. estimates would be "exceedingly laborious". Estimating θ by the method of moments (equating the observed total number of sterile plates to the expected number) can be done in less than 5 minutes per series by a table which he provided, now Table VIII2 in Fisher and Yates. Further, he showed in 1922 that the method of moments has an asymptotic efficiency of 88%. Thus, although one of the principal points in his 1922 paper was the superiority of M.L. over moments, he recommends moments for this problem for what seemed to him sound practical reasons.

The dilution series example reveals four types of problems that recur throughout non-linear experiments. (1) setting-up one or more criteria by which to judge alternative proposed designs. Often, much weight will be given to getting good estimates of the parameters, (2) deciding how to proceed when initial estimates of the parameters are dubious. The relative feasibility, cost, and performance of sequential and non-sequential methods become important here, (3) any biometrician, at least, would insist with Fisher that the experiment be capable of providing its own internal estimate of C.V.($\hat{\theta}$). Dilution series can do this if the model is correct and if large-sample formulas can be trusted in small samples -- a point that could stand more checking, (4) checks on the correctness of the model. With two-fold dilution, about 7 dilutions should provide P values between 5 and 95%, giving some data for χ^2 and related checks.

3. Other work by Fisher

Fisher's remaining work on non-linear problems mainly involved using the concept of amount of information as helpful in planning data-collection, as illustrated in the last Chapter of his book, The Design of Experiments (1935). He did much work of this kind, which I will not describe, on the estimation of linkage in humans, animals and plants. In plants, for instance, the amount of linkage between two genes can be estimated by forming a double heterozygote and either crossing it with itself (selfing) or backcrossing it. For estimating close linkage from selfing, he showed that formation of the double heterozygote parent in coupling (AABB×aabb) can be 15 times as efficient as its formation in repulsion (AAbb×aaBB), and is nearly as efficient as backcrossing.

Fisher's first paper (1923) on the analysis of variance, dealt with a 12×6 factorial on potatoes. He first presents the standard ANOVA into main effects and interactions. He then remarks that the preceding analysis is given solely for illustration, since the linear model is obviously unsuitable, predicting negative expected yields for some of the plots. As more reasonable, he proceeds to fit a non-linear product model, which can be written

$$E(y_{ij}) = \mu(1+\alpha_i)(1+\beta_j) \ .$$

This requires more work but as anticipated fits better, the S.S. deviations being 847 against 981. From the 1923 paper, I would not have expected Fisher's later ANOVA work to have concentrated so largely on development of the linear model. I am sorry that I never asked him why.

4. Quantal bioassay (non-sequential)

Another earlier non-linear problem on which much research for practical experiments has been done is quantal bioassay under a normal or a logit tolerance distribution -- a problem again with a 0-1 response. We are comparing a Standard (S) with a Test (T) preparation thought to contain the same active ingredient and therefore to act like a dilution or concentration of the Standard. Thus if x is log dose, an amount x of S has exactly the same effect as an amount x-M of T. Here M, the log relative potency of Test to Standard, is the quantity to be estimated.

To illustrate from the normal model, if n subjects are given an amount x of S, the proportion responding is binomial with

$$P = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\{-\tfrac{1}{2}(x-\mu_S)^2/\sigma^2\}dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu_S)/\sigma} Z(t)dt \qquad (4.1)$$

where Z(t) is the ordinate of the Standard normal curve.

For T, the formula differs only in that $\mu_T = \mu_S - M$. Thus the problem is a three-parameter one, with one parameter M to be estimated and two nuisance parameters.

For a single agent, Fisher showed that

$$I(\mu) = nZ^2/PQ\sigma^2$$

which is maximized at P=0.5, x=μ. Thus if μ_S, μ_T and therefore M were known, the optimum experiment would place all subjects at the levels of S and T causing 50% response.

Lacking this knowledge, experimenters use 2 or more levels of each agent (hopefully straddling the 50% response) from which the M.L. estimates of μ_S, μ_T can be obtained.

If Y is the normal deviate corresponding to P in (4.1)

$$Y = \frac{-\mu}{\sigma} + \frac{x}{\sigma} \quad .$$

For a single agent, Fisher (1935) and others -- see Finney (1947) -- showed that M.L. estimates of μ and σ could be obtained iteratively by a weighted linear regression on x of a transform y (the working transform) of the observed proportion $p = r/n$ of responding subjects.

This approach gives two fitted lines

$$\hat{Y}_S = \bar{y}_S + b(x-\bar{x}_S) \tag{4.2}$$

$$\hat{Y}_T = \bar{y}_T + b(x-\bar{x}_T) \ . \tag{4.3}$$

To obtain the same response, $\hat{Y}_S = \hat{Y}_T$, the difference $\hat{M}$ between the required doses $\hat{x}_S$ and $\hat{x}_T$ is, from (4.2) and (4.3),

$$\hat{M} = \bar{x}_S-\bar{x}_T - (\bar{y}_S-\bar{y}_T)/b,$$

where 1/b estimates the assumed common σ. Since b is first estimated separately for Test and Standard, a test of significance of (b_T-b_S) is available and is regarded as an essential check on the basic assumptions before the combined estimate b is made.

Since $\hat{M}$ involves the ratio $(\bar{y}_S-\bar{y}_T)/b$ of two random variables, Finney's criterion (1964) for the choice of levels of x_T and x_S and of n is the half-width of Fieller's (1940) 5% fiducial interval for M, which is found to be

$$\frac{1.96}{b(1-g)}\left[(1-g)\left(\frac{1}{{}_S\Sigma nw} + \frac{1}{{}_T\Sigma nw}\right) + \frac{(\hat{M}-\bar{x}_S+\bar{x}_T)^2}{S_{xx}}\right]^{\frac{1}{2}} \tag{4.4}$$

where $S_{xx} = \Sigma nw(x-\bar{x})^2$ summed over both agents and $g=(1.96)^2/b^2S_{xx}$ is the square of (1.96 times the coefficient of variation of b).

