TESTS FOR DISCRIMINATION BETWEEN MODELS

by

Anthony Curtis Atkinson

A thesis submitted for the Ph. D. degree of London University

Imperial College

1970

ABSTRACT

The thesis is in two main parts, both concerned with the choice between several alternative models for the description of data in the presence of random variation.

In Part 1 the models specifying the expected value of each observation do not contain adjustable parameters. A procedure is developed for testing whether all the models describe the data equally well. Under the customary assumptions about the normal distribution of errors, the test statistic has the F distribution.

A more general problem, that of deciding which of two or more distributions best describe the data, is considered in Part 2. The method followed is to combine the separate distributions into one distribution of which they are special cases and to test hypotheses about the parameters of combination. Examples are given of applications to problems involving continuous, discrete and binary data. Asymptotic results are obtained and compared, for finite samples, with results obtained by simulation.

The supplementary Part 3 contains reprints of four papers, two written with co-authors, which deal with topics in operational research and the design and analysis of experiments.

ACKNOWLEDGEMENTS

The research described here was suggested and supervised by Professor D.R. Cox. I am very grateful to him for his continual interest and guidance which resulted not only in this thesis but also in a most enjoyable two years.

The work was supported by an IBM Fellowship.

CONTENTS

Abstract	2
Introduction	5
Part 1 The Choice Between Prediction Formulae	7
Part 2 A General Method for Discriminating between Models	8
1. Introduction	8
2. The Exponential Class of Distributions	13
3. Some Normal Theory Linear Models	17
4. Dating the Works of Plato	22
5. An Alternative Test Statistic	28
6. Tests of Separate Families	32
7. The Exponential Distribution versus the Lognormal Distribution	38
8. A Test of Equidistance	44
9. Quantal Responses	48
10. Discussion	56
References	57
Part 3 Other Publications	64

INTRODUCTION

The theory of estimation and of testing hypotheses about the values of the parameters in a model of known form is well established. The problem of determining which model describes the data adequately has, in comparison, received less attention. If there is only one model it is meaningful to calculate some measure of the agreement between the model and the data. An example is the χ^2 test of goodness of fit. With two or more models a value of χ^2 could be calculated from the fit of each model. But, since these values are not independent, the interpretation of the results may not be clear. An alternative approach is to base the choice between models on some function of all the models and the data. It is the purpose of the thesis to study this idea in greater detail.

The simplest case of choice between models is when the models do not contain adjustable parameters to be estimated from the data. This problem is discussed in Part 1, where a statistic is developed for testing whether all the models describe the data equally well.

The more important and general case of models containing adjustable parameters is the subject of the weightier Part 2. Here the approach is to combine the models into one model of which the

components are special cases. This brings the problem within the scope of standard statistical techniques so that inferences about the parameters of combination can be made in the usual way.

PART 1

THE CHOICE BETWEEN PREDICTION FORMULAE

One situation in which models without adjustable parameters occur is as a result of simplifying assumptions in engineering calculations. For each experiment there may be several predictions of the outcome, depending upon the particular assumptions made. This example may be helpful in understanding Part 1.

Biometrika (1969), 56, 2, *p.* 337
With 1 *text-figure*
Printed in Great Britain

A test for discriminating between models

BY ANTHONY C. ATKINSON
Imperial College

SUMMARY

Suppose that two or more theoretical formulae are available for predicting the value of a single valued observable response. A method is described for testing whether each formula describes the data equally well. The method is compared with other tests.

1. INTRODUCTION

Scientific experiments are sometimes performed to compare alternative theoretical models. We consider those situations in which there is a single measurable response and the models do not contain adjustable parameters. Given a set of experimental results we want to test whether the formulae differ in predictive power.

It is important to distinguish at least three questions which arise, even if there are only two models. One or more of the questions may be relevant in any particular application. The questions are:

(i) Assuming that one of the models is true, what is the evidence provided by the data as to which is the true one? The Bayesian analysis of Box & Hill (1967) assumes this question.

(ii) If one model is already in use as a predictor, is there any evidence of a departure from it in the direction of a second model? This statement of the problem, which is not symmetrical in the models, is similar to that which arises in the tests of separate families of hypotheses suggested by Cox (1961, 1962).

(iii) Is there any evidence that the models give significantly different fits to the data?

Although this paper describes one method of answering the second question, we are chiefly concerned with answering the third. One test of the hypothesis that all the models fit the data equally well would be to test the homogeneity of the estimates of error variance obtained from the sum of squared deviations between the observations and each model. But, since these estimates are not independent, a simple test of this kind is not appropriate.

Another test, based on regression, has been suggested by Williams (1959*a*, pp. 81–9). This and related work are described in §§ 3 and 4, after the necessary nomenclature has been established.

2. NOMENCLATURE

The data consist of a set of N observations. For every observation there is an estimate of the response from each of p rival formulae or models. Let the observed response for the jth experiment be y_j. Then

$$y_j = \eta_j + e_j, \tag{2·1}$$

where η_j is the true unknown value of the response and the errors e_j are assumed normally and independently distributed with zero mean and variance σ^2.

The prediction of the jth observation from the ith model is f_{ij}. The $p \times N$ matrix of these predictions is called F'.

Write $$C = F'F = \left\{ \sum_{k=1}^{N} f_{ik} f_{kj} \right\}$$

and let $$C^{-1} = \{a_{ij}\}. \tag{2·2}$$

For conciseness we often write, for example,

$$\sum_{j=1}^{N} y_j f_{ij} = \Sigma y f_i. \tag{2·3}$$

We shall be interested in linear combinations of the models. The average model $\bar{f}$ is defined by

$$\bar{f}_j = \frac{1}{p}(f_{1j} + f_{2j} + \ldots + f_{pj}). \tag{2·4}$$

For convenience, the sum of squared deviations about this model is denoted by

$$(Y\bar{F})^2 = \sum_{j=1}^{N} (y_j - \bar{f}_j)^2. \tag{2·5}$$

Another linear combination of the models can be found by treating the observations as the dependent variable and the predictions as independent variables in a multiple regression analysis. The model so obtained is called $\hat{f}$ and the fitted coefficients are represented by the $p \times 1$ vector $\hat{B}$.

A third model f^* is formed by regression on the models subject to the restriction that the regression coefficients sum to unity. Let J be a $p \times 1$ vector of ones and let $B^* = \{b_i^*\}$ be the vector of fitted coefficients. Then

$$J'B^* = \sum_{i=1}^{p} b_i^* = 1. \tag{2·6}$$

The sum of squared deviations about this model is $(YF^*)^2$. By analogy, the sum of squared deviations about the ith model is called $(YF_i)^2$.

3. TWO TESTS FOR TWO MODELS

The two tests described in this section are mentioned by Williams. Both are concerned with the choice between two models. Hoel (1947) answers the second question of § 1, whether there is evidence of a departure from f_1 in the direction of f_2. Williams & Kloot (1953) are concerned with the third question, whether each formula provides an equally good prediction. As would be expected, these two formulations lead to different test statistics.

3·1. *Hoel's test*

Hoel develops a likelihood ratio test for the hypothesis

$$H_0: \eta_j = f_{1j}$$

against the alternative $$H_1: \eta_j = f_{2j}. \tag{3·1·1}$$

The t test which results is that if

$$\frac{\Sigma(y - f_1)(f_1 - f_2)\sqrt{(N-1)}}{\sqrt{[\Sigma(y-f_1)^2 \Sigma(f_1-f_2)^2 - \{\Sigma(y-f_1)(f_1-f_2)\}^2]}} < -t_{2\alpha}, \tag{3·1·2}$$

H_0 will be rejected at the significance level α. Squaring the left-hand side of (3·1·2) and rearranging the terms we may write the test statistic as

$$(N-1)\frac{(YF_1)^2-(YF^*)^2}{(YF^*)^2}, \tag{3·1·3}$$

where F^* represents the linear combination of the models found by regression when the coefficients are constrained to sum to one. The test is thus a comparison of the sum of squares of y about f_1 with the sum of squares about f^* using the residual sum of squares from the constrained regression as an estimate of error.

If the test is significant it does not necessarily follow that the original formula should be replaced by the alternative. The rejection of the hypothesis is in the direction of f_2. The original formula f_1 could be the better of the two, although both formulae are significantly worse than some linear combination of the two.

All the tests described in this paper have the same form as Hoel's test. They are F tests of the form

$$\frac{(YF_A)^2-(YF_B)^2}{(YF_B)^2} \tag{3·1·4}$$

suitably adjusted for degrees of freedom, where $(YF_A)^2$ is the sum of squares about some hypothesized formula and $(YF_B)^2$ is about some 'best' formula found by regression.

3·2. *A test of Williams and Kloot*

Hoel's statistic tests whether f_1 should be replaced by a linear combination of the formulae. In contrast, the test of Williams & Kloot (1953) is symmetric in the two formulae. They test whether one model provides a significantly better fit to the data than does the other. Hoel's test would be appropriate if f_1 were in use for prediction and it was desired to decide whether the prediction could be improved by including f_2 in the calculations. Williams & Kloot's test is appropriate if neither formula is currently in use and it is required to choose between them for the future. The goodness of fit in this test is measured by the difference in the sums of squared deviations of the observations from the two models.

Consider

$$(YF_1)^2-(YF_2)^2 = \Sigma f_1^2-\Sigma f_2^2-2\Sigma y(f_1-f_2). \tag{3·2·1}$$

This is a linear function of the observations with variance $4\sigma^2\Sigma(f_1-f_2)^2$. The test compares

$$\frac{\{(YF_1)^2-(YF_2)^2\}^2}{4\Sigma(f_1-f_2)^2} = \frac{\{\Sigma y(f_1-f_2)-\frac{1}{2}(\Sigma f_1^2-\Sigma f_2^2)\}^2}{\Sigma(f_1-f_2)^2} \tag{3·2·2}$$

with an estimate of the variance of the observations.

If this estimate is not available, for example, from replication, an estimate may instead be based on the residual sum of squares from the regression of the observations on the predictions. Williams & Kloot use $(YF^*)^2$ because they are concerned with choosing between two formulae for interpolation in series of correlated observations. Only linear combinations of the formulae in which the coefficients sum to unity will allow for non-zero mean and trend in the observations. The use of $(YF^*)^2$ is therefore necessary in this application.

This test statistic is intuitively appealing. The sum of squared deviations of each formula is an obvious measure of how well the observations are described. By considering the difference of the two sums a direct comparison is made between the formulae and an answer obtained to the third question of § 1. Whether, if the two formulae do differ significantly

the better formula provides an adequate explanation of the data is another question which is to be answered separately, perhaps by a further statistical test.

The use of $(YF^*)^2$ for estimating error is dictated by the special situation which Williams & Kloot are considering. In the absence of special conditions an estimate based on $(Y\hat{F})^2$ is to be preferred, for f^* is a special case of $\hat{f}$. If $(YF^*)^2$ is a good estimator of the error variance, so will $(Y\hat{F})^2$ be. The converse does not hold.

The proposed statistic for general use is thus (3·2·2) divided by $(Y\hat{F})^2$ and suitably adjusted for degrees of freedom. In § 6 we show that both this statistic and that of Williams & Kloot are of the form of (3·1·4).

4. WILLIAMS'S TEST FOR MANY MODELS

In section 5·9 of his book, Williams describes a test of the homogeneity of the sums of squared residuals designed to determine whether all the models fit the data equally well. This test is derived from one of Wilks's (1946), who considered a sample x from a p-variate normal population. The likelihood ratio test was of the hypotheses that (i) the variances of the x_i are equal and (ii) the covariance of x_i and x_j are equal $(i \neq j)$.

As Williams states: '...Wilks's test, or any other test of homogeneity of variances, is not strictly applicable here, since the different sums of squares are not actually variance estimates'. Williams therefore uses the analogy only to suggest a suitable form of statistic.

Wilks's criterion, apart from factors independent of y, reduces to the ratio $(YF^*)^2/(Y\bar{F})^2$. The suggested test statistic is

$$\frac{N-p+1}{p-1}\frac{(Y\bar{F})^2-(YF^*)^2}{(YF^*)^2}. \tag{4·1}$$

This F ratio tests whether the average formula $\bar{f}$ gives a significantly worse fit to the data than the constrained least squares formula f^*, once again using $(YF^*)^2$ as an estimate of error. The divisors for the mean squares result from the facts that $\bar{f}$ contains no adjustable parameters and f^* contains $(p-1)$ parameters.

To calculate this quantity from the data we have

$$(Y\bar{F})^2-(YF^*)^2 = (Y\bar{F})^2-(Y\hat{F})^2-\{(YF^*)^2-(Y\hat{F})^2\}, \tag{4·2}$$

$$(YF^*)^2-(Y\hat{F})^2 = (\hat{B}-B^*)'\,C(\hat{B}-B^*) \tag{4·3}$$

and (Plackett, 1960, pp. 52–3)

$$\hat{B}-B^* = C^{-1}J(J'C^{-1}J)^{-1}(J'\hat{B}-1), \tag{4·4}$$

whence

$$(Y\bar{F})^2-(YF^*)^2 = -\frac{2}{p}J'F'Y+\frac{1}{p^2}J'CJ+Y'FC^{-1}F'Y-(Y'FC^{-1}J-1)(J'C^{-1}J)^{-1}(J'C^{-1}F'Y-1). \tag{4·5}$$

Two objections may be made to this test. The general use of $(YF^*)^2$ as an estimate of error has already been discussed in § 3·2. The other point is that the usefulness of the average formula as a predictor is normally not of interest.

A more appealing test would be based on the differences in the sums of squared deviations of the formulae from the observations, as was the test of Williams & Kloot for two models. It is to such a test that we now turn.

5. The tilde test

The test we propose involves a comparison of the sums of squared deviations of the formulae from the observations. These quantities are intuitively a good measure of the adequacy of the various formulae. Only if they are all equal is the non-centrality parameter of the distribution of the statistic zero. Then all the formulae describe the data equally well, or equally badly. Whether the data are adequately described could be the subject of a further statistical test. Or it could depend on other considerations. One formula might be better than the rest and quite adequate for preliminary forecasts even though it did not fit the data to within experimental error.

The test involves another linear combination of the models, which we denote by $\tilde{f}$. As it is convenient to have a name for this test, we call it after the swung dash or tilde used to distinguish it. For the reasons given in § 3·2, $(Y\hat{F})^2$ is used as an estimate of error. The test is of the form given in (3·1·4).

We require linear combinations of the models which lie equally far from each model in the sense of sums of squares. Then we have

$$\Sigma(f_i-\tilde{f})^2 = l \quad \text{(for all } i\text{)}. \tag{5·1}$$

A geometrical argument may help to explain the basis of the test. Suppose, for example, there are three formulae. Then in the N-dimensional space of the observations the formulae define a triangle. The minimum value of l, say l_0, occurs when $\tilde{f}$ is the intersection of the perpendicular bisectors of the sides of the triangle. Larger values of l generate the line through the point of intersection perpendicular to the plane of the triangle.

There are p parameters in $\tilde{f}$. For a given value of l, the relationships in (5·1) define all of these.

Let
$$T = \{\Sigma y f_i - \tfrac{1}{2}\Sigma f_i^2\} \tag{5·2}$$

and let the numerator of the test statistic be S, where

$$S = (Y\tilde{F})^2 - (Y\hat{F})^2. \tag{5·3}$$

We find that value of l which minimizes $(Y\tilde{F})^2$ by differentiation with respect to l and setting the derivative equal to zero. Geometrically, for three models, this is the value of l appropriate to the foot of the perpendicular from the observations on to the line defined above. Substitution in (5·3) yields

$$S = T'\{C^{-1} - C^{-1}J(J'C^{-1}J)^{-1}J'C^{-1}\}T. \tag{5·4}$$

We now consider the expected value of (5·4) under the hypothesis of equality of the sums of squared deviations between each model and the true response. This hypothesis may be written

$$H_0\colon \Sigma(\eta - f_i)^2 = l \quad (i = 1, 2, \ldots, p) \tag{5·5}$$

or
$$\Sigma f_i^2 - 2\Sigma \eta f_i = k \quad (i = 1, 2, \ldots, p). \tag{5·6}$$

Then from (5·2) we have
$$T = F'e - \tfrac{1}{2}Jk, \tag{5·7}$$

where e is the vector of errors. Substitution in (5·4) gives

$$E(S) = E[e'F\{C^{-1} - C^{-1}J(J'C^{-1}J)^{-1}J'C^{-1}\}F'e]. \tag{5·8}$$

Since the errors are independent, (5·8) becomes

$$E(S) = \sigma^2 \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij} \left\{ a_{ij} - \frac{\sum_{k=1}^{p} a_{ik} \sum_{k=1}^{p} a_{jk}}{\sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}} \right\}. \tag{5·9}$$

By the definition of an inverse

$$\sum_{k=1}^{p} c_{ik} a_{kj} = \begin{cases} 1 & (i=j), \\ 0 & \text{(otherwise)}. \end{cases} \tag{5·10}$$

From (5·10) we obtain $$E(S) = (p-1)\sigma^2. \tag{5·11}$$

We have assumed that the errors are normally and independently distributed. Under the null hypothesis the numerator of the test statistic will have a central χ^2 distribution on $(p-1)$ degrees of freedom, however well or badly the models fit the data. Unless special circumstances, such as those of § 3·2, suggest otherwise, we take $(Y\hat{F})^2$ to be an estimate of error. Then the tilde test, which is the ratio

$$\frac{N-p}{p-1} \frac{(Y\tilde{F})^2 - (Y\hat{F})^2}{(Y\hat{F})^2}, \tag{5·12}$$

can be tested by the F distribution with $(p-1)$ and $(N-p)$ degrees of freedom.

We now compare this test with that of Williams when there are two rival models.

6. WILLIAMS'S AND THE TILDE TESTS FOR TWO MODELS

We adopt a geometrical approach to the form of the two tests when there are only two rival models.

The models and the observations define the points F_1, F_2 and Y in N-dimensional space. All linear combinations of the two models lie in the plane containing the origin, F_1 and F_2. The foot of the perpendicular from Y to the plane is $\hat{F}$. The average model $\bar{F}$ lies half-way between F_1 and F_2. The constrained model F^* is the point on the line through F_1 and F_2 nearest to Y. Therefore $\hat{F}F^*$ is perpendicular to F_1F_2. The line through $\bar{F}$ perpendicular to F_1F_2 is the locus of models with equal sums of squared residuals about the two formulae. The point on this line nearest to Y, and so to $\hat{F}$, is $\tilde{F}$. Therefore $\hat{F}\tilde{F}$ is perpendicular to $\bar{F}\tilde{F}$. These relationships are shown in Figure 1.

It follows that $$(Y\bar{F})^2 - (YF^*)^2 = (\bar{F}F^*)^2 = (\tilde{F}\hat{F})^2 = (Y\tilde{F})^2 - (Y\hat{F})^2. \tag{6·1}$$

Thus the numerators of the two tests are identical when there are only two models. This is not, however, true for greater numbers of formulae.

It can be shown that (6·1) reduces to (3·2·2), the numerator of Williams & Kloot's test. Thus the numerators of all three tests are identical when there are only two models. Furthermore, since both tests associated with Williams use $(YF^*)^2$ as an estimate of error, they are identical and so Williams & Kloot's test is of the form of (3·1·4).

Although Williams's test and the tilde test have the same numerator, the estimate of error in the denominator is different. For the reason given in § 3·2, the tilde test seems preferable.

7. The tests under alternative hypotheses

The denominators of both tests are estimates of error. In considering the properties of the tests we assume that some satisfactory and appropriate estimate is available from replication or as a residual from regression. We consider only the numerators of the tests.

In comparing two test statistics for the same hypothesis it is customary to select the test which has the greater power against some alternative of interest. In the present case this is not possible as the two tests are concerned with different hypotheses. The choice between the tests must therefore depend on the relative importance of the adequacy of the average model and the differences in the sums of squared deviations.

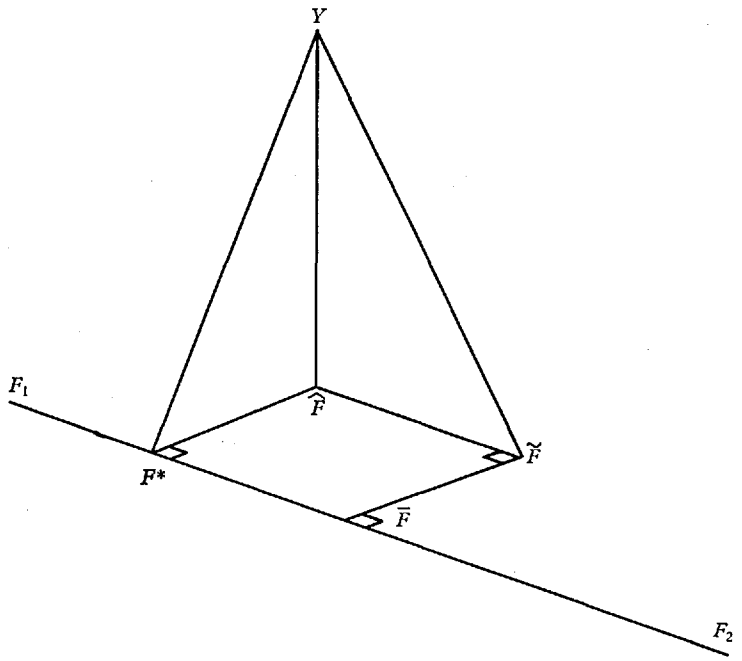

Fig. 1. Geometrical interpretation of Williams's and the tilde tests for two models. The origin is not shown. The vector of observations is Y; the models under comparison are F_1, F_2; their average is $\bar{F}$ and $\bar{F}\tilde{F}$ is the line in the plane of F_1, F_2 equidistant from F_1 and F_2.

It is, however, of interest to look at the values of the tests under departures from the null hypotheses.

Suppose for Williams's test that the true model is f^*. Then

$$\eta = FB^* \quad \text{and} \quad J'B^* = 1. \tag{7·1}$$

Substitution for Y in (4·5) yields, after taking expectations,

$$E\{(Y\bar{F})^2-(YF^*)^2\} = \left(B^*-\frac{J}{p}\right)' C\left(B^*-\frac{J}{p}\right)+E[e'F\{C^{-1}-C^{-1}J(J'C^{-1}J)^{-1}J'C^{-1}\}F'e]. \tag{7·2}$$

From (5·8) it follows that

$$E\{(Y\bar{F})^2-(YF^*)^2\}=\left(B^*-\frac{J}{p}\right)'C\left(B^*-\frac{J}{p}\right)+(p-1)\sigma^2. \tag{7·3}$$

For the non-centrality parameter to be zero each regression coefficient must have a value of $1/p$. The test measures how far the parameters depart from this value.

For the tilde test let

$$\Sigma(\eta-f_i)^2=l_i \quad (i=1,2,\ldots,p)$$

and

$$L=\{l_i\}. \tag{7·4}$$

Then, by comparison with (5·7), we have

$$T=F'e-\frac{Jk}{2}-\frac{L}{2}, \tag{7·5}$$

whence

$$E\{(Y\tilde{F})^2-(Y\hat{F})^2\}=(p-1)\sigma^2+\tfrac{1}{4}[L'\{C^{-1}-C^{-1}J(J'C^{-1}J)^{-1}J'C^{-1}\}L]. \tag{7·6}$$

Since the value of k does not enter this non-centrality parameter, it is a quadratic form in the differences among the l_i. For two formulae (7·6) reduces to

$$E\{(Y\tilde{F})^2-(Y\hat{F})^2\}=\sigma^2+\frac{(l_1-l_2)^2}{4\Sigma(f_1-f_2)^2}. \tag{7·7}$$

For Williams's test let $m=b_1^*-\frac{1}{2}=\frac{1}{2}-b_2^*$, by definition. Then (7·3) becomes

$$E\{(Y\bar{F})^2-(YF^*)^2\}=\sigma^2+m^2\Sigma(f_1-f_2)^2. \tag{7·8}$$

From the results of § 6 these two expressions are identical. The relationship between m and the differences amongst the l_i is thus

$$\left(\frac{l_1-l_2}{m}\right)^2=4\{\Sigma(f_1-f_2)^2\}^2. \tag{7·9}$$

8. A NUMERICAL EXAMPLE

As an example of the use of the proposed test we reanalyse a set of data given by Williams (1959*a*, p. 87). This consists of 33 readings of the failing loads of columns of silver quandong with predictions from three formulae for each reading. These results are summarized in Table 1.

Table 1. *Uncorrected sums of squares and products for Williams's data*

	f_1	f_2	f_3	y
f_1	1·929219	1·771904	2·008474	1·958492
f_2	—	1·628134	1·843793	1·798355
f_3	—	—	2·092331	2·039330
y	—	—	—	1·989298
$YF_i^2\times10^6$	1,533	20,722	2,969	—

The last row of Table 1 gives the sums of squared deviations from the three formulae multiplied by 10^6. From this information it seems evident that the second formula is very much poorer than the other two.

The total sum of squares of the observations may be broken into three parts; the regression sum of squares due to the equidistant formula $\bar{f}$ with 1 degree of freedom and the numerator

and denominator of the tilde test. The results of calculating this F ratio, equation (5·12), are given as an analysis of variance in Table 2. The value of 1432 is very strong evidence indeed that the three formulae do not fit the data equally well.

Table 2. *Analysis of variance for three models*

Source	D.F.	SS ($\times 10^6$)	MS ($\times 10^6$)	F
Difference between formulae $(Y\tilde{F})^2-(Y\hat{F})^2$	2	65,886	32,943	1,432
Residual $(Y\hat{F})^2$	30	690·0	23·00	

The analysis based on Williams's test gives a mean square residual of 30·36 and an F ratio of 29·61. This suggests that the tilde test is more sensitive for detecting differences between the formulae.

The large value of the test statistic is obviously due to the inadequacy of the second formula. We can now repeat the analysis to see whether the other two formulae fit the data equally well. The value of $(Y\tilde{F})^2-(Y\hat{F})^2$, formula (3·2·2), is $112{\cdot}02\times 10^{-6}$, giving an F value of 4·17 on 1 and 30 degrees of freedom. This is, to 3 significant figures, the 5 % value of F. The conclusion is that there is some quite strong, but not overwhelming, evidence that formula 1 describes the data better than formula 3.

In order to perform these calculations it was necessary to work to 10 significant figures. This is not surprising when we consider the form of the matrix $F'F$, which consists of sums of squares and products which are expected to be very similar. It does suggest that numerical problems may arise in evaluating the test statistic for large numbers of alternative formulae and that consideration should be given to reducing the ill-conditioning of the matrix before inversion is attempted, for example by the addition and subtraction of rows and columns one from another.

In this analysis of Williams's data we have considered only whether the experimental results are consistent with the hypothesis of equidistance of the three models. This is, of course, but a part of the complete analysis. As was mentioned in § 5, even if the models were equidistant from the data, all models might be unsatisfactory. A full analysis would therefore include a test of the adequacy of one or more of the models in describing the observations.

9. Hotelling's test for models with parameters

The tests described here have been concerned with models which do not contain parameters to be estimated from the data. In this section the tilde test is compared with a test developed by Hotelling (1940) for the comparison of regression variables. Hotelling's test is discussed by Healy (1955) and identically in the book and paper by Williams (1959*a*, *b*).

There are n observations on a dependent variable y. Associated with each observation there are p regressor variables $x_1, x_2, \ldots, x_p$. It is desired to choose only one of these variables for use as a predictor. The statistic proposed by Hotelling tests whether there is any evidence that the regressor variables differ in predictive power.

The test depends on the fact that the regression sum of squares associated with each x is the square of a linear function of y

Let
$$z_i = \frac{\Sigma y(x_i-\bar{x}_i)}{\sqrt{\{\Sigma(x_i-\bar{x}_i)^2\}}} \quad (i=1,2,\ldots,p). \tag{9·1}$$

Then z_i is the square root of the regression sum of squares due to x_i. Let the elements of the inverse of the sample correlation matrix be r^{ij}. Then the test statistic is

$$S' = \sum_{i=1}^{p}\sum_{j=1}^{p} r^{ij}z_iz_j - \frac{\left(\sum_{i=1}^{p}\sum_{j=1}^{p} r^{ij}z_i\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r^{ij}}. \tag{9·2}$$

If the y's are normally and independently distributed, this sum of squares with $(p-1)$ degrees of freedom provides a criterion for testing the reality of the differences amongst the z_i. Hotelling suggests using the residual from multivariate regression on all the variables as an estimate of error, thus obtaining an F ratio on $(p-1)$ and $(N-p-1)$ degrees of freedom.

This test statistic depends upon the z_i, functions of the regression sums of squares linear in y and on the inverse of the correlation matrix of the x's. Now consider the tilde test. Equation (5·4) may be written as

$$S = \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}t_it_j - \frac{\left(\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}t_i\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}}. \tag{9·3}$$

Here the a_{ij} are the elements of the inverse of the matrix $F'F$ and

$$t_i = \Sigma yf_i - \frac{\Sigma f_i^2}{2} \quad (i=1,2,\ldots,p) \tag{9·4}$$

is a function, linear in y, of the sum of squares explained by the ith formula.

From equations (9·2) and (9·3) it can be seen that Hotelling's test and the tilde test are similar in form. Both contain a function of the sum of squares explained by each model which is linear in y and the inverse of a matrix which depends on the relationship between the different models. In Hotelling's test only the angles between the models are of interest as the presence of adjustable parameters allows for scaling. Here the correlation matrix is appropriate, whereas, in the tilde test, both scale and direction are important. Finally, both tests use the residual sum of squares from an unconstrained regression on all the models as an estimate of error. It is therefore reasonable to consider the tilde test as the analogue of Hotelling's test in the absence of adjustable parameters.

I am grateful to Professor D. R. Cox for his guidance of my work on this topic. This research was supported by an IBM Fellowship.

REFERENCES

BOX, G. E. P. & HILL, W. J. (1967). Discrimination among mechanistic models. *Technometrics* **9**, 57–71.

COX, D. R. (1961). Tests of separate families of hypotheses. *Proc. Fourth Berkeley Symp.* **1**, 105–23.

COX, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B **24**, 406–24.

HEALY, M. J. R. (1955). A significance test for the difference in efficiency between two predictors. *J. R. Statist. Soc.* B **17**, 266–8.

HOEL, P. G. (1947). On the choice of forecasting formulas. *J. Am. Statist. Ass.* **42**, 605–11.

HOTELLING, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Ann. Math. Statist.* **11**, 271–83.

PLACKETT, R. L. (1960). *Principles of Regression Analysis.* Oxford University Press.
WILKS, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. Math. Statist.* **17**, 257–81.
WILLIAMS, E. J. (1959*a*). *Regression Analysis.* New York: Wiley.
WILLIAMS, E. J. (1959*b*). The comparison of regression variables. *J. R. Statist. Soc.* B **21**, 396–9.
WILLIAMS, E. J. & KLOOT, N. H. (1953). Interpolation in a series of correlated observations. *Aust. J. Appl. Sci.* **4**, 1–17.

[*Received November* 1968. *Revised February* 1969]

PART 2

A GENERAL METHOD FOR DISCRIMINATING BETWEEN MODELS

1. INTRODUCTION

1.1. The Problem

The probability density function (p. d. f.) of an observed random vector $Y = (Y_1, \ldots, Y_n)$ is unknown. A set of p. d. f. 's $f_1(\underline{y}, \underline{\theta}_1), \ldots, f_p(\underline{y}, \underline{\theta}_p)$ is under consideration for the description of the data. What inferences can be drawn as to which, if any, of the models are adequate?

In this general formulation of the problem no restriction is placed on the form of the separate p. d. f. 's nor on the associated vectors of parameters. The values of none, some or all of the parameters may be specified, the values of the remainder requiring estimation from the data. Attention is, however, confined to those situations in which each experiment results in a single valued observation. We also assume that numerical information on the prior

probabilities of each model and on the joint prior distribution of the parameters is not available, so that it is not possible to use Bayes's Theorem.

In many, but not all, cases we consider situations in which there are only two alternative models. If these both belong to the same family of distributions, the problem is one to which the Neyman-Pearson theory of hypothesis testing applies. Alternatively the models may belong to separate families in the sense that for any parameter value $\underline{\theta}_{10}$, the p.d.f. $f_1(\underline{y}, \underline{\theta}_{10})$ cannot be approximated arbitrarily closely by $f_2(\underline{y}, \underline{\theta}_2)$. An example is when the alternatives are the exponential and the log normal distributions. A theory of hypothesis testing for separate families has been developed by Cox (1961, 1962).

As an alternative to these tests, Cox suggests combining the two hypotheses in a general model of which they would both be special cases. The p.d.f. could, for example, be taken as proportional to

$$\left\{f_1(\underline{y}, \underline{\theta}_1)\right\}^{\lambda} \left\{f_2(\underline{y}, \underline{\theta}_2)\right\}^{1-\lambda} \qquad (1.1)$$

and inferences about λ made in the usual way. It is the purpose

of the present part to examine this idea in greater detail.

1.2. Some General Remarks

The customary, and often satisfactory, way of deciding which models are adequate is to treat each model in isolation, analysing the residuals or computing some measure of fit such as χ^2 for each alternative. With only one model the value of χ^2 has a clear interpretation. But, with several models, the calculated values of χ^2 will not be independent. Any conclusions should take account of this, although, in extreme cases, the correct conclusion will be clear.

A closely related alternative for two models is to calculate the ratio of maximized likelihoods. But, if the models belong to separate parametric families, the interpretation of this quantity is not always unambiguous. If one model contains more adjustable parameters than the other, the likelihood ratio will be biased in its favour. Thus, in using the ratio, the difference in the number of parameters needs to be considered. On a more formal level, it is not possible to interpret the ratio in terms of the chi-squared distribution.

Such difficulties recommend the study of the properties of

combinations of distributions. An alternative to the exponential combination (1.1) is the mixture distribution

$$\lambda f_1(\underline{y}, \underline{\theta}_1) + (1-\lambda) f_2(\underline{y}, \underline{\theta}_2) \, . \tag{1.2}$$

One disadvantage of this linear combination is that, for λ less than zero or greater than one, it can lead to negative probabilities. For discrete distributions with a finite number of classes, the exponential combination cannot lead to such anomalies, although, for some continuous distributions, not all values of λ are admissible. Since the exponential combination is additive in log likelihoods, it also has the advantage of greater mathematical convenience for many frequently occurring distributions.

1.3. Plan of the Second Part

Although in particular applications there may be other methods of combination which are physically meaningful, we shall be concerned only with combinations of the general form of (1.1). In the next section we consider some properties of the combined p.d.f.. Thereafter we are concerned with making inferences about λ. Testing the hypothesis that the value of λ is zero or one is equivalent to testing for departures from one model in the direction of the other. The hypothesis $\lambda = {}^1/2$ implies that both models are equidistant from the data, where distances are measured in the manner appropriate

to the definition of the combined p. d. f. In this special sense, the hypothesis implies that both models fit the data equally well.

In the first two examples of the use of the method, both the component and combined p. d. f.'s are normal and the resulting tests are based on the F distribution. Usually the resulting distributions are not so tractable. In § 4 the test is based on the asymptotic distribution of the likelihood ratio. In general this involves the simultaneous estimation of the values of λ and the parameters of the component distributions which maximize the likelihood. A test statistic involving reduced computation is developed in § 5 and applied in the following section to two component p. d. f.'s belonging to separate families. The resulting statistic is shown to be asymptotically equivalent to Cox's test of separate families of hypotheses. § 7 is devoted to examining the small sample differences between the two tests. The two final applications of the procedure are to a test of the hypothesis that $\lambda = {}^{1}/_{2}$ and to the analysis of binary data.

2. THE EXPONENTIAL CLASS OF DISTRIBUTIONS

The general form of the combined p. d. f. is proportional to

$$\prod_{i=1}^{p} \left\{ f_i(\underline{y}, \underline{\theta}_i) \right\}^{\lambda_i} . \tag{2.1}$$

In order for (2.1) to have the properties of a density, a normalizing constant has to be introduced. Thus, for the combination of two models we write

$$f_\lambda(\underline{y}) = \frac{\left\{ f_1(\underline{y}, \underline{\theta}_1) \right\}^{\lambda_1} \left\{ f_2(\underline{y}, \underline{\theta}_2) \right\}^{\lambda_2}}{\int \left\{ f_1(\underline{z}, \underline{\theta}_1) \right\}^{\lambda_1} \left\{ f_2(\underline{z}, \underline{\theta}_2) \right\}^{\lambda_2} d\underline{z}} \tag{2.2}$$

where, for discrete variables, the integration in the denominator is replaced by a summation.

In many cases this form of the model contains redundant parameters. For two models three distinct situations can occur:

1. Both λ_1 and λ_2 can be estimated.
2. Only one of the λ's can be estimated, either because (2.2) reduces to the form given in § 1, or because one of the λ's is redundant.
3. Neither λ can be estimated. The combined p. d. f. in this case has only one general form and two special cases corresponding to the constituent distributions.

Which of these three possibilities occurs depends both upon

the form of the distributions being combined and on whether the values of the parameters are specified or are to be estimated.

To examine this point further we consider the general exponential class of distributions, reserving the more usual word family to describe p. d. f.'s differing only in the values of the parameters. For simplicity we discuss only the case in which the variables y_j are independently and identically distributed, the constituent distributions each containing one parameter θ_i. We write

$$f_i(y_j, \theta_i) = \exp\left\{A_i(\theta_i)B_i(y_j) + C_i(\theta_i) + D_i(y_j)\right\}. \qquad (2.3)$$

Raising this p.d.f. to the power λ_i results in multiplication of the exponent by λ_i. Because of the normalizing factor in the denominator of (2.2), the combined p. d. f. is not a function of $C_i(\theta_i)$.

For some of the more frequently encountered distributions of the exponential class, $D_i(y_j)$ is zero. Such distributions are reproductive with respect to exponentiation. If in such cases $\lambda_i A_i(\theta_i)$ can take on all the values which $A_i(\theta_i)$ can, then it is not possible to estimate λ_i. But if $D_i(y_j)$ is not zero or if the value of θ_i is specified, then λ_i may be estimated. Examples of distributions which are reproductive with respect to exponentiation are, for continuous variables, the normal and the exponential, for discrete variables the geometric and for quantal response data, the logistic. Problems involving all

these distributions are considered in the following sections.