In designing an experiment, the number of levels k, their spacing d, and the sample size n at each level must be chosen. From previous work on the Standard, good initial estimates of

β and μ_S should usually be available and an initial estimate M_o is assumed. The strategy is to make $x_T = x_S - M_o$ at any level. This should make $(\hat{M}-\bar{x}_S+\bar{x}_T)$ in (4.4) small and the corresponding term in (4.4) is often negligible. In this event, with n constant, (4.4) becomes

$$\frac{(1.96)}{b}\left[\frac{2}{n(1-g)W}\right]^{\frac{1}{2}} \qquad (4.5)$$

where $W = \Sigma w$ over the k levels for one agent. Regarding the quantity multiplying (1.96) in (4.5) as a kind of effective standard error of $\hat{M}$, Finney (1964, 496-7) tabulates $b^2V_E(\hat{M})$ for $k = 2,3,4$, total number of subjects $N = 2kn = 48, 240$, and a range of choices of levels which give P values centred about 50%. A similar table is given for the logistic model in which logit P is assumed linear in x.

These tables provide estimated optimum spacings and the corresponding $b^2V_E(\hat{M})$ for 2,3,4, levels and N = 48, 240. Similar tables for other sample sizes and numbers of levels could easily be provided.

The optimum levels assume good initial guesses. The only work that I have seen allowing poor guesses is by Brown (1966). Using the simpler Spearman-Kärber estimates of μ_S, μ_T, he recommends choices of n, d, k_S and k_T (which he allows to differ), in order to give a desired width of 95% confidence interval for M. This approach is similar to Fisher's in the dilution series. Naturally, more levels are required to ensure coverage of the 50% dose: Brown's worked example gives $k_S = 10$, $k_T = 22$.

Thus, based on the Fieller criterion, available methods furnish

(1) a near-optimum experiment, assuming good initial estimates of σ, μ_S and M, and using large-sample theory,

(2) assuming the model correct, Fieller's limits for the sample data, as a measure of the precision of $\hat{M}$,

(3) for more than 2 levels per agent, tests of the adequacy of the model. The χ^2 for deviations from the model has (2k-3)d.f. These split into 1 d.f. for non-parallelism, 1 d.f. for combined curvature, and (2k-5) d.f. for other sources. Fortunately, as Finney shows, k=4 does not demand more subjects than k=2.

I know of no intensive study of the robustness of the presumed optima to poor initial guesses at the parameter values. Extensions of Finney's tables to more spacings and more sample sizes would reveal the effects of wrong spacing, through a bad guess at σ, on $b^2V_E(\hat{M})$. For r>1, it looks from his tables that the effects are more serious if the guess is σ/r than if it is $r\sigma$, and more serious with fewer levels, as would be expected. Sample size charts by Healy (1950) indicate for k=3 the effects of wrong centering of the doses (through a poor guess at μ_S). More work on robustness and on the small-sample performance of the recommended plans and formulas would be useful.

5. Quantal bioassay (sequential)

A well-known method, the Up and Down or Staircase method (Dixon and Mood, 1948, Dixon, 1965, 1970), was devised for experiments in which it is convenient to test subjects one at a time, determining the level of the agent for the next subject after seeing the result (0 or 1) for the previous subject. For a given dose spacing d, the rule for a single agent (Standard or Test) is the very simple one

$$x_{u+1} = x_u + d \quad (\text{if } y_u = 0); \qquad x_{u+1} = x_u - d \quad (\text{if } y_u = 1) .$$

The idea is, of course, to concentrate dose levels in the neighborhood of μ, the median of the response y, which the method is designed to estimate. The nominal sample size N is defined as the number of trials, beginning with the first pair in which a reversal (0 to 1 or 1 to 0) occurs. The estimate $\hat{\mu}$ of μ is the mean of the last N values of x_u, with an adjustment (Dixon, 1970) depending on the numbers of 0's and 1's that were obtained. The mean square error of $\hat{\mu}$ is approximately $2\sigma^2/N$ when d lies between the limits $d = 2\sigma/3$ and $d = 3\sigma/2$, with $d = \sigma$ recommended as the most accurate spacing. This work is based on exact small-sample computations.

A single sequence provides no usable estimate of σ, which is undesirable if we wish to attach an estimated r.m.s. error $\sqrt{2}\hat{\sigma}/\sqrt{N}$ to $\hat{\mu}$. Dixon (1970) recommends that the experiment be run in independent sequences with N(say)=6 in each sequence. If there are r of these in parallel under the same operating conditions, this speeds up completion of the experiment and allows $V(\hat{\mu})$ to be estimated from $\Sigma(\hat{\mu}_j - \hat{\mu})^2/r(r-1)$. Alternatively, other relevant variables may be changed from one set to another, permitting the effects of these variables on μ to be investigated by analysis of variance techniques.

For a logistic model, when a single (longer) sequence is being used, Wetherill (1966) has proposed a change intended to make the accuracy of $\hat{\mu}$ more robust against a poor initial guess and use of a d too large. After 6 changes of response type have occurred, estimate $\hat{\mu}$, and restart near $\hat{\mu}$ using half the original spacing. Here there remains the problem of an estimate of σ from the data.

Another sequential plan, using the Robbins-Monro stochastic approximation process, attempts to do better than the Up and Down by steadily shortening the steps as the sequence proceeds. If a group of n subjects are tested at each step, the level of x for the (u+1)th experiment is

$$x_{u+1} = x_u - \frac{c}{u}(p_u - \frac{1}{2}) .$$

When the experiment is terminated, the estimate $\hat{\mu}$ is the level at which the next experiment would have been conducted (Cochran and Davis, 1965). With g steps, the asymptotic formula for $V(\hat{\mu})$ is $\pi\sigma^2/2ng$, the value it would have if all trials could be conducted at the optimum 50% level. To guard against a poor initial guess at μ, a 'delayed' version was also suggested in which the step size c remains unchanged until both deaths and survivals have been obtained. A modification with a similar purpose has been proposed by Kesten (1958).

For small experiments with N = ng = 12, where T is the number of steps = 3, 4, 6, or 12, Davis (1971) has compared the M.S.E.'s of $\hat{\mu}$ for three versions of the Robbins-Monro, two of the Up and Down, and a non-sequential experiment using the Spearman-Kärber estimate, for normal, logistic, uniform and exponential tolerance distributions. This is the first broad comparison of the performances of different plans in small samples. It is reasurring that the recommended step size and the asymptotic formulas for $V(\hat{\mu})$ both perform well for starts within about 1.5σ -- about all that can be expected for N=12. Overall, delayed versions of the Up and Down and the Robbins-Monro performed best, both easily beating the non-sequential methods.

6. Single continuous-variable response-a criterion

For the $\underline{u}$th observation or trial $(u=1,2,...,N)$ the model now becomes

$$y_u = f(\underline{\xi}_u;\underline{\theta}) + \varepsilon_u = f(\xi_{u1},...,\xi_{uk};\ \theta_1,...,\theta_p) + \varepsilon_u \ .(6.1)$$

Here, ξ_{ui} denotes the level at which the value of the variable ξ_i is set by experimenter in the $\underline{u}$th trial. There are $\underline{k}$ such factors or variables, while $\underline{p}$ is the number of parameters involved in the model. In the simplest models the ε_u are assumed independently $N(o,\sigma^2)$.

The paper that provided the impetus to intensive work is that of Box and Lucas (1959). Much related earlier work, dealing primarily with the linear case, had been done by Kiefer (1959), Elfving (1952), and Chernoff (1953), who considered the choice of a criterion and the finding of the design points (levels of the factors ξ_{iu}).

The criterion proposed by Box and Lucas assumes interest in all the parameters. It maximizes the generalization of Fisher's amount of information, or equivalently minimizes the asymptotic formula for Wilks' generalized variance of the M.L. estimates of the θ_j. From (6.1) the log likelihood is

$$L = -\frac{1}{2\sigma^2}\sum_{u=1}^{N}(y_u-f_u)^2 \quad .$$

It follows that the information matrix is

$$E\left(\frac{-\partial^2 L}{\partial\theta_i\partial\theta_j}\right) = \frac{1}{\sigma^2}\sum_{u=1}^{N}\left(\frac{\partial f_u}{\partial\theta_i}\right)\left(\frac{\partial f_u}{\partial\theta_j}\right) = \frac{1}{\sigma^2}(X'X)$$

where X is the N×p matrix

$$(x_{uj}) = \frac{\partial f_u}{\partial\theta_j} \quad .$$

The x_{uj} are known when the factor levels ξ_{ui} and the θ_j are known. The criterion - choose design points ξ_{ui} to maximize $|X'X|$ -assumes initial guesses θ_j for practical use. Other attractive features of this criterion (summarized by M.J. Box and Draper (1971)) are as follows.

(1) It minimizes the volume of the asymptotic confidence region for the θ_j (Kiefer, 1961).

(2) For response functions locally linear in the neighborhood of the M.L. estimates, it maximizes the joint posterior probability of the θ_j, given a non-informative prior $\Pi d\theta_j$. (Draper and Hunter, 1966).

(3) It is invariant under changes of scale of the θ_j.

7. Finding the design points - non-sequentially

Given a criterion, the next step is the complex one of finding design points that satisfy the criterion for a specified N trials. In earlier work, Chernoff (1953) considered the case where our interest is in $s \leq p$ of the parameters, the remaining (p-s) being nuisance parameters. His criterion was different - minimizing the average of the asymptotic variances of the $\underline{s}$ M.L. estimates. Following Elfving (1952), he showed that an optimum design needs at most $s(2p-s+1)/2$ points, becoming $p(p+1)/2$ when $s=p$, and p when $s=1$.

As a start, Box and Lucas (1959) assumed initial guesses $\underline{\theta}_o$ and sought an optimum set of levels when $N=p$, i.e. when there are only as many trials as parameters to be estimated. They point out unappealing features of this decision: no test of the fit of the model, no attempt at robustness against poor initial guesses, to which might be added no data for an experimental estimate of σ^2.

One advantage with $N=p$ is that $(X'X)$ is square, so that $|X'X| = |X|^2$ and it suffices to maximize $|X| = |x_{ui}|$. Illustrative examples worked by Box and Lucas include the exponential growth or decay curve, the Mitscherlich equation, and the two-factor function

$$f(\xi_1,\xi_2,\cdot\theta_1,\theta_2) = \exp(-\theta_1\xi_1 e^{-\theta_2\xi_2}) .$$

Depending on the complexity of the problem, methods available for solution are

1. Geometric or analytic,
2. Calculate $|X|$ for a grid of values of the ξ_{iu}, fit a quadratic to this grid and seek a maximum (with trouble possible if $|X|$ has more than one turning value)
3. Various computer iterative hill-climbing techniques.

As a simple example with an analytic solution, consider the exponential decay curve

$$f_u = \theta_1 e^{-\theta_2 t_u}$$

where t_u (time) is used for ξ_{u1}. The region feasible for experiments is $t(\min) \leq t \leq t(\max)$. For this f_u,

$$|X| = \begin{vmatrix} e^{-\theta_2 t_1} & -t_1\theta_1 e^{-\theta_2 t_1} \\ e^{-\theta_2 t_2} & -t_2\theta_1 e^{-\theta_2 t_2} \end{vmatrix}$$

$$= \theta_1(t_1-t_2)e^{-\theta_2(t_1+t_2)} \quad .$$

This can be written

$$|X| = \{\theta_1(t_1-t_2)e^{-\theta_2(t_1-t_2)}\}\{e^{-2\theta_2 t_2}\} \quad .$$

For given (t_1-t_2) and with $\theta_2>0$, we want $t_2=t(\min)$. The first curly bracket is maximized when

$$t_1-t_2 = 1/\theta_2, \text{ giving } t_1 = t(\min)+1/\theta_2$$

or $\quad t_1 = t(\max)$, whichever is smaller.

Coming to the case of a single non-sequential experiment with $N>p$, Atkinson and Hunter (1968) found in several chemical examples worked by computer maximizing that with N a multiple of p, the optimum plan consisted simply of N/p replications at each of the p optimum sets of levels for the case N=p. This result certainly simplifies the finding of optimum plans. Although a counter example showed that the result does not hold in general, they proved, as a sufficient condition, that the result will hold if the region of experimentation lies within a certain ellipsoid in the x-space (a point that can be checked by the experimenter.)