As an example of the ideas discussed in this section we look at the problem given by Cox (1962) of determining whether a set of observations comes from the Poisson or geometric distributions. We have for the component p. d. f.'s

$$f_1(y, \theta_1) = \frac{e^{-\alpha}}{y!}\ \alpha^y$$

and $$f_2(y, \theta_2) = \frac{\beta^y}{(1+\beta)^{y+1}} \quad . \tag{2.4}$$

These may be rewritten in the exponential class form as

$$f_1(y, \theta_1) = \exp\left\{-\alpha + y \log \alpha - \log(y!)\right\}$$

$$f_2(y, \theta_2) = \exp\left\{y \log\left(\frac{\beta}{1+\beta}\right) - \log(1+\beta)\right\}. \tag{2.5}$$

For the Poisson distribution, $D_1(y)$ is non zero, so that λ_1 can be estimated. For the geometric distribution $D_2(y)$ is zero and, if β is only constrained to be non-negative,

$$A_2(\theta_2) = \log\left(\frac{\beta}{1+\beta}\right)$$

can take any real value. Therefore λ_2 cannot be estimated. The combined p. d. f. can be written as

$$f_\lambda(y, \underline{\theta}) = \frac{\frac{\gamma^y}{(y!)^\lambda}}{\sum_{i=0}^{\infty} \frac{\gamma^i}{(i!)^\lambda}} \quad . \tag{2.6}$$

For $\lambda = 1$ this reduces to the Poisson distribution and for $\lambda = 0$ to the geometric. In order for λ_2 to be estimable it is necessary that values of both α_o and β_o should be specified. An example of testing a hypothesis about the value of λ for this combined p.d.f. is given in § 8.

3. SOME NORMAL THEORY LINEAR MODELS

The combined p. d. f. is of interest as a means of making inferences about the adequacy of the component models by testing hypotheses about the values of the λ's. In this section we consider two examples in which both the component and combined distributions are normal. Although it is not possible to obtain estimates of the λ's, it is possible to test some hypotheses about the values of the parameters.

3.1. The Selection of Regression Variables

Suppose that there are two sets of regression variables. In practice these sets may be overlapping and both contain many variables. For simplicity we suppose that there are only two variables x_1 and x_2. The component p. d. f.'s are normal, means b_1x_1 and b_2x_2 and with common variance σ^2. If we write

$$c_1 = \frac{b_1 \lambda_1}{\lambda_1 + \lambda_2}, \quad c_2 = \frac{b_2 \lambda_2}{\lambda_1 + \lambda_2} \quad \text{and} \quad \tau^2 = \frac{\sigma^2}{\lambda_1 + \lambda_2} \tag{3.1.1}$$

the combined p. d. f. (2.2) becomes

$$f_\lambda(y, \underline{\theta}) \propto \exp\left\{-\frac{1}{2}\frac{(y - c_1x_1 - c_2x_2)^2}{\tau^2}\right\} \tag{3.1.2}$$

which is the regression of y on x_1 and x_2.

The hypothesis that regression on x_1 alone is adequate is the same as the hypothesis $\lambda_2 = 0$. From (3.1.1) the likelihood ratio test of this hypothesis compares the difference in the sum of squares

due to regression on x_1 alone and on both x_1 and x_2 with some suitable estimate of error. Use of the combined p.d.f. thus leads to the standard regression procedure. Owing to overparameterization it is not possible to develop a test of equidistance from the two models expressed as the hypothesis $\lambda_1 = \lambda_2$.

3.2. The Choice of a Prediction Formula

Suppose that, from theoretical considerations or some other source, p formulae are available for predicting the value of a random variable. As a result of n observations it is desired to test whether the formulae vary in predictive power. This situation has been discussed by Williams (1959, pp. 83-9) and by Atkinson (1969).

It is assumed that the observations are independently and normally distributed with constant variance. There are p component p.d.f.'s

$$f_i(y_j) \propto \exp\left\{ -\frac{1}{2} \frac{(y_j - f_{ij})^2}{\sigma^2} \right\} \qquad i = 1, 2, \ldots, p \qquad (3.2.1)$$

where f_{ij} is the prediction of the response for the jth experiment from the ith formula. Writing

$$\kappa_i = \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i , \qquad (3.2.2)$$

$$\text{whence} \quad \sum_{i=1}^{p} \kappa_i = 1 , \qquad (3.2.3)$$

$$\text{and} \qquad \tau^2 = \sigma^2 \Big/ \sum_{i=1}^{p} \lambda_i \tag{3.2.4}$$

the combined p. d. f. is

$$f_\lambda(y_j) \propto \exp\left\{ -\frac{1}{2\tau^2} \left(y_j - \sum_{i=1}^{p} \kappa_i f_{ij}\right)^2 \right\} . \tag{3.2.5}$$

The mean of this normal distribution is a linear combination of the predictions with coefficients summing to unity.

Under the null hypothesis that each formula provides an equally good explanation of the data, all the λ's have the same value. The likelihood ratio test of this hypothesis is a comparison of the difference in the residual sum of squares about the combined p. d. f. (3. 2. 5) when all the λ's are identical and when they are free to vary. The latter quantity is found by regressing the observations on the predictions subject, from (3. 2. 3), to the constraint that the regression coefficients sum to unity. Let the resulting linear combination of the predictors be f^{κ}_j. The required sum of squares can be written as

$$(YF^{\kappa})^2 = \sum_{j=1}^{n} (y_j - f^{\kappa}_j)^2 . \tag{3.2.6}$$

From (3. 2. 2) , under the null hypothesis, all the κ's have the value $1/p$. The mean of the combined p. d. f. is therefore

$$\bar{f}_j = \frac{1}{p} \sum_{i=1}^{p} f_{ij} \qquad (3.2.7)$$

where $\bar{f}_j$ is the average predicted response for the jth observation. The sum of squared residuals about this model is written as

$$(Y\bar{F})^2 = \sum_{j=1}^{n} (y_j - \bar{f}_j)^2 . \qquad (3.2.8)$$

If τ^2 is to be estimated from the data, the likelihood ratio test of the hypothesis that all the λ's have the same value leads to the statistic

$$\frac{n-p+1}{p-1} \frac{(Y\bar{F})^2 - (YF^*)^2}{(YF^*)^2} \qquad (3.2.9)$$

which, under the null hypothesis, has the F distribution on p-1 and n-p+1 degrees of freedom.

This test statistic was derived by Williams by analogy with a test developed by Wilks for testing hypotheses about p-variate normal distributions. The present derivation provides a maximum likelihood basis for the statistic.

If there are only two models the combined p. d. f. (3.2.5) may be rewritten as

$$f_\lambda(y_j) \propto \exp\left[-\frac{1}{2\tau^2}\left\{y_j - \lambda f_{ij} - (1-\lambda) f_{2j}\right\}^2\right], \qquad (3.2.10)$$

a one parameter model of the form suggested in § 1. With two models we can test not only whether each model fits the data equally well ($\lambda = {}^{1}/_{2}$), but also whether there is evidence of a departure from one model in the direction of the other. To test the hypothesis $\lambda = 1$, that the first model is adequate, we have the test statistic

$$(n-1)\ \frac{(YF_1)^2 - (YF^{*})^2}{(YF^{*})^2}\ , \qquad (3.2.11)$$

where

$$(YF_1)^2 = \sum_{j=1}^{n} (y_j - f_{1j})^2 \qquad (3.2.12)$$

is the sum of squared residuals about the first model.

This test statistic was originally derived by Hoel (1947) and obtained in the form of (3.2.11) by Atkinson. It tests whether there is a linear combination of the two formulae, with coefficients summing to unity, which provides a significantly better fit to the data than does the first formula. Similarly the test of $\lambda = 0$ is equivalent to testing whether there are departures from the second formula. Both of these test statistics are asymmetrical in the formulae. If it were required to test whether the formulae fitted the data equally well, Williams's test of $\lambda = {}^{1}/_{2}$ would be appropriate.

4. DATING THE WORKS OF PLATO

4.1. Introduction

In the preceding section it was not possible to estimate the values of the λ's because of the overparameterization of the combined p.d.f.. It was, however, possible to obtain test statistics with a known and tabulated distribution. In this section we consider an example involving the combination of two p.d.f.'s where both λ_1 and λ_2 can be estimated. But, in order to test hypotheses about these parameters, it is necessary to use the asymptotic distribution of the log ratio of maximized likelihoods.

The example concerns a problem in the chronology of Plato's works which was discussed by Cox and Brandwood (1959), whose general approach we follow. The problem is that between writing the Republic (R) and the Laws (L), Plato wrote several shorter dialogues, the order of five of these being uncertain. The works in question are the Critias (C), Philebus (F), Politicus (P), Sophist (S) and Timaeus (T).

In an attempt to order these seven works, philologists have studied the distribution of long and short syllables at the ends of the sentences. Only the last five syllables are counted, the sentence endings thus being broken up into 32 classes. The statistical problem

is to use these observed frequencies to order the works in decreasing affinity with R and increasing affinity with L. Provided Plato's literary style changed monotonically with time, the resultant ordering of the works will be an estimate of the order in which they were written.

4.2. Theory

We assume that the frequency distribution of sentence endings is multinomial with 32 classes and that the probabilities associated with each class for the reference populations R and L are known without error. Let these probabilities for the ith class be θ_i and ϕ_i respectively. We take the combined probability as

$$\theta_{\lambda i} = \frac{\theta_i^{\lambda_1} \phi_i^{\lambda_2}}{\sum_{i=1}^{32} \theta_i^{\lambda_1} \phi_i^{\lambda_2}} \qquad i = 1, 2, \ldots, 32 . \tag{4.2.1}$$

The log likelihood of the observations is, ignoring a constant term,

$$L = \sum_{i=1}^{32} n_i \log \theta_{\lambda i} , \tag{4.2.2}$$

where n_i is the observed number of sentences having an ending in the ith class for the particular shorter work.

To date the work we obtain estimates of λ_1 and λ_2 by maximizing

(4.2.2). This approach differs from that of Cox and Brandwood who considered only the one parameter family formed by the constraint $\lambda_1 + \lambda_2 = 1$. They thus forced the combined probability to lie on the line in Λ space through the two reference populations. Use of the two parameter form enables us to determine whether one or both of the reference populations are irrelevant.

A further difference is that instead of estimating the value of the parameter directly, they used a sufficient statistic, which is equivalent to working with a monotone function of λ.

To test hypotheses about the values of the λ's we use the likelihood ratio test. Let the likelihood contain p parameters, the values of ν of which are specified by the hypothesis under test. Call the ratio of the restricted to the overall maxima of the likelihood l. This will have a value less than unity. If, as in the example of this section, the exact distribution of l is unknown, we use the result that, under the null hypothesis, -2 log l is asymptotically distributed as χ^2 on ν degrees of freedom (Kendall and Stuart, 1961, pp. 230-1).

4.3. Results

We first test whether the reduced model used by Cox and Brandwood provides an adequate explanation of the data. In this model the combined probability is

$$\theta_{\lambda i} = \frac{\theta_i^{1-\lambda}\,\phi_i^{\lambda}}{\sum_{i=1}^{32} \theta_i^{1-\lambda}\,\phi_i^{\lambda}} . \qquad (4.3.1)$$

Table 1

Maximum likelihood estimates of the parameters in this model and the full model (4.2.1) are given in Table 1. The last column is minus twice the log likelihood ratio which, under the hypothesis $\lambda_1 + \lambda_2 = 1$, would be distributed as χ^2_1. The results indicate very strongly that the works are not adequately described by being assumed to lie between R and L. In fact the results suggest that R is irrelevant and that the only meaningful ranking is in order of similarity to L. For the data on which this analysis is based see Table 1 of Cox and Brandwood.

To test the hypothesis that the works are not related to R, the likelihood was maximized with λ_1 equal to zero. The resulting estimates of λ_2 are given in Table 2. These are close to the estimates

Table 2

for the two parameter model given in Table 1. Comparing the likelihoods for these two maximizations, the value of χ^2 on five degrees of freedom is 4.51. This is in excellent agreement with the hypothesis that it

is only possible to order the works in order of similarity to L.

4.4. Discussion

The assessment of the importance of these results must be left to those with greater knowledge of Greek philosophy. We consider only the statistical aspects of the analysis.

In order to place a work with reference to L and R it is necessary that the parent distributions are distinguishable and that neither is uniform. For suppose that $\theta_i = \theta$, all i. Then in (4.2.1) $\theta_{\lambda i}$ would not depend on θ or λ_1, and the only possible ordering is in similarity to the second population. Thus, for good discrimination, we require distributions with high, or low, frequencies for different classes.

If all 32 sentence endings had the same probability of occurrence, the average frequency would be 3.125 °/o. In the reference populations only 4 classes of ending occur with frequencies greater than 6 °/o and none with less than $^1/_2$ °/o. Of these four, three are in L. The frequencies in these classes for the reference populations and the shorter works are given in Table 3.

Table 3

These frequencies appear to explain most of the observed

features of the data. The only class in R with a high frequency occurs with a low frequency in the other six works. As a result λ_1 would be expected to be near zero, and R will be irrelevant as a reference population. The endings that occur with a high frequency in L occur with near average frequency in R. In S, C, P and F these endings occur with increasing frequency, whereas in T they occur at frequencies very near the average. The value of λ_2 for T is therefore also near zero. The fact that the ordering we have obtained of R, T, S, C, P, F, L is the same as that obtained by Cox and Brandwood is explained by the fact that most of the frequencies in R fluctuate closely about the value of $3.125^{o}/o$. Including R in the analysis does not appreciably alter the values of the θ_{λ_i} for the various classes.

5. AN ALTERNATIVE TEST STATISTIC

5.1. The Need for a Simpler Test

In order to use the log likelihood ratio as a test statistic it is necessary to perform two maximizations of the likelihood, one overall and one under the hypothesis being tested. In the analysis of the data from Plato's works with two parameters and a discrete distribution with 32 classes, the arithmetic involved in the maximizations was not too heavy. But consider testing hypotheses about the values of the λ's in a combined p. d. f. of the form of (2. 2) in which both component distributions contain a vector of parameters requiring estimation. The calculation of maximum likelihood estimates of λ_1 and λ_2 would involve a multivariable function maximization with, except in fortunate cases, a complicated numerical integration at each point at which the function is evaluated. Even with powerful computing facilities, such a problem could be formidable. We therefore consider an alternative to the maximum likelihood ratio test which involves appreciably less computation.

5.2. An Asymptotically Normal Test Statistic

The parameters of the combined p. d. f. (2. 2) are divisible into a vector $\underline{\lambda}$ about which it is desired to test some hypothesis and a vector $\underline{\theta}$ related to the parameters of the component p. d. f.'s. The

values of these parameters are not specified by the hypothesis to be tested and they are therefore nuisance parameters.

In the remaining sections of the paper we shall be concerned with combinations of two distributions which only contain one parameter λ. An asymptotically normal statistic for testing hypotheses about the value of a single parameter in the presence of nuisance parameters was suggested by Bartlett (1953) and, independently, by Neyman (1959). For ease of exposition we assume that L, the log likelihood of the observations, contains only the scalar nuisance parameter θ.

$$\text{Let } I_{11} = -E\left(\frac{\partial^2 L}{\partial \lambda^2}\right), \quad I_{12} = -E\left(\frac{\partial^2 L}{\partial \theta \partial \lambda}\right)$$

$$\text{and } I_{22} = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right), \qquad (5.2.1)$$

where the derivatives are evaluated and expectations calculated under the null hypothesis, using any locally $\sqrt{n}$ consistent estimate of θ (Neyman, 1959). The proposed test statistic is

$$T = \frac{\frac{\partial L}{\partial \lambda} - \frac{I_{12}}{I_{22}} \frac{\partial L}{\partial \theta}}{\sqrt{\left(I_{11} - \frac{I_{12}^2}{I_{22}}\right)}}. \qquad (5.2.2)$$

This is the derivative of the log likelihood with respect to the parameter of interest adjusted for regression on the partial derivative with respect to the nuisance parameter, all divided by the appropriate

standard error. The adjustment for regression ensures that the statistic is asymptotically unbiased for all ~~admissible~~ allowable estimates of θ. Under the null hypothesis the statistic will be asymptotically distributed with the standard normal distribution.

In most applications we shall use maximum likelihood estimates of the nuisance parameter. Using the value of θ which maximizes L as a function of both θ and λ provides a test which is asymptotically equivalent to the χ^2 test based on the likelihood ratio, and involves as much computation. An alternative is to use the maximum likelihood estimate of θ under the null hypothesis. If this is that λ is zero or one, we are concerned only with the component distributions, not with the combination. One result of using this estimate of θ is that the calculated value of the second term of the numerator of (5.2.2) is zero.

In § 8 we test the hypothesis $\lambda = 1/2$. The calculation of the maximum likelihood estimate of the nuisance parameter under this null hypothesis is usually not trivial. We therefore make use of the broader class of estimates which are locally $\sqrt{n}$ consistent. For discussion of this idea and of optimality properties of the test procedure see Neyman (1959).

The extension of the method to a vector of m nuisance parameters is straightforward. The numerator of the statistic is the

multiple regression of $\partial L/\partial\lambda$ on the set of partial derivatives $\partial L/\partial\theta_i$, $i = 1, 2, \ldots, m$. To calculate the regression coefficients we write

$$X'Y = \left\{ - E \left(\frac{\partial^2 L}{\partial\lambda \, \partial\theta_i} \right) \right\}$$

$$\text{and} \qquad X'X = \left\{ - E \left(\frac{\partial^2 L}{\partial\theta_i \, \partial\theta_j} \right) \right\} \quad , \tag{5.2.3}$$

when, in the standard notation of regression analysis, the vector of coefficients is $(X'X)^{-1} X'Y$. For $m = 1$ this reduces to the statistic given in (5.2.2).

A further extension of the theory of these tests is to cases in which there is a vector of parameters $\underline{\lambda}$. Instead of the asymptotically normal statistic (5.2.2), the test statistic has an asymptotic χ^2 distribution. Since in what follows we are only concerned with examples in which λ is scalar, we do not here develop the theory of such tests.

6. TESTS OF SEPARATE FAMILIES

Having obtained a convenient form of test statistic, we now apply the results of § 5 to tests of hypotheses about constituent p. d. f.'s which belong to separate families.

We assume that the observations are independently and identically distributed. Let the two constituent p. d. f.'s be $f(y, \underline{\alpha})$ and $g(y, \underline{\beta})$. The combined p. d. f. is of the one parameter form

$$f_\lambda(y) = \frac{\left\{f(y, \underline{\alpha})\right\}^\lambda \left\{g(y, \underline{\beta})\right\}^{1-\lambda}}{\int \left\{f(z, \underline{\alpha})\right\}^\lambda \left\{g(z, \underline{\beta})\right\}^{1-\lambda} dz} \qquad (6.1)$$

We are interested in testing the hypothesis $\lambda = 1$. That is, we are testing for departures from the first model in the direction of the second. For ease of exposition we assume that α & β are scalar parameters.