M.J. Box (1968a,1970a) considered also the case: N not a multiple of p. In some problems he found that replications of the N=p solution differing by at most 1 could be proved to be optimal. In others, while this could not be proved, a computer search was unable to locate anything superior to the near-equal-replication solution. He also considered a one-factor, two-parameter problem with ξ_{u1} = time = t_u, where different trials cost different amounts. The problem was to maximize $|X'X|$ subject to a fixed cost $C = \Sigma c_u$. The optimum again consisted of experiments at only two times t_1, t_2, but with the difference that t_1 and t_2 changed both with N and C and the numbers of replications were no longer near-equal, so that more computing effort was necessary.

The counter-example by Atkinson and Hunter is the linear fitting of a bivariate regression, $f_u = \theta_1\xi_{1u}+\theta_2\xi_{2u}$, with the region of experimentation $0\leq\xi_{iu}\leq 1$. For N=p=2, the optimum design is at the levels (1,0) and (0,1), which gives

$$X'X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \; |X'X| = 1 \; .$$

With N=6, three replications of this plan give

$$X'X = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}; \; |X'X| = 9 \; .$$

But two replications of the three-point plan (1,0), (0,1), (1,1) give

$$X'X = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}; \; |X'X| = 12 \; .$$

The key ellipse in this example is the circle $\xi_1^2 + \xi_2^2 = 1$, and the point (1,1) in the experimental region lies outside this circle.

The preceding results on the best set of design points are conceptually similar to Fisher's original optimum for the dilution series problem, and assume in effect good initial estimates of the θ_j. With poor initial guesses, the resulting plan will not be optimal in any real sense. I have come across no work analogous to Fisher's, where we start with a wider spread than p points with the object of guaranteeing a specified value of $|X'X|$ starting from initial θ_j assumed known initially only to lie within a certain region.

8. Finding the design points sequentially

As would be expected, the methods start with p points, determined by first guesses θ_{oj}, and leading to M.L. estimates $\hat{\theta}_{1j}$ of all the parameters. Box and Hunter (1965a) discuss how to add points one at a time. If (N-1) steps have been completed, so that $\hat{\theta}_{N-1,j}$ are known, then $|X'X|$ as a function of the x's for the $\underline{N}$th point takes the form

$$|X'X|_N = \begin{vmatrix} c_{11}+x_{1N}^2 & c_{12}+x_{1N}x_{2N} & \cdots & c_{1p}+x_{1N}x_{pN} \\ c_{12}+x_{1N}x_{2N} & c_{22}+x_{2N}^2 & \cdots & \\ \cdot & \cdot & \cdot & \cdot \\ c_{1p}+x_{1N}x_{pN} & \cdots & & c_{pp}+x_{pN}^2 \end{vmatrix}$$

where the c_{ij} are known. The criterion is computed for all points of a grid of values of the $\xi_{u1} \ldots \xi_{uk}$ and a quadratic fitted to find the maximizing values.

M.J. Box (1970a) adds sequential sets of n=p points, each put at the best p design points as estimated from the M.L.$\hat{\theta}$ obtained from the combined trials conducted to date. After a time, both the M.L.$\hat{\theta}$ and the indicated set of p design points for the rth set begin to change little from those in the (r-1)th set. Box introduces a criterion R_1 as a guide to the time when it is no longer worth changing points. A second quantity R_2 compares the $|X'X|$ value given by all trials conducted to date with the value that $|X'X|$ would have if it had been possible to use our current estimate of the best design points in all trials. Thus R_2 indicates the amount lost owing to poor initial guesses at the θ_j. In the simulated example (3 parameters, 2 factors), some values of R_1 and R_2 are as in Table 8.1.

Table 8.1. Values of R_1, R_2 in sequential plan

Set	2	3	4	5	6
R_1	1.40	1.09	1.06	1.04	1.02
R_2	0.78	0.86	0.86	0.88	0.91

In order to study the effect on the sequential process of having initial prior information of different amounts about different θ_j, Draper and Hunter (1967a) took a multinormal prior

$$(2\pi)^{-\frac{1}{2}p}|\Omega|^{-\frac{1}{2}}\exp\{-\tfrac{1}{2}(\theta-\theta_o)'\Omega^{-1}(\theta-\theta_o)\}$$

where Ω is the p×p matrix of variances and covariances and the $\underline{\theta}_o$ are initial guesses. In the case where $\underline{N}$ trials had already been completed at chosen levels $\underline{\xi}_u$, they discussed where to put a further $\underline{n}$ trials. Their criterion was to maximize the posterior distribution of $\underline{\theta}$ after (N+n) trials with respect both to $\underline{\theta}$ and to the values $\underline{\xi}_u$ (u=N+1,...,N+n). Assuming $f(\underline{\xi}_u,\theta)$ to be locally linear, this leads to the approximate criterion: maximize

$$|X'X + \sigma^2\Omega^{-1}| \qquad (8.1)$$

with respect to $\underline{\theta}$ and $\underline{\xi}_u$ (u=N+1,...,N+n). One hurdle is that in (8.1) the values of $\underline{\theta}$ are hidden in X'X and their maximizing values after (N+n) trials depend on observations not yet taken. The natural suggestion is to use $\hat{\theta}_N$ in maximizing (8.1) with respect to the levels $\underline{\xi}_u$.

The principal value of this type of prior is likely to be the light it throws on how the design would be affected by different amounts of prior information about the different θ_j. As an illustration they work a problem with θ_1,θ_2 independent normals $(0,\sigma_j^2)$, N=0, n=2, and a <u>single</u> ξ_{u1} (=t_u, a time variable). As σ_1,σ_2 vary from 0 to ∞, three basic design types predominate: (t_1,t_2) = (1.2,6.9) for little prior information, (t_1,t_2) = (1.2,1.2), where the experiment concentrates on estimating θ_1, and (t_1,t_2) = (6.9,6.9), where the emphasis is on θ_2. Further illustrations of this type would be of interest.

9. Tests of fit of the model

As a dividend, the sequential approach might provide some data for a test of fit of the model (at least assuming σ^2 known), since $\underline{y}_u$ will have been determined in general at N>p design points. If, however, the successive design points vary over only a restricted part of the experimental region, examination of the residuals may tell us little. Experience with a wider range of non-linear models may throw more light on this issue.

I know of no work in which an N>p experiment was deliberately planned ab initio with a test of lack of fit as one objective. Two suggestions have been made by Box and Lucas (1959).

First, having computed the combinations of levels of the $\underline{\xi}_u$ needed to maximize $|X'X|$, the experimenter might examine where they occur in the $\underline{\xi}$ space of interest, and add extra points where he is most worried that the model may be incorrect.

Secondly, the experimenter may sometimes be reasonably sure that if the model is incorrect, a more general model with say one or two extra parameters gives an adequate fit. The example cited is where the model (with a single ξ variable) is

$$f_u = 1-e^{-\theta_1\xi_u}$$

where in fact the more general model

$$f_u = \frac{\theta_1}{(\theta_1-\theta_2)}\{e^{-\theta_2\xi_u} - e^{-\theta_1\xi_u}\}$$

might be required. The experiment might be planned to estimate θ_1 and θ_2 and test the N.H.θ_2=0, which makes the original model correct.

10. Discrimination between specified models

An approach by D. R. Cox (1961, 1962) for discrimination between two models, used a test of significance and was asymmetric: the hypothesis that model 1 is correct was chosen as the null hypothesis. The test criterion was a modified form of the likelihood ratio, maximizing the asymptotic power against model 2 as the alternative. Later, in planning an experiment to discriminate between the probit and the logit models from observations confined to 3 log dosages, Chambers and Cox (1967) used a compromise symmetric form of this approach. They first chose a criterion asymptotically powerful against A.H = logistic, given NH = probit. For this criterion, they determined the optimum three dosage levels and the proportions of the observations to be put at each level. Then they reversed the procedure, having A.H. = probit, N.H. = logistic. Fortunately, the optimum doses did not differ greatly in the two cases, so that a good compromise design could be constructed. Unfortunately, as Chambers and Cox note, this plan put the majority of the observations at a high dose level with expected percent killed over 99.6%. Thus the experiment would require large samples, as will not surprize those who have worked with both probits and logits. This approach might end, of course, by rejecting neither model, one specific model, or both models.

An alternative approach, Box and Hill (1967), is symmetric and extends to more than two specific models. For two models, the approach supposes that _n_ observations have already been taken (at least enough to estimate any parameters involved) and considers where best to put the (n+1)th for maximum discrimination. At first sight one might be inclined to seek the point (levels of the factors) for which $|\hat{Y}_1-\hat{Y}_2|$ is maximized, where $\hat{Y}_1$ and $\hat{Y}_2$ are estimated from the results for the first _n_ observations. But as Box and Hill note, the precision of estimation of $|\hat{Y}_1-\hat{Y}_2|$ is also relevant.

Prior probabilities Π_{io} are first assigned to each model. With two models, the choice might be $\Pi_{io} = 1/2$ and with m models, $\Pi_{io} = 1/m$. After n runs, the posterior probability for the ith model is

$$\Pi_{in} = \Pi_{i,n-1}p_i \;/\; \Sigma\Pi_{i,n-1}p_i$$

where p_i is the probability density function of the nth observation y_n under model i.

The criterion chosen for discrimination uses Shannon's (1948) concept of entropy, also known as the Kullback-Liebler information (1951). For m models the entropy is $-\sum_{i=1}^{m} \Pi_i \ln \Pi_i$.

This has its maximum value when $\Pi_i = 1/m$ and becomes steadily smaller as the Π_i becomes unequal, i.e. as discrimination improves. Hence the (n+1)th observation is chosen at levels ξ which will maximize the expected decrease in entropy from the nth to the (n+1)th experiment.

For two models the resulting discrimination criterion is shown to be

$$D = \Pi_{1n}\Pi_{2n}\left(\int p_1 \ln(p_1/p_2)dy_{n+1} + \int p_2 \ln(p_2/p_1)dy_{n+1}\right).$$

If we can further assume that the models are locally linear, with deviations ξ_u that are $N(o,\sigma^2)$, where σ^2 is known, the criterion becomes, for two models $Y^{(1)}$ and $Y^{(2)}$,

$$D = \tfrac{1}{2}\Pi_{1n}\Pi_{2n}\left\{\frac{(\sigma_1^2-\sigma_2^2)^2}{(\sigma^2+\sigma_1^2)(\sigma^2+\sigma_2^2)} + \left(\hat{Y}^{(1)}_{n+1}-\hat{Y}^{(2)}_{n+1}\right)^2\left(\frac{1}{\sigma^2+\sigma_1^2} + \frac{1}{\sigma^2+\sigma_2^2}\right)\right\},$$

where σ_i^2 is the approximate variance of $\hat{Y}_{n+1}^{(i)}$. For m criteria, D is the corresponding expression summed over all pairs of models.

One of the examples worked is a simulated example, with 4 parameters and 2 factors, to distinguish among the first- to fourth-order reaction curves. The experiment starts with a grid of 4 points to estimate all 4 parameters needed. The results proceed as in Table 10.1.

Table 10.1. Example of discrimination among models

n	ξ_1	ξ_2	Π_1	Π_2	Π_3	Π_4
1	25	575				
2	25	475				
3	125	575				
4	125	475	.01	.43	.50	.06
5	125	600	.00	.56	.43	.01
6	125	600	.00	.86	.13	.00
7	50	450	.00	.97	.02	.00
8	100	600	.00	1.00	.00	.00

From the beginning, the competition is between the second- and third-order curves, the second-order soon establishing itself as correct.

Box and Hill also work an example in which (i) all models are generalizations of model 1 and (ii) model 1 is correct so that all models are correct. Here the entropy criterion seems to be given an impossible task, but by their largest n(15), it is tending towards selection cf the simplest of the correct models - an admirable performance. However, in more recent examples of this situation in which n was continued to large values, Siddik (1972) found that the posterior probability of the simplest correct model rose to a value 0.85 to 0.95, but then fluctuated erratically around that value. While it still can be conjectured that the criterion will operate well in practical experiments, its large-sample performance needs further study.