The log likelihood of one observation is

$$L = \lambda \log f(y, \alpha) + (1-\lambda) \log g(y, \beta)$$

$$- \log \int \left\{f(z, \alpha)\right\}^\lambda \left\{g(y, \beta)\right\}^{1-\lambda} dz \qquad (6.2)$$

To form the test statistic we differentiate (6.2) with respect to the parameters α, β and λ. First, we introduce the notation of Cox (1962) and let

$$F_i = \log f(y, \alpha) , \quad F_{i,\alpha} = \frac{\partial}{\partial \alpha} \{\log f(y, \alpha)\} ,$$

$$G_i = \log g(y, \beta) , \quad F = \log f(z, \alpha) \text{ etc.} \tag{6.3}$$

with $$\int H(z)\, f(z, \alpha) dz = E_\alpha(H) . \tag{6.4}$$

Then on the usual assumption that differentiation with respect to the parameters commutes with integration we obtain, upon evaluation at $\lambda = 1$,

$$\frac{\partial L}{\partial \lambda} = F_i - G_i - E_\alpha (F-G) \tag{6.5}$$

$$\frac{\partial L}{\partial \alpha} = F_{i,\alpha} \tag{6.6}$$

and $$\frac{\partial L}{\partial \beta} = 0 . \tag{6.7}$$

If V_α and C_α stand for variance and covariance under the null hypothesis, as E_α stands for expectation in (6.4), substitution in (5.2.2) and summation over the sample yields the test statistic

$$T = \frac{\sum_{i=1}^{n} \left\{ F_i - G_i - E_\alpha(F-G) - \frac{C_\alpha(F-G, F_\alpha)}{V_\alpha(F_\alpha)} F_{i,\alpha} \right\}}{\sqrt{\left[n \left\{ V_\alpha(F-G) - \frac{C_\alpha^2(F-G, F_\alpha)}{V_\alpha(F_\alpha)} \right\} \right]}} \tag{6.8}$$

To use (6.8) it is necessary to estimate α and β under the null hypothesis $\lambda = 1$. The estimate of α is the customary maximum likelihood estimator $\hat{\alpha}$ satisfying

$$\sum_{i=1}^{n} F_{i,\hat{\alpha}} = 0 . \qquad (6.9)$$

The estimate of β is not determined by the theory of § 5, for, when $\lambda = 1$, the likelihood is independent of β. Substitution of any preassigned value of β in (6.8) yields a statistic which is asymptotically standard normal. But the purpose of the statistic is to test for departures from the first distribution in the direction of the second. We therefore use as an estimate of β that value which best describes the data under the null hypothesis. This quantity, β_{α}, is the limit to which $\hat{\beta}$ converges when the null hypothesis is true, where $\hat{\beta}$ is the maximum likelihood estimate when $\lambda = 0$. Since α is estimated, we replace this value by $\beta_{\hat{\alpha}}$ in calculating the test statistic. From (6.9) we can write the numerator of (6.8) in an obvious extension of our notation as

$$T'(\beta_{\hat{\alpha}}) = \sum_{i=1}^{n} \left[F_i(\hat{\alpha}) - G_i(\beta_{\hat{\alpha}}) - E_{\hat{\alpha}} \left\{ F(\hat{\alpha}) - G(\beta_{\hat{\alpha}}) \right\} \right] . \qquad (6.10)$$

This expression is of order $\sqrt{n}$ in probability.

As written in (6.8) the test statistic is identical with that developed by Cox (1962) for tests of separate families of hypotheses. But Cox defines G_i as $\log g(y, \hat{\beta})$. The numerator of his test statistic is thus

$$T'(\hat{\beta}) = \sum_{i=1}^{n} \left[F_i(\hat{\alpha}) - G_i(\hat{\beta}) - E_{\hat{\alpha}} \left\{ F(\hat{\alpha}) - G(\beta_{\hat{\alpha}}) \right\} \right] . \qquad (6.11$$

Under the null hypothesis, the statistic with (6.10) as numerator would have zero expectation if the true value of α were known. Cox's statistic has this property only asymptotically as $\hat{\beta} \to \beta_{\hat{\alpha}}$. If the value of α is estimated both statistics will be biased, but the bias of $T(\beta_{\hat{\alpha}})$ will be the less. That the two are asymptotically equivalent is shown by considering the difference

$$T'(\beta_{\hat{\alpha}}) - T'(\hat{\beta}) = \sum_{i=1}^{n} \left\{ G_i(\hat{\beta}) - G_i(\beta_{\hat{\alpha}}) \right\} . \qquad (6.12$$

Expanding $G_i(\beta_{\hat{\alpha}})$ in a Taylor's series about $\hat{\beta}$, (6.12) becomes

$$T'(\beta_{\hat{\alpha}}) - T'(\hat{\beta}) = (\beta_{\hat{\alpha}} - \hat{\beta}) \sum_{i=1}^{n} G'_i(\hat{\beta})$$

$$- \frac{1}{2} (\beta_{\hat{\alpha}} - \hat{\beta})^2 \sum_{i=1}^{n} G_i''(\hat{\beta}) + \ldots \qquad (6.1$$

By the definition of the maximum likelihood estimate $\hat{\beta}$, the first term of (6.13) is zero. We also have the asymptotic result that

$$\mathrm{var}(\hat{\beta}) = \frac{-1}{\sum_{i=1}^{n} G_i''(\hat{\beta})} , \qquad (6.1$$

so that the second term of (6.13) is of order 1 in probability. Since the two numerators are of order $\sqrt{n}$ in probability, their asymptotic

equivalence is demonstrated. In the next section we consider the small sample differences between the two statistics.

We have so far developed a test for departures from $f(y, \alpha)$ in the direction of $g(y, \beta)$ by considering the hypothesis $\lambda = 1$. A test of the hypothesis $\lambda = 0$ is developed in an entirely analogous way, with the roles of the two distributions being interchanged.

There are two main generalizations of (6.8). If the observations are not identically distributed, the argument proceeds in a similar way to give equation 18 of Cox (1962). The other main generalization is if $\underline{\alpha}$ is a vector parameter: the previous argument applies unaltered if $\underline{\beta}$ is a vector.

We suppose for simplicity that the observations are identically distributed. The test statistic involves the multiple regression of (6.5) on the set of p partial differentials of the form of (6.6) where p is the number of nuisance parameters. Let

$$F_{i,\alpha_j} = \frac{\partial}{\partial \alpha_j} \left\{ \log f(y, \underline{\alpha}) \right\} \tag{6.15}$$

be the elements of the 1 x p vector X_i. If we retain the convention in which subscript i's refer to observed quantities, we have, by analogy with (5.2.3) that

$$X'X = \left\{ C_\alpha \left(F_{\alpha_j}, F_{\alpha_k} \right) \right\}$$

$$\text{and} \qquad X'Y = \left\{ C_{\alpha}(F\text{-}G,\ F_{\alpha_j}) \right\}. \tag{6.16}$$

The numerator of the test statistic is now written as

$$\sum_{i=1}^{n} \left\{ F_i\text{-}G_i\text{-}E_{\alpha}(F\text{-}G) - X_i(X'X)^{-1}\ X'Y \right\} \tag{6.17}$$

with variance

$$n \left\{ V_{\alpha}(F\text{-}G) - Y'X\ (X'X)^{-1}\ X'Y \right\}. \tag{6.18}$$

Note that (6.17) is the correct version of equation (20) of Cox (1962) which is missing some terms. In the applications in the paper, the correct form has however been employed.

It would be possible, in principle, to write down the general test statistic when $\underline{\alpha}$ is a vector and the observations are not identically distributed. This results in even more rebarbative notation, whilst involving no new ideas. We proceed instead to an example of the use of the test statistic.

7. THE EXPONENTIAL DISTRIBUTION VERSUS THE LOG NORMAL DISTRIBUTION

7.1. Theory

The purposes of this section are twofold: firstly to give an example of the use of the test statistic and secondly to study the small sample properties of the two forms which arose in the previous section. As an example we test the hypothesis that the distribution is exponential against the alternative that it is log normal. Thus we have

$$f(y,\alpha) = \frac{1}{\alpha} e^{-y/\alpha} \qquad (7.1.1)$$

and

$$g(y,\underline{\beta}) = \frac{1}{y(2\pi\beta_2)^{1/2}} \exp\left\{-\frac{(\log y-\beta_1)^2}{2\beta_2}\right\}. \qquad (7.1.2)$$

For a full discussion of this problem see Cox (1962) and Jackson (1968). The estimates of the parameters that will be required are

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i \,, \quad \hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} \log y_i$$

and

$$\hat{\beta}_2 = \frac{1}{n}\sum_{i=1}^{n} (\log y_i - \hat{\beta}_1)^2 \,. \qquad (7.1.3)$$

We defined $\underline{\beta}_{\hat{\alpha}}$ as the limit to which $\hat{\underline{\beta}}$ converged under the null hypothesis. Thus

$$\beta_{1,\alpha} = E_\alpha(\log y) = \log \alpha + \psi(1)$$

and $$\beta_{2,\alpha} = V_\alpha(\log y) = \psi'(1), \qquad (7.1.4)$$

where $\psi(x) = \frac{d \log \Gamma(x)}{dx}$, etc. For tabulation of these functions and recurrence relationships see Abramowitz and Stegun (1965, § 6).

Consider first the numerator of the test statistic (6.8). If we do not specify which estimate of $\underline{\beta}$ we employ we can write

$$\frac{T'(\underline{\beta})}{n} = \frac{\sum\left[F_i(\hat{\alpha}) - G_i(\underline{\beta}) - E_{\hat{\alpha}}\left\{F(\hat{\alpha}) - G(\beta_{\hat{\alpha}})\right\}\right]}{n} \qquad (7.1.5)$$

$$= \hat{\beta}_1 - \beta_{1,\hat{\alpha}} + \frac{1}{2}\log\left(\frac{\beta_2}{\beta_{2,\hat{\alpha}}}\right) + \frac{1}{n}\sum_{i=1}^{n}\frac{(\log y_i - \beta_1)^2}{2\beta_2} - \frac{1}{2}. \qquad (7.1.6)$$

This expression does not involve the value of $\hat{\alpha}$ directly because the exponential distribution is reproductive with respect to exponentiation. The asymptotic variance of (7.1.6) is

$$\frac{1}{n}\left[\psi'(1) - \frac{1}{2} + \frac{\psi''(1)}{\psi'(1)} + \frac{\psi'''(1)}{4\{\psi'(1)\}^2}\right]$$

$$= \frac{0.2834}{n}. \qquad (7.1.7)$$

If we estimate $\underline{\beta}$ by $\underline{\beta}_{\hat{\alpha}}$ we obtain

$$\frac{1}{n}T'(\beta_{\hat{\alpha}}) = \hat{\beta}_1 - \beta_{1,\hat{\alpha}} + \frac{1}{2\psi'(1)}\left\{\hat{\beta}_2 - \beta_{2,\hat{\alpha}} + (\hat{\beta}_1 - \beta_{1,\hat{\alpha}})^2\right\}, \qquad (7.1.8)$$

whereas, following Cox, the test statistic is

$$\frac{1}{n} T'(\hat{\beta}) = \hat{\beta}_1 - \beta_{1,\hat{\alpha}} + \frac{1}{2} \log \left(\frac{\hat{\beta}_2}{\beta_{2,\hat{\alpha}}} \right) . \qquad (7.1.9)$$

To test the hypothesis that the observations are a random sample from the exponential distribution we compare the ratio of (7.1.8) or (7.1.9) to the square root of (7.1.7) with the standard normal distribution. Large negative values of the statistic are evidence of a departure from the null hypothesis in the direction of the log normal alternative.

To compare the two forms of the test statistic it is convenient to rewrite them at greater length as

$$\frac{1}{n} T'(\beta_{\hat{\alpha}}) = \hat{\beta}_1 - \beta_{1,\hat{\alpha}} + \frac{1}{2} \left\{ \frac{\hat{\beta}_2 - \psi'(1)}{\psi'(1)} + \frac{(\hat{\beta}_1 - \beta_{1,\hat{\alpha}})^2}{\psi'(1)} \right\} \qquad (7.1.10)$$

and

$$\frac{1}{n} T'(\hat{\beta}) = \hat{\beta}_1 - \beta_{1,\hat{\alpha}} + \frac{1}{2} \log \left\{ 1 + \frac{\hat{\beta}_2 - \psi'(1)}{\psi'(1)} \right\} . \qquad (7.1.11)$$

Since $x \geq \log (1+x)$, (7.1.10) is never less than (7.1.11), as would be expected from the general comparison of the two alternatives given in (6.13). We have already shown that, in general, the two are asymptotically equivalent. In this example we have that both expressions are of order $(1/\sqrt{n})$ in probability, except for the last term of (7.1.10) which is of order $(1/n)$. Under the null hypothesis we also have that $\hat{\beta}_2$ tends to $\psi'(1)$. Then, for large n, we have approximately that

$$\log\left\{1 + \frac{\hat{\beta}_2 - \psi'(1)}{\psi'(1)}\right\} = \frac{\hat{\beta}_2 - \psi'(1)}{\psi'(1)} \qquad (7.1.12)$$

so that the two forms of the statistic are asymptotically equivalent.

7.2. Empirical Results

The small sample distribution of Cox's statistic $T(\hat{\beta})$ was studied by Jackson (1968) empirically, by simulation on a computer, and analytically by study of higher terms in the Taylor's series expansion of the statistic about $(\alpha, \underset{\alpha}{\beta})$. To compare the two forms of the statistic we use only the empirical method of simulation.

Let u be a uniformly distributed random variable between 0 and 1. Then y = -log u will have the exponential distribution (7.1.1) with $\alpha = 1$. To calculate the test statistics we calculate the estimates of the parameters from the relationships (7.1.3). The results of simulations for five different sample sizes are summarized as the first four moments of the distributions of the two statistics in Table 4.

Table 4

The most striking feature of these results is that the approach to asymptotic normality is disappointingly slow. The sampling moments of $T(\hat{\beta})$ are in good agreement with those calculated by Jackson, who also studied the test of $\lambda = 0$, the log normal null hypothesis. The test statistic for that case also approached normality gradually.

The results of Table 4 do not help in choosing between the two statistics. $T(\beta_{\hat{\alpha}})$ is unbiased, as is expected from § 6, and is also preferable on the basis of variance. But the large values of the higher moments offset this preference. The histograms of the results of 1000 simulations for n = 20 showed that the distribution of $T(\hat{\beta})$ was slightly skewed, whereas that of $T(\beta_{\hat{\alpha}})$ was virtually symmetrical and centred about the origin, with a few outlying high values. $T(\beta_{\hat{\alpha}})$ is thus preferable apart from the outliers which are caused by values of y near zero giving large values of $\hat{\beta}_2$. These enter $T(\beta_{\hat{\alpha}})$ directly whereas in (7.1.9) for $T(\hat{\beta})$ they occur as $\log \hat{\beta}_2$, and so will have less effect on the value of the statistic.

The sensitivity of the test statistics to small values of y follows from the form of the two distributions, the exponential being a maximum at the origin whereas the log normal goes to zero. It is however often an undesirable feature, as these very small values are liable to relatively large recording errors. This sensitivity also places very stringent requirements on the random number generator used in these simulations, for values of u very close to one will have a disproportionate effect on the values of the statistics.

Because of the lack of bias and the shape of the distribution, $T(\beta_{\hat{\alpha}})$ is to be preferred. In order to justify this preference it is necessary that the small sample properties of the two statistics be

studied for examples less critically dependent on a small subset of the observations.

8. A TEST OF EQUIDISTANCE

The tests of the previous section were designed to detect departures from one distribution in the direction of an alternative. We now consider the choice between two distributions which enter the test symmetrically. That is, we test the hypothesis $\lambda = {}^1/2$. As an example we take as the constituent distributions the Poisson and the geometrical. The combined distribution $f_\lambda(y, \theta)$ is given in (2.6).

We define

$$E_\lambda \left\{ H(y) \right\} = \sum_{y=o}^{\infty} H(y)\, f_\lambda(y, \theta) \qquad (8.1)$$

with similar definitions for variances and covariances. Proceeding as before we obtain the test statistic

$$T = \frac{- \sum \log y! + n\, E_\lambda(\log y!) + \dfrac{C_\lambda(y, \log y!)}{V_\lambda(y)} \left\{ \sum y - n\, E_\lambda(y) \right\}}{\sqrt{n} \left[V_\lambda(\log y!) - \dfrac{\left\{ C_\lambda(y, \log y!) \right\}^2}{V_\lambda(y)} \right]^{1/2}} \qquad (8.2)$$

For testing the hypotheses $\lambda = 1$ and $\lambda = 0$, (8.2) reduces to the forms given by Cox (1962).

To test the hypothesis $\lambda = {}^1/2$ we calculate expectations etc. from (8.1) with $\lambda = {}^1/2$, employing some suitable estimate of the nuisance parameter γ. We could find this by maximizing the likelihood of the observations under the null hypothesis. In general, although

not in this case, the computation involved would be appreciable. An alternative is to use some locally consistent estimate which is easier to calculate.

The simplest of such estimates is to assume that the value of the ~~parameter~~ estimate varies linearly with λ, to estimate the parameter when λ is zero and one, and to interpolate to give an estimate for other values of λ. For the general combined p. d. f. (6. 1) this procedure is not possible, for the value of the likelihood does not depend on $\underline{\beta}$ when $\lambda = 1$. But in the combined distribution of (2. 6), γ is clearly defined for both $\lambda = 0$ and $\lambda = 1$. In fact we have the estimates

$$\lambda = 1 \ : \ \text{Poisson} \ : \ \hat{\gamma}_1 = \bar{y}$$

and

$$\lambda = 0 \ : \ \text{Geometric} : \hat{\gamma}_0 = \frac{\bar{y}}{1+\bar{y}} \tag{8.3}$$

We therefore take as our estimate of γ for some specified value of λ

$$\hat{\gamma}_\lambda = \lambda \hat{\gamma}_1 + (1-\lambda) \hat{\gamma}_o \, . \tag{8.4}$$

As a numerical example we consider the simulated data given by Cox in his Table 2. There are 30 observations with $\sum y_i = 26$ and $\sum \log y_i! = 5.950$. The estimates of the parameters are $\hat{\gamma}_1 = 0.8667$ and $\hat{\gamma}_o = 0.4643$. For testing the hypothesis $\lambda = {}^1/_2$ we take the

average of these values, 0.6655, as the estimate of γ. For comparison the maximum likelihood estimate has the value 0.6568 indicating that, for this example, the value of the parameter varies almost linearly with λ, a result which was confirmed by estimation for a series of values of λ.

Given these values of λ and γ the quantities required for evaluation of (8.2) can be calculated without difficulty as the series converge quite rapidly. The results of these calculations for $\lambda = 0$ and 1 as well as $^1/2$ are given in Table 5. There are some differences between these values and those given by Cox, although the general inferences remain unchanged.

Table 5

These results indicate that there is some evidence of departure from the hypothesis $\lambda = {}^1/2$ in the direction of the Poisson distribution, which agrees with the conclusions reached by Cox. In fact, the maximum of the likelihood as a function of γ and λ lies on the far side of the Poisson distribution, away from the geometric. This is because, for these data, the expected numbers of zeroes and observations greater than two are both less for the fitted Poisson distribution than for the fitted geometric. The observed frequencies deviate from the Poisson estimates in the same way.

There are two features of this application which are of importance to the theory of these tests. One is the use of the estimate of γ defined by (8.4). If a combined model is such that all the nuisance parameters are estimable for λ equals zero or one then linear interpolation may provide satisfactory estimates for testing $\lambda = {}^{1}/_{2}$. The other feature is that the value of the test statistic appears to vary linearly with λ. We consider this point further in § 9.2.

9. QUANTAL RESPONSES

9.1. Theory

In this section we consider the simplifications which occur when the observations take only the values 0 or 1 (Cox, 1962).

At k levels of a variable x_i, often called the dose level, n_i experiments are performed. Of these U_i are successful and therefore $n_i - U_i$ fail. The observed number of success is distributed binomially with index n_i. The purpose of the experiment is to determine the relationship between the dose level and the parameter of the binomial distribution.