A succeeding paper by Hill, Hunter and Wichern (1968) recognizes that the best choice of the ξ levels for discrimination will not in general be those that give the best parameter estimation for the correct model, and seeks to reconcile these conflicting aims. If we knew that model $\underline{j}$ was the correct model, we would choose the ξ to maximize $\Delta_j = |X'X|_j$ for model j. Call this value $\Delta_{j,\max}$ and let Δ_j denote the value of the estimation criterion Δ_j for any other choice of levels $\underline{\xi}$. Similarly, let $D_{\max}$ be the maximum expected decrease in entropy, and D the decrease obtained from any other setting of the $\underline{\xi}$. The criterion which these authors suggest for choosing the $\underline{\xi}$ levels is

$$C = w_1 D/D_{\max} + w_2 \sum_{j=1}^{m} \Pi_{jn}\Delta_j/\Delta_{j,\max} \quad .$$

The w_1 and w_2 are weights ($w_1+w_2=1$) which can be changed, as the sequence of runs proceeds, to give increasing weight to good parameter estimation when it becomes clearer that one model is being selected by the discrimination technique. For w_1 they suggest, as one possibility,

$$w_1 = \{m(1-\Pi_{bn})/(m-1)\}^{\lambda}$$

where Π_{bn} is the probability assigned to the best model before the (n+1)th observation is taken. The quantity λ is a positive power that controls the rate of decrease of w_1, the weight assigned to the discrimination criterion. Initially, if all $\Pi_{io} = 1/m$, w_1 is unity and all emphasis is given to good discrimination. As Π_{bn} approaches 1, so does w_2, emphasis shifting to estimation for the most likely model.

11. Model building

It is more difficult to do justice to the work here, since the strategies will change as the accumulated data suggest new ideas to experimenter and statistician.

As one approach, Box and Hunter (1962, 1965b) consider the case where the experimenter has at best a tentative model which describes $f(\underline{\xi},\underline{\theta},t)$. If there are $\underline{k}$ factors $\underline{\xi}$ which the experimenter can manipulate, they suggest running the reaction, with measurements of response at certain fixed times, for a 2^k factorial or fractional factorial in the levels of the ξ_i, widely separated as far as operating restrictions permit. For each combination of the factor levels they estimate each θ_j and do a standard factorial analysis into main effects and interactions for each $\hat{\theta}_j$. There are two objectives in this procedure: (i) if the model is correct, the $\hat{\theta}_j$ should not change systematically with time or with the changes in the levels of the ξ_i, since the θ_j should be constant, and (ii) the way in which the $\hat{\theta}_j$ change may enable the experimenter to specify a vague model more completely, or may suggest relations among the θ_j and the ξ_i that make sense mechanically.

In their simulated example they use letters A, B, (the initial concentrations of two reactants), C (the concentrations of a catalyst), and D (the temperature), to denote the factors instead of our $\xi_1,\ldots,\xi_4$. The tentative model was

$$E(y) = \frac{(B)k_1}{k_1-k_2}(e^{-k_2t} - e^{-k_1t}) \quad .$$

The initial experiment was a 2^4 factorial in A...D, measured at 5 times. The nature of the reaction suggested that

$$k_1 = (A)^{p_1}(C)^{q_1}\alpha_1 e^{-\beta_1/T}$$

$$k_2 = (A)^{p_2}(C)^{q_2}\alpha_2 e^{-\beta_2/T}$$

where T is the absolute temperature. (There are now 8 parameters to be estimated). If this suggestion is correct, a factorial analysis of ℓnk_1 and ℓnk_2, which was then carried out, should show no effects of B and no interactions involving A, C, and D. The analysis confirmed the model, as did careful examination of the residuals (y-f) for the 2^4 runs at the 5 times conducted initially. Finally, the 8 parameters were estimated from the combined data. The second paper (Box and Hunter, 1965) gives further discussion of the examination of residuals, contour diagrams, and plots of the likelihood function as diagnostic aids. The necessity for repeated interchange of ideas between experimenter and statistician is stressed.

An interesting review of approaches and problems in model-building by M.J. Box (1968b) presents his experiences, with discussion from the audience.

12. More than one measured response

In some chemical reactions it is possible to measure more than one response $y_{u\ell}$ $(\ell=1,2,\ldots,L)$ which provides information about some or all of the parameters θ_j. The simplest example quoted is the one-parameter exponential

$$y_{1u} = e^{-\theta\xi_u} + \varepsilon_{1u}; \quad y_{2u} = 1 - e^{-\theta\xi_u} + \varepsilon_{2u}$$

where the single factor ξ_u represents time. Note that y_{1u}, y_{2u} do not add to 1 because of the experimental errors $\varepsilon_{\ell u}$.

In a general approach the model is

$$y_{\ell u} = f_\ell(\underline{\xi}_u;\underline{\theta}) + \varepsilon_{\ell u} .$$

It has not been considered realistic to assume $\varepsilon_{\ell u}, \varepsilon_{mu}$ independent. Instead, they are given a multivariate normal distribution with variance-covariance matrix $\sigma_{\ell m}$.

The first paper on this problem, Box and Draper (1965) did not assume the $\sigma_{\ell m}$ known in advance, and merely assigned a 'non-informative' prior distribution to the $\sigma_{\ell m}$. Later papers took the more tractable problem in which the $\sigma_{\ell m}$ are assumed known, and will be considered first.

With known $\sigma_{\ell m}$, Draper and Hunter (1966) assigned a Bayesian prior $\Pi d\theta_j$ and followed the method which led to the $|X'X|$ criterion for a single response, as mentioned in Section 7. It helps to write

$$v_{\ell m} = \sum_{u=1}^{N} \{y_{\ell u} - f_{\ell u}\}\{y_{mu} - f_{mu}\} \quad . \tag{12.1}$$

They find that the posterior probability is

$$p(\underline{\theta}|\underline{y}) = c \exp\{-\tfrac{1}{2} \sum_{\ell}^{L} \sum_{m}^{L} \sigma^{\ell m} v_{\ell m}\} \; . \tag{12.2}$$

In this approach the θ_j would be estimated by minimizing

$$\Sigma\Sigma\sigma^{\ell m} v_{\ell m} \tag{12.3}$$

that is, by the natural extension of the method of least squares to the case of multivariate normal deviations.

By an extension of the univariate method, the further assumption that the response functions are approximately linear in the vicinity of the M.L. estimates leads to the criterion: choose the design points to maximize

$$\Delta = \left| \sum_{\ell=1}^{L} \sum_{m=1}^{L} \sigma^{\ell m} X_\ell' X_m \right| \qquad (12.4)$$

where for given ℓ, X_ℓ is the N×p matrix

$$X_\ell = (x_{\ell uj}) = \frac{\partial f_\ell}{\partial \theta_j} (\underline{\xi}_u;\underline{\theta}) \quad .$$

The matrices X_ℓ should strictly be evaluated at the M.L. $\hat{\underline{\theta}}$ after the experiment has been completed, which cannot be done when the experiment is being planned. If no trials have been conducted, the suggestion is to compute the X_ℓ for initial guesses θ_o; if N trials have been done and a further $\underline{n}$ are being planned, use the X_ℓ at the M.L. estimates after $\underline{N}$ trials.

Illustrations were given for a two-response, one-parameter problem and by M.J. Box (1970a) for a two-response, two-parameter and for a two-response, four-parameter problem. Draper and Hunter's interest was to see how the optimum plan and the value of the criterion Δ in (9.4) varied with σ_{11}, σ_{22}, and ρ, while Box considered whether replications of the optimum N=p plan were still to be recommended. For N a multiple of p, equal replications of this optimum were the best he could find. For N not a multiple of p, the best of the near-equal replications was not optimal, but near enough as a good start in a computer search for anything better. M.J. Box comments that this search may not be worth the trouble, though further experience is needed.

Draper and Hunter (1967b) have also extended to this case the single-response work reported in section 8 for a multinormal prior. With two responses, for instance, the criterion to be maximized is

$$\Delta = |\sigma^{11}X_1'X_1+\sigma^{22}X_2'X_2+\sigma^{12}(X_1'X_2+X_2'X_1)+\Omega^{-1}|$$

where Ω is the prior covariance matrix of the θ_j.

Returning to the case of a 'non-informative' prior that leads to the criterion (12.4), M.J. Box (1970b) considered two practical complications. (1) The response variables may not be measured directly but computed from other prime variables, measured directly, whose values change as the design points change,(2) The factor levels $\underline{\xi}$ may be themselves subject to error (a familiar problem in experimentation). Consequences are that the σ_{ij} vary with the design points and that it becomes less reasonable to think of 'dependent' variables $y_{\ell u}$ and 'independent' variables ξ_{iu}. Nevertheless, by assuming that the basic measurements are independent, with known variances, he has developed a computer program (essentially involving a known σ_{ij} changing with the design points). An example illustrates the application of this technique.

As mentioned, Box and Draper (1965) considered the case where the $\sigma_{\ell m}$ are not known in advance. They assigned the prior $\Pi d\theta_j$ to the parameters and the prior

$$p(\sigma^{\ell m}) = |\sigma^{\ell m}|^{-\frac{1}{2}(L+1)}$$

which is the multivariate extension of assigning a uniform prior to (log σ) in the univariate case. They find the posterior

$$p(\underline{\theta}|\underline{y}) = C|v_{\ell m}|^{-\frac{1}{2}N}$$

where C is a constant. The θ_j would then be estimated by minimizing $|v_{\ell m}|$. At first sight this criterion seems rather different from the criterion (11.3): minimize

$$\Sigma\Sigma\sigma^{\ell m}v_{\ell m}$$

which emerged when the $\sigma_{\ell m}$ were assumed known. Box and Draper show, however, that there is a natural resemblance. Let $V_{\ell m}$ be the cofactor of $v_{\ell m}$. Now $|v_{\ell m}|$ can be calculated by multiplying the elements of $v_{\ell m}$ in any <u>single</u> row or column by their cofactors and adding. It follows that

$$|v_{\ell m}| = \Sigma\Sigma\frac{V_{\ell m}}{L}v_{\ell m} \quad . \qquad (11.5)$$

Thus the weights $\sigma^{\ell m}$ in (11.3) are replaced by weights proportional to the M.L. estimates of the $\sigma^{\ell m}$.

The two simulated examples worked both involve only a single ξ_u variate (time). One example has two responses, one parameter, one has 3 responses, 2 parameters. It is now necessary to take N>p in order to obtain estimates of the weights $V_{\ell m}$. The values chosen were N=10 for the L=?, p=1 example and N=12 for the L=3, p=2 example, no attempt being made to find optimum values of $\underline{t}$ (design pointc). From the worked examples a recommendation is made to plot the complete posterior functions for the θ_j obtained (i) from each individual response function (ii) from each pair and (with L=3) from the three combined. These plots indicate the type and amount of information supplied about the respective θ_j by individual responses and combinations of them. They can also reveal deficiencies in the model, e.g. when the posterior from y_{1u} has little overlap with that for y_{2u}.

13. Comments

From the work on the 0-1 and the continuous-variable response, we now have a good grasp of the multiple desirable objectives in a non-linear experiment, a body of techniques that concentrate on a general-purpose criterion for giving good estimates of all the parameters, a method for discriminating among specified models, and an attack on the problem of model-building. Since much of the continuous-variable work is recent, with simulated examples, I would expect a period of digestion by experimenters in industry, with feedback on features that they like and don't like and additional properties desired. Further, as the groups at Wisconsin and I.C.I. warn us, the industrial workers have still harder problems awaiting attack.

It is easy to list much additional related work that would be relevant. To mention a few areas:

(1) Criteria and designs for the estimation of only some of the parameters, the others being regarded as nuisance parameters.

(2) Compromise designs, non-optimal by any single criterion, that cope with several different objectives.

For instance, a non-sequential plan might deliberately start with N>p distinct points, in order to provide (i) some robustness against poor initial $\underline{\theta}_o$, (ii) either a check on the correctness of the model if an outside estimate of σ^2 is available, or (iii) an internal estimate of σ^2 if the model can be assumed correct. Something to provide both a check and an estimate of σ^2 might I suppose be possible by a development analogous to Tukey's 1 d.f. for non-additivity under the linear model.

(3) Since the approach and formulas are to a large extent asymptotic, checks by computer studies on the small sample performance of the 'optimum' plans and formulas.

(4) Finally, and in no invidious sense, I hope that more people will enter this field, with a resulting broader range of problems attacked, of techniques developed, and of viewpoints. The discussion in the Royal Statistical Society, following Kiefer's (1959) presentation of his work on optimum linear plans, revealed doubts about the wisdom of concentrating on optimizing any single criterion. Reasons advanced were that optimizing may require mathematical assumptions or restrictions found unreasonable in many applications, that the experimenter's aims may change when he begins to see some results, and that, in sequential experiments, rules leaving flexibility of judgment to the experimenter and therefore sounding vague to some degree may be better

than fixed rules laid down by a statistician's criterion. While this part of the discussion was somewhat negativistic in tone, it suggested that approaches from differing viewpoints have an important role.