We assume that there are two component models assigning probabilities θ_i and ϕ_i to success at dose level x_i. In general these probabilities will also depend on vectors of nuisance parameters $\underline{\alpha}$ and $\underline{\beta}$. We take as the combined probability

$$\theta_{\lambda i} = \frac{\theta_i^{\lambda}\phi_i^{1-\lambda}}{\theta_i^{\lambda}\phi_i^{1-\lambda} + (1-\theta_i)^{\lambda}(1-\phi_i)^{1-\lambda}} . \qquad (9.1.1)$$

Because the observations have only two possible values the integral in the denominator of the general combined p.d.f. (2.2) reduces to the summation of (9.1.1). Calculation of maximum likelihood estimates of λ, $\underline{\alpha}$ and $\underline{\beta}$ is thus greatly simplified and an example is given in § 9.2. But first we develop the appropriate asymptotically normal test statistic.

It is convenient to consider the proportion of successes rather than the absolute number. Let

$$P_i = U_i/n_i \,. \tag{9.1.2}$$

The log likelihood of the observations is

$$L = \sum n_i \left\{ P_i \log \theta_{\lambda i} + (1-P_i)\log(1-\theta_{\lambda i}) \right\}. \tag{9.1.3}$$

Differentiating with respect to λ we obtain the quantity

$$T' = \frac{\partial L}{\partial \lambda} = \sum n_i (P_i - \theta_{\lambda i}) \log \left\{ \frac{\theta_i/(1-\theta_i)}{\phi_i/(1-\phi_i)} \right\}. \tag{9.1.4}$$

This is the sum of the differences between the observed and predicted proportions of successes weighted by the number of observations and the log odds ratio from the component probabilities. It forms the basis for an intuitively appealing test statistic.

We require to adjust (9.1.4) for regression on the partial derivatives of the log likelihood with respect to the parameters of θ_i and ϕ_i. Let these be the scalar quantities α and β. Then

$$\frac{\partial L}{\partial \alpha} = \lambda \sum \frac{n_i (P_i - \theta_{\lambda i}) \theta_i'}{\theta_i (1-\theta_i)} \tag{9.1.5}$$

and

$$\frac{\partial L}{\partial \beta} = (1-\lambda) \sum \frac{n_i (P_i - \theta_{\lambda i}) \phi_i'}{\phi_i (1-\phi_i)} \tag{9.1.6}$$

where $\theta_i' = \dfrac{\partial \theta_i}{\partial \alpha}$ and similarly for ϕ_i'.

For testing the hypotheses $\lambda = 0$ or 1, one or the other of these quantities will equal zero. Otherwise, in calculating the value

of the statistic, the constants λ and $1-\lambda$ cancel and can be ignored.

Because of the binary nature of the response, the calculation of variances and covariances is greatly simplified. To calculate the covariance of two quantities Q and R which have the values Q_o and R_o when y = 0 and Q_1 and R_1 when y = 1 we have

$$C_\lambda(Q,R) = \theta_{\lambda i}(1-\theta_{\lambda i})\,(Q_1-Q_o)\,(R_1-R_o)\,. \qquad (9.1.7)$$

These results are sufficient for the construction of the statistic for testing the hypothesis that λ has some specified value. In the special case when $\lambda = 1$ we have that $\theta_{\lambda i} = \theta_i$ and the general results given here reduce to those of Cox (1962).

If $\lambda = 1$, the maximum likelihood equation for estimating α (9.1.5) becomes

$$\sum \frac{n_i(P_i-\hat{\theta}_i)\,\hat{\theta}_i'}{\hat{\theta}_i\,(1-\hat{\theta}_i)} = 0\,. \qquad (9.1.8)$$

In § 7 we estimated β_α by finding the expected value of $\hat{\beta}$ when the true distribution was $f(y,\alpha)$. The analogous estimate in this example is found by replacing the observations in the maximum likelihood equation for $\hat{\beta}$ by the estimated probabilities $\hat{\theta}_i$. Then if $\phi_{\hat{\alpha}i}$ is the probability under the alternative model when $\beta = \beta_{\hat{\alpha}}$ we have

$$\frac{n_i(\hat{\theta}_i - \phi_{\hat{\alpha}i})\,\phi_{\hat{\alpha}i}'}{\phi_{\hat{\alpha}i}\,(1-\phi_{\hat{\alpha}i})} = 0 \qquad (9.1.9)$$

as the relationship defining $\beta_{\hat{\alpha}}$.

9.2. One and Two Hit Models

We compare the test statistic derived from (9.1.4) with the asymptotic distribution of the likelihood ratio for testing hypotheses about λ. We also study the dependence of the value of the statistic and of the estimates of the nuisance parameters on λ.

Suppose we have the alternative models

$$\theta_i = 1 - e^{-\alpha x_i} \tag{9.2.1}$$

and

$$\phi_i = 1 - e^{-\beta x_i} - \beta x_i e^{-\beta x_i} . \tag{9.2.2}$$

The one hit model (9.2.1) represents the hypothesis that the probability of success depends upon the number of units receiving one or more impulses, whereas in (9.2.2) two or more impulses are required.

Maximum likelihood estimates of α and β at various values of λ for some data (Pereira and Kelly, 1957) are given in Table 6.

Table 6

An obvious feature of these results is that $\hat{\alpha}$ and $\hat{\beta}$ do not vary linearly with λ, so that linear interpolation in parameter estimates, as in § 8, is not possible. As $\lambda \to 1$ we have, from (9.1.6) that $\hat{\beta}$ must satisfy

$$\sum \frac{n_i (P_i - \hat{\theta}_i)}{\hat{\phi}_i (1 - \hat{\rho}_i)} \hat{\phi}_i' = 0 . \tag{9.2.3}$$

This is satisfied by $\hat{\beta} \rightarrow 0$ as $\lambda \rightarrow 1$ in such a way that the combined probability (9.1.1) remains defined. The solution of (9.1.9) for $\beta_{\hat{\alpha}}$ gives a value of 0.915. For λ near 1 the probabilities ϕ_i are small adjustments to the fitted θ_i which reduce the lack of fit. The value of $\hat{\beta}$ as $\lambda \rightarrow 0$ reflects this, whereas $\beta_{\hat{\alpha}}$ provides the best fit of the second model to the estimates from the first.

The overall maximum of the likelihood is when $\lambda = 0.755$. The differences in log likelihood and the squared values of the test statistic are given in Table 6. Although these values do not agree closely, the conclusions to be drawn from the two tests are similar, namely that there is evidence of departure from the two hit model in the direction of the one hit model and no evidence of departure in the reverse direction. Linear interpolation in T for $\lambda = {}^1/_2$ gives a value of 0.833 which agrees closely with the value of χ^2 found by maximizing the likelihood.

Further evidence of the linear dependence of T on λ comes from the value of λ for which T is zero. This is 0.756, nearly identical with the value of 0.755 found by maximizing the likelihood. The implication is that, in the neighbourhood of the maximum, the likelihood is sufficiently regular to be represented by a quadratic in λ. Under this condition the values of the statistic at $\lambda = 0$ and 1 may be expected to be sufficient for λ.

9.3. The Logistic Null Hypothesis

One of the models more frequently used in the analysis of binary data is the logistic response curve in which

$$\theta_i = \frac{\exp\{(x_i-\mu)/\tau\}}{1+\exp\{(x_i-\mu)/\tau\}} \quad . \tag{9.3.1}$$

We consider the test of the hypothesis $\lambda = 1$ when the alternative specifies that the probability is ϕ_i.

We obtain maximum likelihood estimates of the nuisance parameters μ and τ by substitution in (9.1.5) which yields the relationships

$$\sum n_i (P_i - \hat{\theta}_i) = 0$$

and

$$\sum n_i x_i (P_i - \hat{\theta}_i) = 0 . \tag{9.3.2}$$

Using the estimates $\hat{\mu}$ and $\hat{\tau}$ the statistic (9.1.4) reduces to

$$T' = - \sum n_i (P - \hat{\theta}_i) \log \left\{ \phi_i/(1 - \phi_i) \right\} . \tag{9.3.3}$$

The probabilities θ_i do not enter this expression in the log odds ratio because the logistic model is reproductive with respect to exponentiation.

Now suppose that there are only 3 dose levels x_1, x_2 and x_3. Whatever the alternative, the combined probability (9.1.1) contains at least three adjustable parameters λ, μ and τ. It is thus possible to describe the data exactly and to obtain estimates of the combined

probabilities identical with the observed values of the P_i. These estimates maximize the value of the likelihood. Using the log likelihood ratio test of § 4 we would compare this maximum value with the maximized value of the likelihood under the logistic null hypothesis. The asymptotically normal test of § 5 is an approximation to this test, so that, from (9.3.3), the most powerful test of the logistic model against any alternative is

$$T' = - \sum n_i (P_i - \hat{\theta}_i) \log \left\{ P_i/(1-P_i) \right\}, \qquad (9.3.4)$$

provided that there are only 3 dose levels. Now let

$$Z_i = \log \left(\frac{P_i}{1-P_i} \right) \qquad (9.3.5)$$

when the test statistic may be written

$$T' = - \sum n_i (P_i - \hat{\theta}_i) Z_i . \qquad (9.3.6)$$

The maximum likelihood equations (9.3.2) yield, upon rearrangement,

$$\frac{n_1(P_1-\hat{\theta}_1)}{x_2 - x_3} = \frac{n_2(P_2-\hat{\theta}_2)}{x_3 - x_1} = \frac{n_3(P_3-\hat{\theta}_3)}{x_1 - x_2} . \qquad (9.3.7)$$

An alternative form of the test statistic is therefore

$$T' \propto (x_2-x_3) Z_1 + (x_3-x_1) Z_2 + (x_1-x_2) Z_3 . \qquad (9.3.8)$$

This is identical with a statistic given by Chambers and Cox (1967) for testing the goodness of fit of the logistic model. If there are only

three dose levels this statistic is essentially unique. For a greater number of dose levels, the form of the statistic must depend upon the particular alternative against which the logistic model is tested.

10. DISCUSSION

The results obtained above demonstrate the usefulness of considering the properties of exponential combinations of p. d. f.'s. What are the properties of other methods of combination?

Two general points in the theory of tests for separate families require further investigation.

1. Is the conclusion of § 7 that $T(\beta_{\hat{\alpha}})$ is preferable to $T(\hat{\beta})$ justified?

2. Establishment of conditions under which the value of the test statistic varies linearly with λ.

We have said nothing about the design of experiments to discriminate between models. This raises the problem of design in the presence of nuisance parameters. Other extensions would be to the study of time series and to multivariate problems. Finally, use of the combined p. d. f. may overcome the difficulty in the Bayesian analysis of problems of this kind in which prior probabilities have to be assigned to the models and prior distributions to the parameters.

REFERENCES

Abramowitz, M. and Stegun, I. A. (eds.) (1965), Handbook of Mathematical Functions. New York: Dover.

Atkinson, A. C. (1969), A test for discriminating between models. Biometrika 56, 337-47.

Bartlett, M. S. (1953), Approximate confidence intervals II. More than one unknown parameter. Biometrika 40, 306-17.

Chambers, E. A. and Cox, D. R. (1967), Discrimination between alternative binary response models. Biometrika 54, 573-8.

Cox, D. R. (1961), Tests of separate families of hypotheses. Proc. 4th Berkeley Symp. 1, 105-23.

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

Cox, D. R. and Brandwood, L. (1959), On a discriminatory problem connected with the works of Plato. J. R. Statist. Soc. B 21, 195-200.

Hoel, P. G. (1947), On the choice of forecasting formulas. J. Am. Statist. Ass. 42, 605-11.

Jackson, O. A. Y. (1968), Some results on tests of separate families of hypotheses. Biometrika 55, 355-63.

Kendall, M. G. and Stuart, A. (1961), The Advanced Theory of Statistics, vol. 1, 2nd ed. London: Griffin.

Neyman, J. S. (1959), Optimal asymptotic tests of composite statistical hypotheses. In Probability and Statistics, The Harald Cramer Volume (ed. U. Granander); Uppsala, Sweden: Almquist and Wiksells (New York : Wiley).

Pereira, H. G. and Kelly, B. (1957), Toxic and infective action of adenovirus. J. Gen. Microbiol. 17, 517-24.

Williams, E. J. (1959), Regression Analysis, New York : Wiley.

TABLE 1

Comparison of two parameter model and reduced model for dating 5 works of Plato.

Work	Full Model		Reduced Model	-2 log l
	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}$	
T	-0.046	0.057	0.250	93.48xxx
S	-0.032	0.244	0.392	74.92xxx
C	+0.089	0.291	0.404	7.72xx
P	-0.160	0.680	0.764	25.48xxx
F	-0.095	0.921	0.949	4.29x

Percentage points of χ^2_1

5%	3.84 x
1%	6.63 xx
0.1%	10.83 xxx

TABLE 2

Test of the irrelevance of R in describing the 5 works.

Work	$\hat{\lambda}_2$	-2 log l
T	0.055	0.241
S	0.241	0.135
C	0.301	0.174
P	0.649	2.804
F	0.897	1.151

TABLE 3

Percentage Distributions of Selected Sentence Endings

Type of Ending	R	T	S	C	P	F	L
-UUU-	4.6	3.3	2.3	6.0	4.0	6.5	8.8
- U - U -	6.4	3.0	2.1	1.3	1.8	2.8	2.4
U -- U -	4.8	3.0	4.6	5.3	4.5	5.3	8.2
- - - U -	4.1	3.8	4.7	2.0	6.8	9.0	8.8

TABLE 4

Moments of the empirical distributions of the two statistics for testing the exponential distribution against the log normal

n		Mean	Variance	γ_1	β_2
20	$T(\hat{\beta})$	-0.411	0.706	0.105	3.758
	$T(\beta_{\hat{\alpha}})$	0.063	0.816	2.200	16.304
50	$T(\hat{\beta})$	-0.269	0.863	0.469	3.261
	$T(\beta_{\hat{\alpha}})$	0.058	0.942	0.991	4.535
100	$T(\hat{\beta})$	-0.256	0.900	0.542	3.968
	$T(\beta_{\hat{\alpha}})$	-0.032	0.934	1.046	5.465
150	$T(\hat{\beta})$	-0.169	0.861	0.461	3.734
	$T(\beta_{\hat{\alpha}})$	0.020	0.989	1.212	7.345
250	$T(\hat{\beta})$	-0.104	0.925	0.369	3.228
	$T(\beta_{\hat{\alpha}})$	0.048	0.955	0.797	4.159

The results for n = 20 are based on 1000 trials, the rest from 500 trials.

TABLE 5

Tests of the Poisson and Geometric Distributions

λ	$E_\lambda(\ln y!)$	$C_\lambda(y, \ln y!)$	$V_\lambda(y)$	$V_\lambda(\ln y!)$	T
0	0.394	1.309	1.618	1.363	1.948
$^1/_2$	0.301	0.697	1.131	0.580	1.318
1	0.233	0.441	0.867	0.321	0.606

$\lambda = 0$ is the geometric distribution and $\lambda = 1$ the Poisson

TABLE 6

One and Two Hit Binary Responses

λ	$\hat{\alpha}$	$\hat{\beta}$	T	T^2	-2 log l
1	0.413	(0.915)	-0.795	0.633	0.868
0.755	0.508	0.556	-	-	-
0.5	0.531	0.752	0.833^{*}	0.694^{*}	0.632
0	(0.403)	0.905	2.461	6.057	4.930

The parameter estimates in parentheses are $\alpha_{\hat{\beta}}$ and $\beta_{\hat{\alpha}}$. The asterisked value of T for $\lambda = {}^{1}/2$ is calculated by interpolation. The overall maximum of the likelihood is given by $\lambda = 0.755$.

PART 3

OTHER PUBLICATIONS

These notes give the background to the writing of the four papers which follow.

1. 'The Design of Experiments for Parameter Estimation' is a generalization of a dissertation submitted for the Diploma of Imperial College in statistics. The work was supervised by Dr. W. G. Hunter.

2. 'A Mathematical Basis for the Selection of Research Project' resulted from two years' work on operations research in the American chemical industry. Dr. Bobis, my group leader, was responsible for initiating the work and obtaining the co-operation and interest of the users of the method. The development of the model was a joint undertaking, in which I provided the mathematical formulation and the programming. Owing to problems with company clearance the published paper does not contain any numerical examples.

3. 'The Use of Residuals as a Concomitant Variable' was intended to form part of this thesis, but was abandoned as the problem was not suitable for longer treatment.

4. The purpose of 'Constrained Maximisation and the Design of Experiments' is purely didactic. I hope it may help to reduce the amount of time spent searching over grids to find the maximum of functions.

TECHNOMETRICS VOL. 10, NO. 2 MAY 1968

The Design of Experiments for Parameter Estimation[1]

ANTHONY C. ATKINSON[2]
American Cyanamid Company

WILLIAM G. HUNTER
The University of Wisconsin

This paper is concerned with the design of experiments to estimate the parameters in a model of known form, which may be nonlinear in the parameters. This problem was discussed in detail by Box and Lucas for the case where N, the number of experiments, is equal to p, the number of parameters. The present work is an extension to cases where N is greater than p. Conditions are established under which, when the number of experiments is a multiple of the number of parameters, replication of the best design for p experiments is an optimal design for N experiments. Several chemical examples are discussed; in each instance, the best design consists of simply repeating points of the original design for p experiments. An example is also mentioned where the best design does *not* consist of such replication.

1. OUTLINE OF THE PROBLEM

We shall be concerned with the design of experiments to estimate the parameters in a model of known form when there is a single measurable response for each set of experimental conditions. As an example, suppose an experimenter is studying a system in which a raw material A reacts to form a product B which, in turn, decomposes to give a substance C. An experiment consists of measuring η, the amount of B present, after the reaction has proceeded for a time ξ_1 . For each experiment this is the only response which is observed. If further it is known that both reactions are irreversible and first-order and that initially at time $\xi_1 = 0$ there is one unit of A present, none of B, and none of C, then the relationship between the response η and the time ξ_1 will be of the form

$$\eta = \frac{\theta_1}{\theta_1 - \theta_2} \{\exp(-\theta_2\xi_1) - \exp(-\theta_1\xi_1)\} \tag{1.1}$$

where θ_1 and θ_2 are the rate constants of the two steps of the reaction which could be represented schematically as

$$A \xrightarrow{\theta_1} B \xrightarrow{\theta_2} C.$$

Received December 1966; revised May 1967.
[1] This research was supported in part by the National Science Foundation under Contract NSF-GP 2755.
[2] Presently at Imperial College, London, S. W. 7.

The problem we are considering is: at what times should measurements of η be taken in order to obtain estimates of the parameters θ_1 and θ_2 which are as precise as possible?

In general the model, which need not be linear in either the parameters or the variables, may be written as

$$\eta = f(\xi, \theta) \tag{1.2}$$

where

η = the true value of the measured response y,
ξ = the vector of k controllable variables $(\xi_1, \xi_2, \cdots, \xi_k)$,
θ = the vector of p parameters $(\theta_1, \theta_2, \cdots, \theta_p)$.

The program of N experiments may be represented by the $N \times k$ design matrix $\mathbf{D} = \{\xi_{iu}\}$. The ith member of the uth row of this matrix gives the level of the ith variable for the uth experiment. Frequently the values which the variables can take will be restricted by physical constraints on the system. Obviously in the example above, for instance, negative values of ξ_1 have no physical meaning. The subspace of the k-dimensional Ξ-space defined by the constraints will be known as the region of operability $R(\xi)$. Outside this region experiments cannot be performed.

2. The Design Criterion

The approach we shall use in designing experiments in nonlinear situations is that of Box and Lucas (5). In their paper, they examine applications of the method to several models (including the preceding example) when the number of experiments equals the number of parameters; that is, when $N = p$. In Section 3 the method is applied to two examples when N is greater than p. In the present section, the logic of the design criterion is presented, followed by a simple example of its application.

The particular nonlinear functions which we shall consider arise from developing models based on chemical kinetics. In such situations, it is usual for the experimenter to have some idea of the values of the parameters before the experiment commences. This is also, of course, generally true in fields other than chemical kinetics. We assume that the response function is approximately linear in θ near these preliminary estimates θ^0, that is, that it can be expanded to a Taylor's series in $(\theta - \theta^0)$ to give

$$f(\xi, \theta) \doteqdot f(\xi, \theta^0) + \sum_{i=1}^{p} \frac{\partial f(\xi, \theta)}{\partial \theta_i}\bigg|_{\theta=\theta^0} (\theta_i - \theta_i^0). \tag{2.1}$$

Now let

$$\mathbf{Z} = \{z_u\} = \{f(\xi_u, \theta) - f(\xi_u, \theta^0)\}, \tag{2.2}$$

$$\mathbf{X} = \{x_{iu}\} = \left\{\frac{\partial f(\xi_u, \theta)}{\partial \theta_i}\bigg|_{\theta=\theta^0}\right\}, \tag{2.3}$$

$$\mathbf{B} = \{b_i\} = \{\theta_i - \theta_i^0\}. \tag{2.4}$$

The subscript u designates the run number, $u = 1, 2, \cdots, N$. Under these conditions the experimental results can be reduced to the matrix form $\mathbf{Z} \doteqdot \mathbf{XB}$.

The problem is therefore transformed to that of finding the most efficient design in the X-space to estimate the vector of coefficients $\mathbf{B}$ in a linear model, where $\mathbf{X}$ is an $N \times p$ matrix of partial derivatives of the response. Each x_{iu} will, in general, be a function of the vector of the preliminary estimates of the parameters $\boldsymbol{\theta}^0$ and also of the vector of process variables ξ_u. The vector $\boldsymbol{\theta}^0$ may represent initial estimates of the parameters available before any data are obtained from the current investigation, or it could be composed of the current best estimates of the parameters (say, least squares estimates) based on some data that have already been collected.

Due to experimental error, the measured response of the uth experiment will not be η_u the true response but rather y_u where

$$y_u = \eta_u + e_u . \tag{2.5}$$

Assuming that the errors have zero mean, are independent and have constant variance,

$$E(y_u) = \eta_u = f(\xi_u , \boldsymbol{\theta})$$

and

$$E(y_u - \eta_u)(y_v - \eta_v) = \begin{cases} \sigma^2 & u = v \\ 0 & u \neq v \end{cases}.$$

Under these conditions if the model is in fact linear in the parameters it is known that the least squares estimates $\hat{\boldsymbol{\theta}}$ resulting from minimizing the sum of squares

$$\sum_{u=1}^{N} \{y_u - f(\xi_u , \boldsymbol{\theta})\}^2 \tag{2.6}$$

have the variance-covariance matrix $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. If, further, the errors are normally distributed, the boundary of a region with confidence coefficient $1 - \alpha$ in the space of parameters is formed by the values of $\boldsymbol{\theta}$ which satisfy the relationship

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) = s^2 p F_\alpha(p, \nu) \tag{2.7}$$

where $F_\alpha(p, \nu)$ is the α percent point of the F-distribution with p and ν degrees of freedom and s^2 is an independent estimate of the error variance σ^2 based on ν degrees of freedom. The boundary of such a region is hyper-ellipsoidal, the volume of the hyper-ellipsoid depending upon the value of the determinant $|\mathbf{X}'\mathbf{X}|$. For given values of $F_\alpha(p, \nu)$ and s, the volume will decrease as the value of the determinant increases. In the nonlinear case these results are approximately true, the accuracy depending upon the degree of nonlinearity of the function.

If it be assumed that the values of all the parameters are of equal interest to the experimenter, a reasonable design criterion is that it should minimize the volume of the confidence region for the estimates of the parameters. This

is achieved exactly if the model is linear and approximately if the model is nonlinear by choosing the experiments to maximize $|\mathbf{X}'\mathbf{X}|$. This criterion is directly related to Wilks' generalized variance.

The design criterion is, therefore, to find that set of N experimental conditions which maximizes Δ_N where

$$\Delta_N = |\mathbf{X}'\mathbf{X}| = \begin{vmatrix} \sum_{u=1}^{N} x_{1u}^2 & \sum_{u=1}^{N} x_{1u}x_{2u} & \cdots & \sum_{u=1}^{N} x_{1u}x_{pu} \\ \vdots & & & \\ \sum_{u=1}^{N} x_{iu}x_{1u} & \sum_{u=1}^{N} x_{iu}x_{2u} & \cdots & \sum_{u=1}^{N} x_{iu}x_{pu} \\ \vdots & & & \\ \sum_{u=1}^{N} x_{pu}x_{1u} & \sum_{u=1}^{N} x_{pu}x_{2u} & \cdots & \sum_{u=1}^{N} x_{pu}^2 \end{vmatrix}. \tag{2.8}$$

Because of the nature of the partial derivatives and the constraints on the variables which define the region of operability $R(\xi)$, not all values of x_{iu} will be attainable. Those values which are available for experimentation will be said to define the attainable region $R(\mathbf{x})$, a subspace of the p-dimensional X-space. The design problem then becomes that of selecting N points in $R(\mathbf{x})$ which maximize Δ_N.

For the special case where the number of experiments is equal to the number of parameters, $\mathbf{X}$ is a square $p \times p$ matrix and $|\mathbf{X}'\mathbf{X}| = |\mathbf{X}|^2$ where

$$|\mathbf{X}| = \begin{vmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ \vdots & & & \\ x_{1u} & x_{2u} & \cdots & x_{pu} \\ \vdots & & & \\ x_{1p} & x_{2p} & \cdots & x_{pp} \end{vmatrix}. \tag{2.9}$$

In the p-dimensional X-space, the value of this determinant is proportional to the volume of the simplex formed by the origin and the p experimental points. Thus, an optimal design will be one for which the volume of the simplex is maximized. It follows, then, that for an $N = p$ design to be optimal, the p experimental points must lie on the boundary of $R(\mathbf{x})$.

In some cases, more than one set of experimental conditions may give the same value for the determinant. If this happens, the sets of experimental designs will be considered equally good, regardless of the shapes of the approximate confidence regions to which they give rise.

2.1. *A Linear Example*

In order to illustrate the use of the design criterion and the geometrical interpretation of the determinant, a simple linear model will be considered.

Suppose that

$$\eta = \theta_1\xi_1 + \theta_2\xi_2 \tag{2.10}$$

with the restrictions that

$$0 \leq \xi_1 \leq 1 \quad \text{and} \quad 0 \leq \xi_2 \leq 1.$$

The region of operability $R(\xi)$ is therefore the unit square, sides parallel to the axes, with one corner at the origin.

Applying the design criterion, we obtain

$$x_1 = \frac{\partial \eta}{\partial \theta_1} = \xi_1 \tag{2.11}$$

$$x_2 = \frac{\partial \eta}{\partial \theta_2} = \xi_2 \tag{2.12}$$

and

$$\Delta_2 = \begin{vmatrix} \xi_{112} + \xi_{122} & \xi_{11}\xi_{21} + \xi_{12}\xi_{22} \\ \xi_{11}\xi_{21} + \xi_{12}\xi_{22} & \xi_{21}^2 + \xi_{22}^2 \end{vmatrix}. \tag{2.13}$$

In the linear case, preliminary estimates of the parameters are not needed because the x_i's and hence the determinant Δ are not functions of the θ's. In this case, the X-space and Ξ-space are identical. From the geometrical interpretation of the determinant when $N = p$, the optimal design for two experiments consists of two points in $R(\mathbf{x})$ which, together with the origin, form a triangle of greatest area. These points must be somewhere on the boundary of $R(\mathbf{x})$. For this example, there are an infinity of designs giving the same maximum value of the determinant, namely one experiment at $(x_1, x_2) = (0, 1)$ and the other anywhere between and including (1, 0) and (1, 1) or one at (1, 0) and the other anywhere between and including (0, 1) and (1, 1). All these designs give a value of Δ_2 equal to one.

2.2. *The General Design Procedure*

If, as was not the case for the preceding example, the model is nonlinear, it is necessary to have preliminary estimates of the parameters in order to apply the design criterion. If the estimates upon which the design is based are poor, the design may be inefficient, the robustness of the criterion to poor estimates depending upon the particular model being studied. Hence, in most cases in order to obtain the maximum information per experiment, p experiments could be performed, from which the parameters could be re-estimated. Thereafter, the experiments could be designed one at a time, using the current best estimates of the parameters, which could be recalculated after each experiment. This sequential procedure has been discussed elsewhere (4).

In what follows it will be assumed that N experiments, where N is no less than p and may be greater than p, are to be designed on the basis of the preliminary estimates θ^0. The results are not analyzed until all N runs have been performed. Although ordinarily one would like to proceed sequentially and analyze the results as they are obtained so that all available information would be used in designing the next experiment, there are situations in which this is not feasible. In other words, it may be inefficient to go to the computer after each experiment to re-estimate the parameter values and determine the best

settings for the experimental variables for the next experiment. In situations of this kind, it will be reasonable to design experiments more than one at a time at each stage. For example, many practical examples exist where N runs can conveniently be made in one day and the computer can then be used overnight to provide new estimates for the parameters. In these circumstances, the computer could also be used to find the optimal set of new design points for further runs, perhaps to be performed the following day. There are situations, then, in which it is the best policy to design experiments in groups of $N > p$, the case we discuss in this paper.

3. Application of the Design Criterion

3.1. *Two Consecutive First-Order Reactions—Model* 1

In order to illustrate the use of the design criterion in a nonlinear situation, we return to the first example we mentioned, one that was studied in considerable detail by Box and Lucas (5). The model describes the amount of B present after the reaction $A \rightarrow B \rightarrow C$ had been underway for some time ξ_1. Suppose that the preliminary estimates of the parameters are $\theta_1^0 = 0.7$ and $\theta_2^0 = 0.2$. Then

$$\eta = 1.4\{\exp(-0.2\xi_1) - \exp(-0.7\xi_1)\}, \tag{3.1}$$

$$x_1 = \frac{\partial \eta}{\partial \theta_1} = (0.8 + 1.4\xi_1)\exp(-0.7\xi_1) - 0.8\exp(-0.2\xi_1), \tag{3.2}$$

$$x_2 = \frac{\partial \eta}{\partial \theta_2} = (2.8 - 1.4\xi_1)\exp(-0.2\xi_1) - 2.8\exp(-0.7\xi_1). \tag{3.3}$$

These three functions of ξ_1 are plotted by Box and Lucas.

Using a computer function maximization routine, the optimal design for two experiments was found to consist of terminating the reaction at times of 1.23 and 6.86 units, giving a value of 0.6568 for the determinant Δ_2. In order to appreciate the design situation, it is profitable to consider the X-space as shown in Figure 1. For any value of ξ_1 the values of x_1 and x_2 are fixed, so that $R(\mathbf{x})$ is a curved line. The points of the optimal design, shown by heavy dots labelled 1 and 2, together with the origin, form the triangle of maximum area within $R(\mathbf{x})$.

Since there is only one controllable variable, the value of the determinant for N experiments is a function of only N variables, the times at which the runs are terminated. The optimal designs for various numbers of experiments from 3 to 20 were also found using a computer function maximization routine. The results are shown in Table 1. In each case, the optimal design was found to consist of experiments solely at those two times which form the optimal design for two experiments. When N was even, equal numbers of experiments at each point maximized the value of the determinant. For odd N an extra experiment at either of the two sets of conditions gave the same maximal value of the determinant. These results were checked by performing the maximization starting from a variety of randomly chosen points.

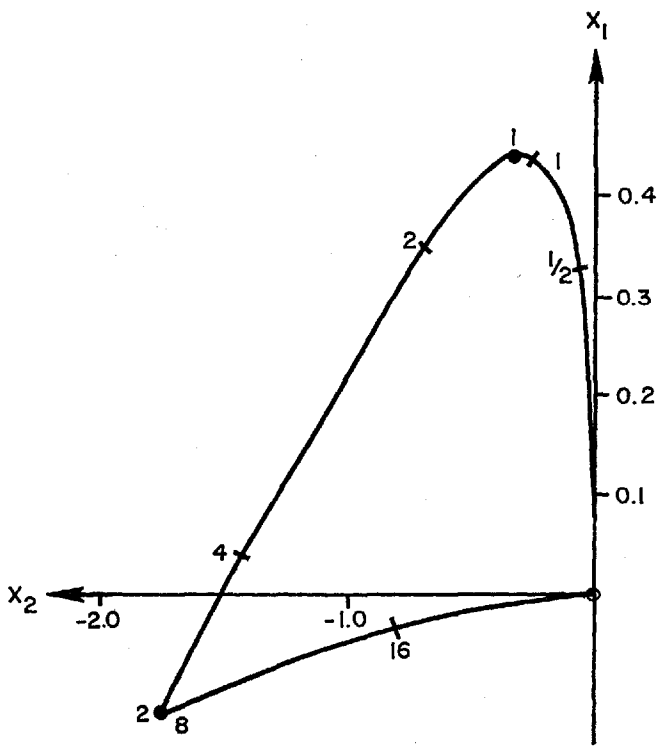

FIGURE 1
$R(\mathbf{x})$ for Model 1. Heavy dots indicate the optimal design, dashes indicate values of ξ_1.

3.2. *The Catalytic Dehydration of Hexyl Alcohol—Model 2*

The result that the optimal design for $N = np$ ($n = 2, 3, 4, \cdots$) experiments consists of n replications of the best design for p experiments was obtained with a second nonlinear model.

TABLE 1
Optimal design for up to 20 experiments for Model 1

	Number of experiments at		
N	$\xi_1 = 1.23$	$\xi_1 = 6.86$	Δ_N
2	1	1	0.6568
3	$\begin{bmatrix}1\\2\end{bmatrix}$	$\begin{bmatrix}2\\1\end{bmatrix}$	1.3135
4	2	2	2.6271
5	$\begin{bmatrix}2\\3\end{bmatrix}$	$\begin{bmatrix}3\\2\end{bmatrix}$	3.9406
6	3	3	5.9110
10	5	5	16.4193
20	10	10	65.6774

The rate η of catalytic dehydration of certain tertiary and long chain primary alcohols to an olefin and water can be written as

$$\eta = \frac{\theta_3\theta_1\xi_1}{1 + \theta_1\xi_1 + \theta_2\xi_2} \tag{3.4}$$

where

ξ_1 = the partial pressure of alcohol,
ξ_2 = the partial pressure of olefin,
θ_1 = the adsorption equilibrium constant for alcohol,
θ_2 = the adsorption equilibrium constant for olefin,
θ_3 = the effective reaction rate constant.

This model was used elsewhere (4) in an example illustrating a sequential design procedure.

Each experiment involves setting values of the partial pressures ξ_1 and ξ_2 and observing the rate η. Since there are three parameters in the model, $R(\mathbf{x})$ is three-dimensional and finding the optimal design for three experiments involves a search in nine dimensions. To facilitate visual presentation of the results and to better appreciate their significance, we decided to proceed as if

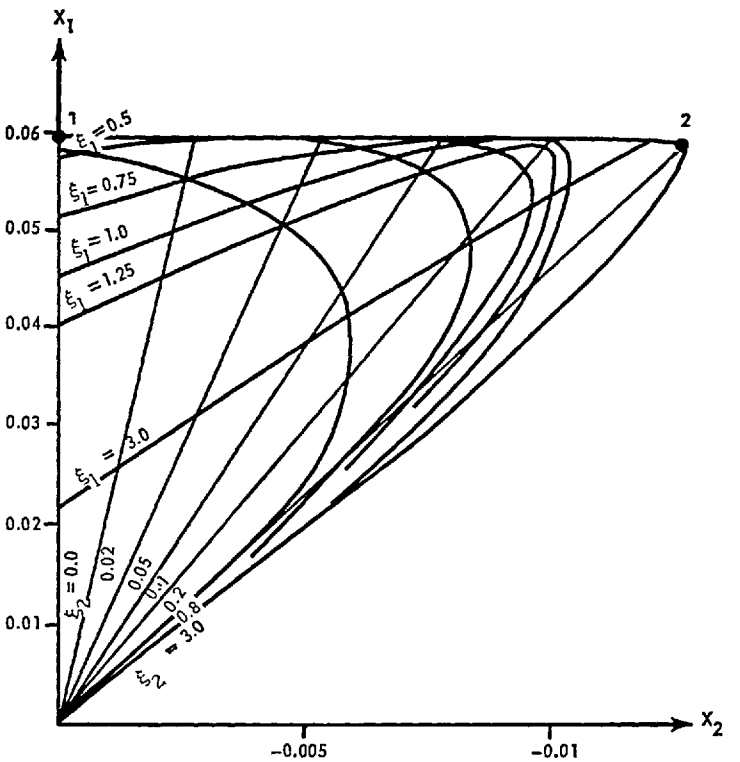

FIGURE 2
$R(\mathbf{x})$ for Model 2, the catalytic dehydration of hexyl alcohol. The conditions for the optimal design are shown by the heavy dots.

the value of θ_3 were known, and experiments were to be performed for the estimation of θ_1 and θ_2.

Following reference (4), we take as our initial values of the parameters $\theta_1 = 2.9$, $\theta_2 = 12.2$ and θ_3 (known) $= 0.69$ with the following constraints defining $R(\xi)$: $0 \leq \xi_1 \leq 3$ and $0 \leq \xi_2 \leq 3$. Then

$$\eta = \frac{2.001\xi_1}{1 + 2.9\xi_1 + 12.2\xi_2}, \tag{3.5}$$

$$x_1 = \frac{0.69\xi_1(1 + 12.2\xi_2)}{(1 + 2.9\xi_1 + 12.2\xi_2)^2}, \tag{3.6}$$

$$x_2 = \frac{-2.001\xi_1\xi_2}{(1 + 2.9\xi_1 + 12.2\xi_2)^2}. \tag{3.7}$$

For these parameter values, the resulting $R(\mathbf{x})$ is shown plotted in Figure 2. Loci of constant ξ_2 are straight lines passing through the origin, whereas loci of constant ξ_1 are curved. The optimal design for two experiments, shown by heavy dots in the figure, was found to correspond to the following conditions:

Experiment	ξ_1	ξ_2
(1)	0.3448	0
(2)	3.0	0.7951

These gave a value of 5.6903×10^{-7} for Δ_2. From Figure 2, it can be seen that these conditions define the triangle of maximum area within $R(\mathbf{x})$ when one corner of the triangle is the origin.

The conditions for optimal designs for up to 10 experiments are presented in Table 2. Here $N = np$ ($n = 2, 3, 4, 5$). In each case the optimal design again consists of repeating the two conditions of the optimal design for two experiments.