In a paper delivered at these meetings, Wheeler (1972) maintains that the experimenter should seek a design that will be reasonably efficient under a variety of situations which he judges that he may face. Thus for insurance he may want to fit a model more complex than the one that he hopes is correct, he may fear some loss of observations from accidents, and may want at least a specified number of degrees of freedom for estimation of σ^2. To indicate the inefficiency of any proposed plan, relative to a plan that concentrates solely on efficiency of estimation, Wheeler uses as criterion the relative maximum variance of the predicted response over the experimental region, illustrating how the extensive results on optimum design for linear models, in particular Wynn (1970), provide computer methods for meeting these goals.

I wish to thank P. Morse, S. M. Siddik, and R. E. Wheeler for information about recent work.

References

Atkinson, A. C. and Hunter, W. G. (1968). The design of experiments for parameter estimation. Technometrics, 10, 271-289.

Box, G. E. P. and Draper, N. R. (1965). The Bayesian estimation of common parameters from several responses. Biometrika, 52, 355-365.

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

Box, G. E. P. and Hunter, W. G. (1962). A useful method for model-building. Technometrics, 4, 301-317.

Box, G. E. P. and Hunter, W. G. (1965a). Sequential design of experiments for non-linear models. Proc. IBM Scientific Computing Symposium in Statistics, 113-137.

Box, G. E. P. and Hunter, W. G. (1965b). The experimental study of physical mechanisms. Technometrics, 7, 23-41.

Box, G. E. P. and Lucas, H. L. (1959). Design of experiments in non-linear situations. Biometrika, 46, 77-90.

Box, M. J. (1968a). The occurrence of replications in optimal designs of experiments to estimate parameters in non-linear models. Jour. Roy. Statist. Soc. B., 30, 290-302.

Box, M. J. (1968b). The use of designed experiments in non-linear model building. In The Future of Statistics, Ed. D. G. Watts, Academic Press, New York, 241-268.

Box, M. J. (1970a). Some experiences with a non-linear experimental design criterion. Technometrics, 12, 569-589.

Box, M. J. (1970b). Improved parameter estimation. Technometrics, 12, 219-229.

Box, M. J. and Draper, N. R. (1971). Factorial designs, the |X'X| criterion, and some related matters. Technometrics, 13, 731-742.

Brown, B. W. Jr. (1966). Planning a quantal assay of potency. Biometrics, 22, 322-329.

Chambers, E. A. and Cox, D. R. (1967). Discrimination between alternative binary response models. Biometrika, 54, 573-578.

Chernoff, H. (1953). Locally optimum designs for estimating parameters. Ann. Math. Statist., 24, 586-602.

Cochran, W. G. and Davis, M. (1965). The Robbins-Monro method for estimating the median lethal dose. *Jour. Roy. Statist. Soc.*, *B*, 27, 28-44.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symposium*, 1, 105-123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Jour. Roy. Statist. Soc.*, *B*, 24, 406-424.

Davis, M. (1971). Comparison of sequential bioassays in small samples. *Jour. Roy. Statist. Soc. B*, 33, 78-87.

Dixon, W. J. (1965). The up and down method in small samples. *Jour. Amer. Statist. Ass. 60*, 967-978.

Dixon, W. J. (1970). Quantal-response variable experimentation: the up-and-down method. In *Statistics in Endocrinology*, ed. J. W. McArthur and T. Colton, M.I.T. Press, 251-268.

Dixon, W. J. and Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *Jour. Amer. Statist. Ass. 43*, 109-126.

Draper, N. R. and Hunter, W. G. (1966). Design of experiments for parameter estimation in multiresponse situations. *Biometrics*, 53, 525-533.

Draper, N. R. and Hunter, W. G. (1967a). The use of prior distributions in the design of experiments for parameter estimation in non-linear situations. *Biometrika*, 54, 147-153.

Draper, N. R. and Hunter, W. G. (1967b). The use of prior information in the design of experiments for parameter estimation in non-linear situations: multiresponse case. *Biometrika*, 54, 662-665.

Elfving, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Statist.*, 23, 255-262.

Fieller, E. C. (1940). The biological standardization of insulin. *Jour. Roy. Statist. Soc. Supp.* 7, 1-64.

Finney, D. J. (1947). The principles of biological assay. *Jour. Roy. Statist. Soc. Supp.* 9, 46-91.

Finney, D. J. (1964). Statistical method in biological assay. 2nd. ed. Griffin and Co., London.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.*, London, A, 222, 309-368.

Fisher, R. A. (1935). The case of zero survivors. Ann. App. Biol., 22, 164-165.

Fisher, R. A. and Mackenzie, W. A. (1923). The manurial response of different potato varieties. Jour. Agric. Sci., 13, 311-320.

Healy, M. J. R. (1950). The planning of probit assays. Biometrics, 6, 424-434.

Hill, W. J., Hunter, W. G. and Wichern, D. W. (1968). A joint design criterion for the dual problem of model discrimination and parameter estimation. Technometrics, 10, 145-160.

Kesten, H. (1958). Accelerated stochastic approximation. Ann. Math. Statist., 29, 41-49.

Kiefer, J. (1959). Optimum experimental design. Jour. Roy. Statist. Soc., B, 21, 272-319.

Kiefer, J. (1961). Optimum designs in regression problems. II Ann. Math. Statist., 32, 298-325.

Kullback, S. and Liebler, R. A. (1951). On information and sufficiency. Ann. Math. Statist., 22, 79-86.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Tech. Jour., 27, 379-423 and 623-656.

Siddik, S. M. (1972). Kullback-Liebler information function and the sequential selection of experiments to discriminate among several linear models. Ph.D. thesis,Case Western Reserve University.

Wetherill, G. B. (1966). Sequential methods in statistics. Methuen & Co., London.

Wheeler, R. E. (1972). Efficient experimental design. Submitted to Technometrics.

Wynn, H. P. (1970). The sequential generation of D-optimum experimental designs. Ann. Math. Statist. 41, 1655-1664.