TABLE 2

Optimal design for up to 10 experiments for Model 2

	Number of experiments at conditions		
N	(1)	(2)	$\Delta_N \times 10^7$
2	1	1	5.69
3	1	2	11.4
4	2	2	22.8
5	2 or 3	3 or 2	34.1
10	5	5	142.3

where the settings of the process variables for the experimental conditions are

Condition	ξ_1	ξ_2
(1)	0.3448	0
(2)	3.0	0.7951

4. Conditions for Replication

From the preceding examples, it seems that the optimal design for $N = np$ experiments, $n = 2, 3, \cdots$, may always consist of replications of the optimal design for p experiments. This conclusion has also been suggested by Behnken (1) in connection with a model arising in the study of copolymerization.

In this section we shall show by a counterexample that replication of an optimal p-point does *not* always give the best $N > p$ design, and then proceed to establish conditions on $R(\mathbf{x})$ under which replicated designs are optimal.

4.1. *A Counterexample*

For the linear model

$$\eta = \theta_1\xi_1 + \theta_2\xi_2 \tag{4.1}$$

with the constraints that both controllable variables have values between zero and unity, the attainable region $R(\mathbf{x})$ is a unit square. One of the many optimal designs, all of which give a value of one for Δ_2, consists of experiments at points $(x_1, x_2) = (0, 1)$ and $(1, 0)$, that is, at unity on each of the coordinate axes. The corresponding X matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.2}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.3}$$

We now consider two alternative designs for six experiments. For the first design the preceding design is replicated three times; thus

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, \tag{4.4}$$

whence

$$\Delta_6 = |\mathbf{X}'\mathbf{X}| = 9.$$

For the second design, an experiment is also performed at the point $(x_1, x_2) = (1, 1)$. The X matrix for these three experiments is

$$\mathbf{X} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \tag{4.5}$$

and

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \tag{4.6}$$

Repeating this design once gives six experiments in all for which

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix} \tag{4.7}$$

and

$$\Delta_6 = 12.$$

In this case, the first design for six experiments is clearly poorer although it does consist of replications of an optimal design for two experiments. In the next section we establish conditions on $R(\mathbf{x})$ for replication to provide optimal designs.

4.2. *A Criterion for Replication*

In Section 2 we mentioned that when $N = p$, $\mathbf{X}$ is a square $p \times p$ matrix and $\Delta_p = |\mathbf{X}'\mathbf{X}| = |\mathbf{X}|^2$, where the uth row of $\mathbf{X}$ consists of the co-ordinates of the uth experiment in the p-dimensional $\mathbf{X}$-space. Now suppose that the conditions for the uth experiment were altered as follows. A new $\mathbf{X}$ matrix is formed by deleting the uth row and adding one other, the $(p + 1)$th. Denote the determinant of this matrix by $\delta_{p+1,-u}$. The determinant of the original matrix would, under this scheme, be denoted by $\delta_{p+1,-(p+1)}$. We shall use this nomenclature to state two theorems, the proofs of which are given in the appendix.

Theorem 1. A necessary condition for the optimality of a p-point design is that it shall consist of points on $R(\mathbf{x})$ such that $R(\mathbf{x})$ does not lie outside the p-dimensional parallelepiped defined by the p pairs of planes

$$\delta^2_{p+1,-u} = \Delta_p, \qquad u = 1, 2, \cdots, p. \tag{4.8}$$

The uth of these pairs of planes passes through the points u and u' (the reflection of u in the origin) and is parallel to the plane defined by the origin and the remaining $(p - 1)$ experimental points.

Theorem 2. A sufficient condition for the optimal design for $N = np$ experiments, $n = 2, 3, \cdots$, to consist solely to replications of p points which are optimal for $N = p$ is that $R(\mathbf{x})$ be contained by the p-ellipsoid E, the locus of points $(p + 1)$ satisfying the relationship

$$\sum_{u=1}^{p} \delta^2_{p+1,-u} = \Delta_p. \tag{4.9}$$

The p-ellipsoid E, a transformation of the unit p-sphere centered at the origin, passes through the p experimental points, at each of which the equation of the tangent plane to E is given by Equation 4.8.

When E is transformed so that it becomes the unit p-sphere E^T, the p points defining E^T are a rotation of the simplex consisting of points at unity on the p axes. In terms of the weighing designs discussed by Hotelling (6) and Mood (7), such a design is equivalent to weighing p objects one at a time on a spring balance. These designs are also related to the first-order rotatable designs described by Box (2).

Consider the first-order polynomial which may be written as

$$\eta = \theta_1 + \sum_{i=2}^{p} \theta_i \xi_i \,. \tag{4.10}$$

Then

$$x_1 = \frac{\partial \eta}{\partial \theta_1} = 1 \tag{4.11}$$

and

$$x_i = \frac{\partial \eta}{\partial \theta_i} = \xi_i , \qquad i = 2, \cdots , p. \tag{4.12}$$

The optimal design for p experiments consists of points at the vertices of an orthogonal simplex in the p-dimensional X-space. Alternatively, since $x_1 = 1$ for all design points, the design can also be considered as forming a simplex in $p - 1$ dimensional space. The optimal design will form a regular simplex in $p - 1$ dimensions.

4.3. *Linear Model Example*

Before applying the preceding results to some nonlinear models, we shall apply them to the linear model, Equation 4.1, mentioned earlier.

The equation of E was defined as the locus of points satisfying the relationsip

$$\sum_{u=1}^{p} \delta^2_{p+1,-u} = \Delta_p . \tag{4.13}$$

When $p = 2$ this may be written as

$$\begin{vmatrix} x_{11} & x_{21} \\ x_{13} & x_{23} \end{vmatrix}^2 + \begin{vmatrix} x_{12} & x_{22} \\ x_{13} & x_{23} \end{vmatrix}^2 = \begin{vmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{vmatrix}^2 \tag{4.14}$$

where the point (x_{13} , x_{23}) lies on E. Multiplying out, we obtain

$$(x^2_{21} + x^2_{22})x^2_{13} + (x^2_{11} + x^2_{12})x^2_{23} - 2(x_{11}x_{21} + x_{12}x_{22})x_{13}x_{23} = (x_{11}x_{22} - x_{12}x_{21})^2 \tag{4.15}$$

which is the equation of an ellipse centered at the origin.

One of the possible optimal designs for the linear model consists of experiments at the points $(x_1 , x_2) = (0, 1)$ and $(1, 0)$. Substituting these values in the equation for E gives

$$x_1^2 + x_2^2 = 1 \tag{4.16}$$

as the boundary within which $R(\mathbf{x})$ must lie for the conditions of Theorem 2 to be satisfied. As can be seen in Figure 3(a) the point $(x_1 , x_2) = (1, 1)$ lies *outside* this boundary, so the conditions are not satisfied. We have already shown by example that replications of this design do not yield an optimal design (Section 4.1).

If we suppose instead that the design consisted of experiments at the points $(1, 0)$ and $(1, 1)$, then the equation of E becomes

$$x_1^2 + 2x_2^2 - 2x_1x_2 = 1.$$

Again, part of $R(\mathbf{x})$ lies outside E as can be seen in Figure 3(b) and, although this design is optimal for two experiments, replications of it are not optimal for any even number of experiments.

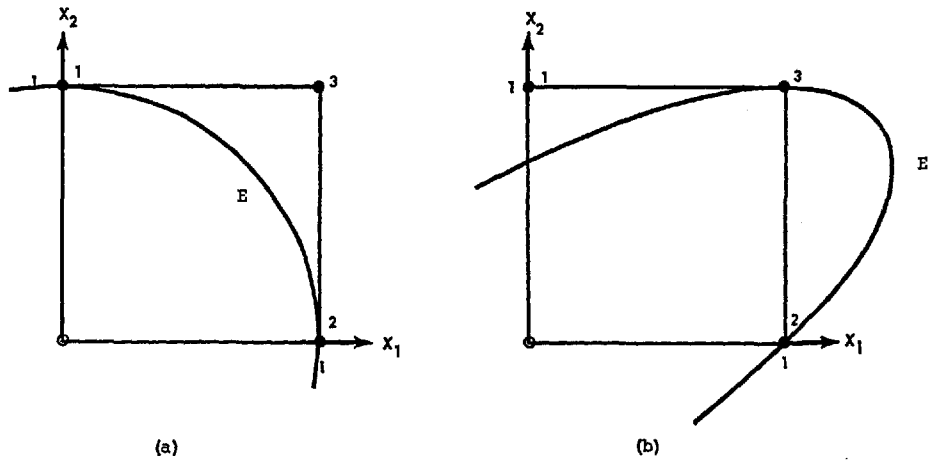

FIGURE 3
The ellipse E for two designs for the linear model $\eta = \theta_1\xi_1 + \theta_2\xi_2$.

5. Application of Results to Nonlinear Models

5.1. *Two Consecutive Irreversible First-Order Reactions—Model* 1

As a first example, we return to the first model discussed in this paper. The model describes the amount of an intermediate product resulting from an irreversible first-order reaction, this product itself being subject to first-order decay. The optimal design consists of experiments at times of 1.23 and 6.86 units. The corresponding values of x_1 and x_2 are:

Experiment	ξ_1	x_1	x_2
1	1.23	0.4406	−0.3405
2	6.86	−0.1174	−1.7485

The equation of E is therefore

$$4.832x_1^2 + 0.2116x_2^2 - 0.1682x_1x_2 = 1.$$

The ellipse E is shown plotted together with $R(\mathrm{x})$ in Figure 4. It can be seen that $R(\mathrm{x})$ is contained by E so that, as a result of the second theorem, optimal designs for an even number of experiments must consist of replications of the two-point design. We have already shown by example that this is so. (Table 1).

5.2. *The Catalytic Dehydration of Hexyl Alcohol—Model* 2

It was shown in Section 3 that the optimal two-run design for two-parameter Model 2 (Equation 3.5) consists of the following experiments:

Experiment	ξ_1	ξ_2	x_1	x_2
1	0.3448	0	0.0595	0
2	3.0	0.7951	0.0588	−0.0127

The attainable region $R(\mathrm{x})$ for this model is plotted in Figure 2. In Figure 5 the boundary of $R(\mathrm{x})$ is shown together with E, by which it is contained.

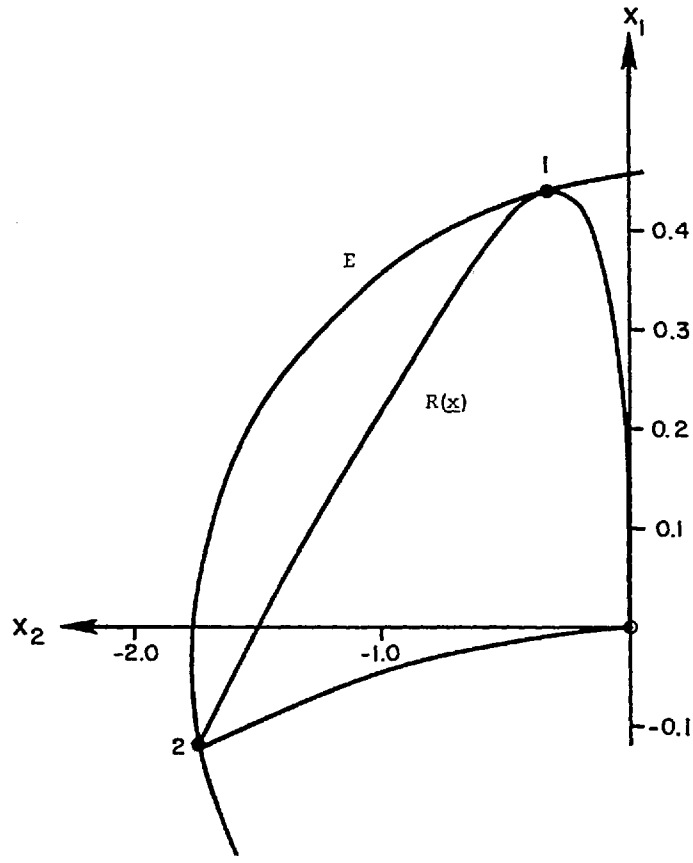

FIGURE 4
$R(\mathbf{x})$ and E for Model 1.

This is in agreement with Theorem 2 and the results of the searches for optimal designs given in Table 2.

5.3. *Two Other Nonlinear Models*

Our results were applied to two further nonlinear models occurring in the literature. One is that developed by Behnken (1) in his study of binary copolymers formed from monomers of differing activities. The other, a first-order decay model with rate a function of temperature, is described by Box and Lucas, who show $R(\mathbf{x})$ in their Figure 4. In both cases $R(\mathbf{x})$ is contained by E and the optimal design for np experiments was found by computer search as before to consist of n replications of the optimal design for p experiments.

6. AN ITERATIVE APPROACH TO THE OPTIMAL DESIGN

With the preceding examples, the main part of this paper is complete. In this section we describe an iteration leading to the optimal design for p experiments.

Since there are k controllable variables, the value of the design criterion for p experiments is a function of $p \times k$ variables. If p and k are not both small, identification of an optimal design can prove difficult. The iteration which we suggest replaces the search in $p \times k$ dimensions with a series of optimizations in k dimensions.

The optimal design consists of a set of points which form a simplex of maximum volume within $R(\mathbf{x})$. If $p - 1$ of the points are fixed, finding conditions for the remaining experiment which maximize Δ_p requires an optimization in only k dimensions. Once this point has been fixed, finding the best conditions for another point cannot lead to a smaller value of the determinant and will usually lead to a greater one. The process is repeated for each of the design points and then continued until the optimal design is achieved.

For this iteration to succeed, it is necessary that the boundary of $R(\mathbf{x})$ be reasonably well behaved. For each of the four models studied, the boundary

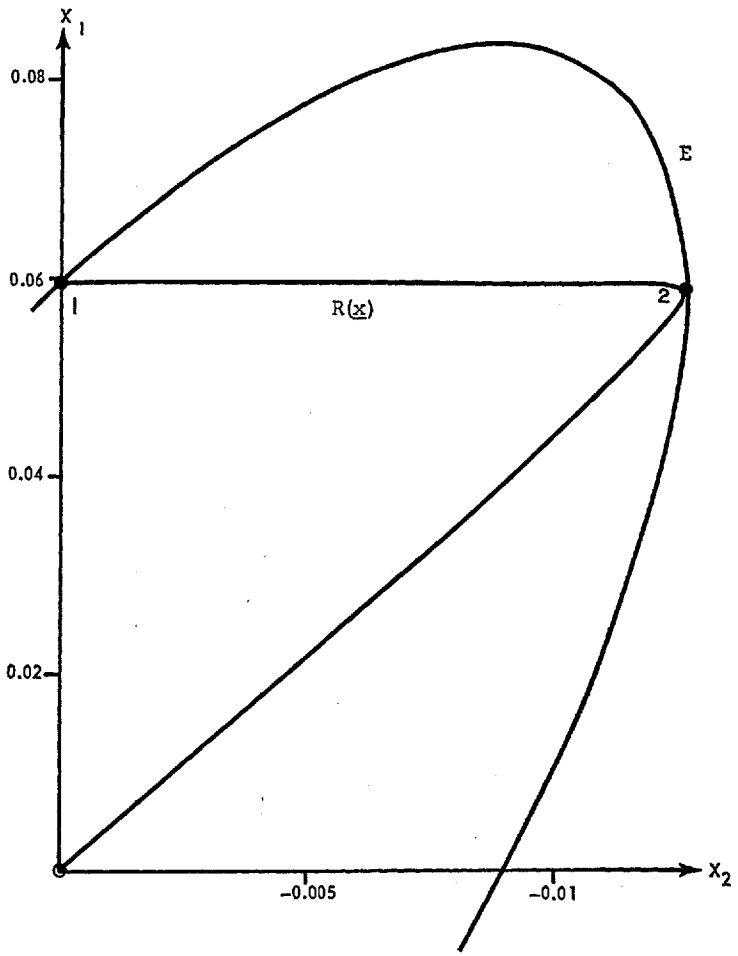

FIGURE 5
The boundary of $R(\mathbf{x})$ and E for Model 2.

is well behaved. Even in cases where $R(\mathbf{x})$ is poorly behaved, selecting starting conditions at random and repeating the process several times should lead to a good design.

As an example of this method, we once again consider the example of two successive first-order reactions, $A \rightarrow B \rightarrow C$. Suppose, as shown in Figure 6, that the first guess of an experimental point was at a time of 30 units. Then the best conditions for a second experiment would be at a time of 1.17. Geometrically, this is the point where the tangent to $R(\mathbf{x})$ is parallel to the vector from the origin to the first point. The second iteration finds the best condition for a second experiment when one is performed at a time of 1.17. This at 6.85 units. The third iteration gave a design consisting of experiments at times of 1.23 and 6.85 units, which is close to the optimal design. The results for this and another series of iterations are shown in Table 3. In both cases the optimal design was achieved in three iterations.

The rate of convergence will depend upon the shape of the boundary of $R(\mathbf{x})$. If $R(\mathbf{x})$ were triangular, only two iterations would be needed, whereas if $R(\mathbf{x})$

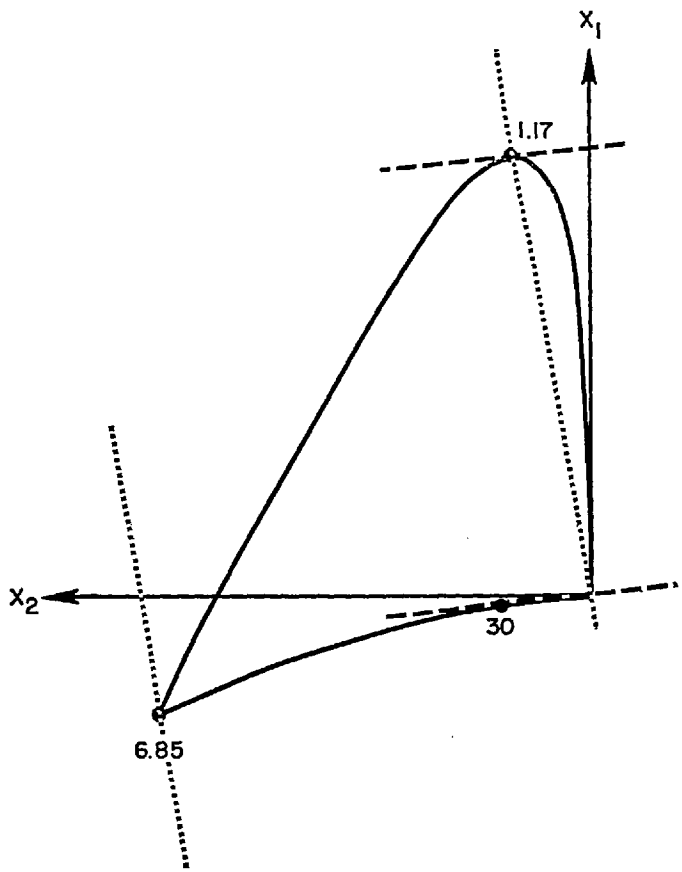

FIGURE 6
An iterative approach to the optimal design for Model 1.
Pairs of similarly dashed lines are parallel.

TABLE 3
Iterative approach to the optimal design for two experiments for Model 1

	ξ_{11}	ξ_{12}	Δ_2
Case 1	1.17	30.0*	0.0019
	1.17*	6.85	0.6551
	1.23	6.85*	0.6568
Case 2	2.0*	6.96	0.4841
	1.23	6.96*	0.6565
	1.23*	6.86	0.6568

* In each step in the iteration, the asterisked value was kept constant, and the value of ξ_1 which gave the optimal design in combination with the fixed value was calculated.

approximated E, convergence would be much slower. However, in the latter case, if convergence were not completely achieved, the designs would only be fractionally less efficient than the optimal design.

7. Conclusions

We have discussed the situation in which experiments are to be performed to estimate the p parameters in a model of known form. We have assumed that a given number of experiments are to be run before any results are available for analysis. Under these conditions we have given a necessary condition for the optimality of a p-point design and established a sufficient condition for replication of the optimal p-point design to be optimal for $N = np$ experiments.

A result of some practical importance that suggests itself on the basis of this work is that in many cases in order to select the best $N > p$ experiments it will be sufficient simply to find the best set of p experiments, the optimal design consisting of replications of these p experiments. The function Δ, then, needs to be maximized in only $p \times k$ dimensions instead of $N \times k$ dimensions. In some circumstances this reduction in the number of dimensions may represent considerable savings. An iterative search procedure for finding the best set of p experiments has been presented.

Appendix

Proofs of Theorems.

Theorem 1. A necessary condition for the optimality of a p-point design is that it shall consist of points on $R(\mathbf{x})$ such that $R(\mathbf{x})$ does not lie outside the p-dimensional parallelepiped defined by the p pairs of planes

$$\delta^2_{p+1,-u} = \Delta_p, \qquad u = 1, 2, \cdots, p. \tag{A.1}$$

Proof. When $N = p$, the value of the design criterion $\Delta_p = |\mathbf{X}|^2$ is proportional to the square of the volume of the simplex formed by the origin and the p experimental points in the X-space. Disregarding, for the moment, any restrictions on the conditions of experimentation due to the boundary of $R(\mathbf{x})$, con-

sider the locus of points $(p + 1)$ such that experimenting at the original p points is as good as experimenting at the set with the point u replaced by the point $(p + 1)$. Using the nomenclature of Section 4.2, the locus of the point $(p + 1)$ will be given by those values satisfying the relationship

$$\delta^2_{p+1,-u} = \delta^2_{p+1,-(p+1)} = \Delta_p. \tag{A.2}$$

The locus will be such that the volumes of the simplexes formed by the two sets of points will be equal and will consist of the pair of planes through the point u and its reflection in the origin, parallel to the plane defined by the origin and the other $(p - 1)$ points. If $R(\mathbf{x})$ is such that, for some u, it lies in part outside this pair of planes, then the squares of the value of the determinant formed by a point in that part of $R(\mathbf{x})$, the other $(p - 1)$ points and the origin will be greater than Δ_p , and so the original p points cannot form an optimal design.

Theorem 2. A sufficient condition for the optimal design for $N = np$ experiments, $n = 2, 3, \cdots$, to consist solely of replications of p points which are optimal for $N = p$ is that $R(\mathbf{x})$ be contained by the p-ellipsoid E defined by Equation 4.9.

Proof. When the number of experiments is greater than the number of parameters, the design criterion Δ_N has a geometrical interpretation (8). The determinant Δ_N is equal to the sum of the squares of the values of the $N!/p!\ (N - p)!$ determinants formed by taking the N rows of $\mathbf{X}$ p at a time, where the absolute value of each determinant is again proportional to the volume in the X-space of the simplex formed by the origin and p of the experimental points. Since our object is to maximize the value of Δ_N , it follows that any optimal design must consist of points on the boundary of $R(\mathbf{x})$.

Now consider a design for $p + 1$ experiments. The matrix $\mathbf{X}$ is $(p + 1) \times p$. The determinant Δ_{p+1} will equal the sum of $p + 1$ terms, each of which will be the square of the determinant of the $p \times p$ matrix formed by removing one row from $\mathbf{X}$. Thus we have

$$\Delta_{p+1} = \sum_{u=1}^{p+1} \delta^2_{p+1,-u}. \tag{A.3}$$

Suppose that the first p experiments are fixed, giving a value of Δ_p for the determinant. Repeating any one of these points will give a value of $2\Delta_p$ for Δ_{p+1} . Then the locus of points $(p + 1)$ which, together with the original design, give the same value of the determinant as replicating one of the original points is given by those values satisfying

$$\sum_{u=1}^{p+1} \delta^2_{p+1,-u} = 2\Delta_p \tag{A.4}$$

or, since $\Delta_p = \delta^2_{p+1,-(p+1)}$,

$$\sum_{u=1}^{p} \delta^2_{p+1,-u} = \Delta_p \tag{A.5}$$

which is the equation of E. Comparison of Equations A.5 and A.1 shows that

E is contained by the pairs of planes, which touch E only at the points u and u'. Therefore, the set of p points defining E is such that the tangent to the hyperellipsoid at each point is parallel to the plane defined by the origin and the other $p - 1$ points. These points then form a set of conjugate points, and the volume of the simplex formed by them and the origin is a simplex of maximum volume of this kind within E. This proves that, when $R(\mathbf{x})$ is contained by E, the design is an optimal design for $N = p$ experiments.

When E is transformed so that it becomes a p-sphere E^T, $R^T(\mathbf{x})$ the transformed $R(\mathbf{x})$ will be contained by E^T, the transformed E, and the p points defining E^T will be the vertices of an orthogonal simplex, the p-dimensional analogue of a right-angled triangle. The problem is thus reduced to that of designing experiments within a p-sphere. It has been shown by Box and Hunter (3) that, for a given N, any design consisting of points on a sphere at the vertices of the regular solids in any combination and in any absolute or relative orientation will have the same maximum value for the determinant Δ_N, provided N can be so divided. Points at the vertices of the orthogonal simplex are members of this class, so that when $N = np$, replication of this design will give an optimal design. If $R^T(\mathbf{x})$ lies inside E^T at all points other than the p defining E^T (and hence if $R(\mathbf{x})$ lies inside E at all points other than the p defining E), then the optimal design will consist of replications of these p points. The proof is thus complete.

Acknowledgements

The greater part of this work was performed when the authors were at Imperial College, London, during the academic year 1964–65. Mr. Atkinson gratefully acknowledges the receipt of a Science Research Council Advanced Course Studentship during this time.

References

1. Behnken, D. W., 1964. Estimation of copolymer reactivity ratios: an example of nonlinear estimation, *Journal of Polymer Science, Part A, 2*, 645–668.
2. Box, G. E. P., 1952. Multifactor designs of first order. *Biometrika, 39*, 49–57.
3. Box, G. E. P. and Hunter, J. S., 1957. Multifactor experimental designs for exploring response surfaces, *Ann. Math. Stat., 28*, 195–241.
4. Box, G. E. P. and Hunter, W. G., 1965. Sequential design of experiments for nonlinear models, *Proceedings of the IBM scientific computing symposium on statistics*, October 1963, 113–137.
5. Box, G. E. P. and Lucas, H. L., 1959. Design of experiments in nonlinear situations, *Biometrika, 46*, 77–90.
6. Hotelling, H., 1944. Some improvements in weighing and other experimental techniques, *Ann. Math. Stat., 15*, 297–306.
7. Mood, A. M., 1946. On Hotelling's weighing problem, *Ann. Math. Stat., 17*, 432–446.
8. Wilks, S. S., 1962. *Mathematical statistics*, John Wiley, New York.

A Mathematical Basis for the Selection of Research Projects

ANTHONY C. ATKINSON AND ARTHUR H. BOBIS

Abstract—**A method is presented for determining the money to be spent on product oriented research programs. The method is used by American Cyanamid's Organic Chemicals Division.**

INTRODUCTION

DURING the past few years many articles have appeared describing mathematical models intended for use in evaluating research projects. These can be broadly separated into two categories: simple models that treat the process of research as if it were static, and complex models that deal with research as a dynamic problem. In reviewing this literature, Baker and Pound [1] state that the simple models ignore what is perhaps the essential aspect of the problem and are, therefore, of questionable utility, while the more complicated models require so much information that their usefulness can never truly be tested. With this paper, we are adding to the proliferation of methods and models. But we believe we have chosen a course that avoids the previous inadequacies and that, while dealing with the real problem, treats it in such a way as to encourage use of the method. The justification for our belief is that the model has been developed in consultation with groups of researchers, and that management finds the results helpful in evaluating and planning research efforts. Our point of departure for this work was the paper by Hess [2].

The purpose of this paper is to describe the mathematics on which the model is based. By mathematics we mean both the algebra and the philosophy that led to the particular algebraic formulation. The description is in five parts:

1) the probability model
2) commercial information
3) the rate of expenditure
4) optimization
5) simulation.

An Appendix describes the distribution used to fit data.

In building a mathematical model of the research process, we have had to make use of estimates of such quantities as the probability that a project will succeed. Although there is no means of checking such a number, and also no "frequentist" interpretation of such probabilities, we have handled subjective probabilities as though they described recurring events. Also, no attention has been given to problems arising from optimistic or pessimistic estimations of such quantities as sales. We have worked on the assumption that decisions have to be made, usually on inadequate data, and that they will be made intuitively on the data, whatever its quality. Anything that can be done to quantify the bases for these decisions and to demonstrate the logical consequences of the assumptions is a step in the right direction.

Twice a year the research managers in the American Cyanamid Company's Organic Chemicals Division collect data of the kind described in this paper and submit it for analysis. This is done at the time budgets are being prepared and again when we are reviewing the results of these expenditures. The computer evaluation, which is an allocation of research funds across the several projects, is used as a guide for the eventual budget. The analysis often points out shortcomings in the information. At this point the necessary modifications are made and the procedure repeated until there is general agreement that the analysis is based on the best possible information. Although the computer allocation and the eventual budget need not be identical, serious departures between the two must be examined.

As will be shown, the model is general; there is nothing that is specific to the needs of the American Cyanamid's Organic Chemicals Division, nor is its use restricted to the chemical industry. It has been formulated so that it can be applied to those research projects whose objective is either the invention of new products or modification of existing products that are sufficiently well defined that sales and selling price are estimatable.

The model as described limits the projects that are considered to those where technical success or failure will be determined within the ensuing five years. It has been our experience that in the product-oriented research projects that have been considered, few have been eliminated from the analysis on this basis.

The analysis that is described is based on accumulated net profits over the next 11 years. This choice is arbitrary and some other basis could have been chosen. Another choice would have been to calculate the returns from each project over the product's expected life. The choice of 11 years has proven satisfactory. Returns beyond the eleventh year are subject to increasing variation, while discounting severely reduces the revenue. The choice of a shorter basis would, it is felt, unduly penalize the long-range project.

Manuscript received February, 1968; revised November, 1968.

A. C. Atkinson was with the American Cyanamid Company, Bound Brook, N. J. He is now with Imperial College, London, S.W.7, England.

A. H. Bobis is with the American Cyanamid Company, Bound Brook, N. J.

I. The Probability Model

The distribution of the cost of completing a project is defined by the least, most likely, and greatest expected costs of completion, Rx_1, Rx_2, and Rx_3. These are the $2\frac{1}{2}$ percent mode and $97\frac{1}{2}$ percent points of the distribution. By completion we mean that the project has reached a stage where it is either manifestly a failure, or where it is bound to succeed, perhaps given some further expenditure. In the absence of any information to the contrary, the distribution of the cost would be expected to be normal. For convenience, we have used the similarly shaped logistic distribution with a transformation to allow for skewness. The probability C that the project will be completed for an expenditure no greater than X is given by

$$C = 1/[1 + \exp(\alpha - \beta X^a)]. \tag{1}$$

The method used for obtaining the constants a, α, and β from Rx_1, Rx_2, and Rx_3 is described in the Appendix.

Although (1) gives the probability of completion as a continuous function of the research expenditure, we have chosen to consider the expenditure as occurring as a series of annual budgets x_i, up to a maximum of 5. Let the total research expenditure through year i be X_i. Then,

$$X_i = \sum_{j=1}^{i} x_j. \tag{2}$$

Rewriting (1), the probability of completion by the end of the ith year C_i becomes

$$C_i = f_1(X_i) \tag{3}$$

and the probability of completion in the ith year c_i is given by

$$c_i = C_i - C_{i-1} = f_1(X_i) - f_1(X_{i-1}) \tag{4}$$

with $C_0 \equiv 0$.

The completion can result either in a success or a failure. The overall probability of success, given project completion P_s, is calculated by multiplying together the probabilities of technical, legal, engineering, and commercial success, since failure in any one of these areas leads to failure for the project. These probabilities refer, respectively, to the event that we can make a satisfactory product in the laboratory, that the route will not be blocked by competitive patents, and that the process can be scaled up. The last probability, that of a commercial success, refers to the chances that we shall be able to sell the product that has been specified.

The probability of success in year i, p_i, is the probability of completion in that year scaled by the overall probability of success. Using, as before, capital letters to denote cumulative probabilities, and small letters to denote annual probabilities, we have

$$p_i = Ps\, c_i \tag{5}$$

$$P_i = Ps\, C_i. \tag{6}$$

The probability that the project fails in year i is, therefore, $c_i - p_i = (1 - Ps)c_i$, and that it is not completed by the end of year i is $1 - C_i$. This is the probability that research expenditure will be required in succeeding years. The situation is illustrated in Fig. 1.

II. Commercial Information

The return from the investment in research comes from sales of the product; we do not consider the possibility of licensing agreements, nor attempt to give any monetary value to the increased knowledge and expertise that may result from a project, even an unsuccessful one.

The sales estimates we require are those in the first, sixth, and eleventh years, assuming that sales could begin on January 1 of year 1. Call these estimates E_1, E_2, and E_3. New product sales in general follow an S-shaped curve, which may be conveniently represented by the cumulative of the logistic distribution, suitably scaled. Let s_j be the sales in year j. Then,

$$s_j = \frac{B}{[1 + \exp(\gamma - \delta j)]} \tag{7}$$

or

$$s_j = f_2(j) \tag{8}$$

where B is the asymptotic value toward which the sales tend. The values of the constants are

$$B = \frac{E_2(E_1E_2 + E_2E_3 - 2E_1E_3)}{(E_2^2 - E_1E_3)} \tag{9}$$

$$\delta = 0.2 \ln\left[\frac{E_3(B - E_2)}{E_2(B - E_3)}\right] \tag{10}$$

$$\gamma = \ln\left[\frac{(B - E_2)}{E_2}\right] - 6\delta. \tag{11}$$

For this model to apply, it is mathematically necessary that $\sqrt{(E_1E_3)} < E_2 < E_3$. Otherwise, the preceding equations involve the logarithms of negative numbers. Physically, this implies that the sales must not dip by the end of the period, nor must they increase too rapidly between the sixth and eleventh years.

Equation (7) gives the sales for any year, assuming that they start at the beginning of the first year. Starting later will not only cause the sales to lag behind but, in a competitive market, could cause the ultimate market share to be reduced. This is estimated from E_4, the sales for year 11 assuming that the product is not available until the beginning of year 6. Each year's delay will cause the ultimate sales B to be reduced by a factor k, so that after five years

$$k^5 = \frac{E_4}{E_2}. \tag{12}$$

Sales in the jth year, when the first year of sales is i, are from (8) and (12) given by

$$s_j = k^{i-1} f_2(j - i + 1). \tag{13}$$

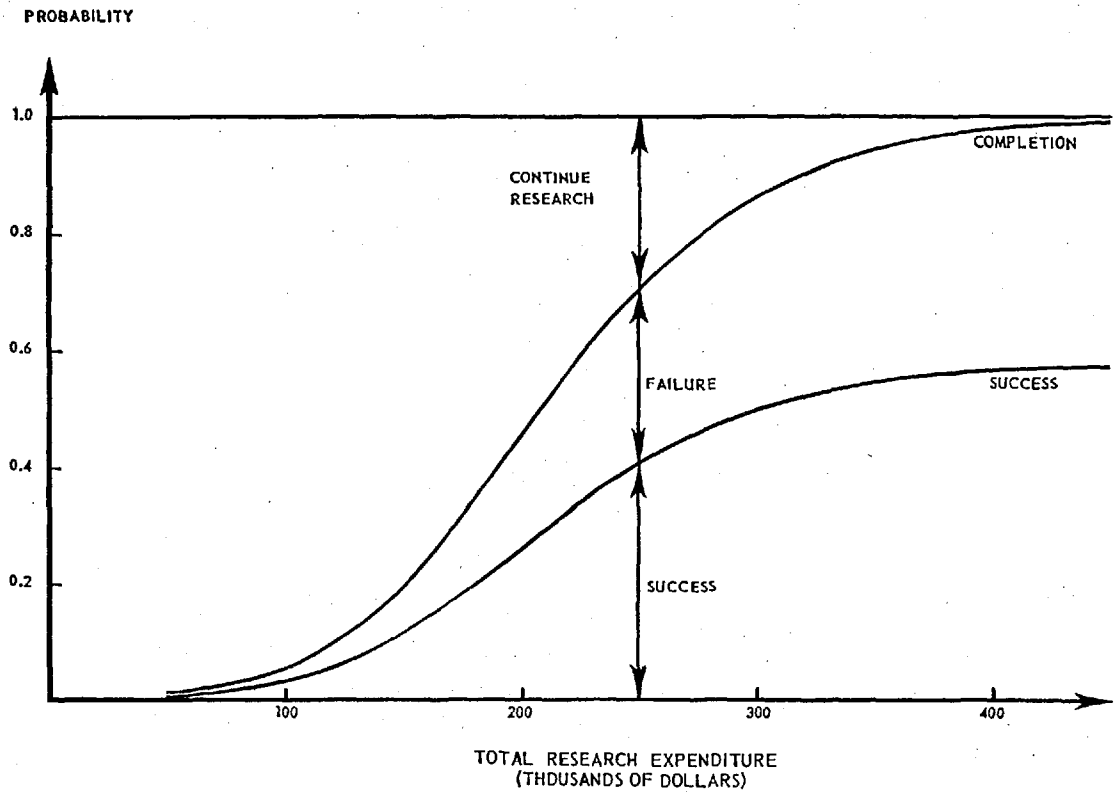

Fig. 1. The probabilities of success, failure, and of continuing research as a function of total research expenditure.

Due to the necessity of test marketing and building a plant, the year in which sales start will be later than the year in which research succeeds by some number of years L. For sales starting the year after the research success, $L = 1$. Then s_{ij} sales in year j following a research success in year i are, from (13),

$$s_{ij} = k^n f_2(j - n), \tag{14}$$

where

$$n = i + L - 1. \tag{15}$$

We have assumed that the selling price in year j, Q_j, varies linearly with time. The relationship is defined by the selling prices in the first and eleventh years. The manufacturing cost M_j may be similarly defined, or may be calculated from the selling price, if one of the research objectives is to achieve a given profit on sales. In this latter case, the desired profit on sales is given by

$$M_j = Q_j\,(1 - W). \tag{16}$$

The final expenses to be considered are overheads and selling costs H, which are a constant fraction of the sales revenue. With this information we are now able to calculate G_{ij}, the money accruing to the company from sales in year j as a result of a research success in year i.

$$G_{ij} = s_{ij}[Q_j(1 - H) - M_j]\left[\frac{1}{(1 + D)}\right]^{j-1} \tag{17}$$

where D is the discount factor.

Summing (17) over the lifetime of the product gives G_i, the payoff from a success in year i.

$$G_i = \sum_{j=i+L}^{N} G_{ij}, \tag{18}$$

where N is the last year in which sales are considered—usually, but not necessarily, year 11.

The expected payoff G is, from (5) and (18),

$$G = \sum_{i=1}^{5} G_i p_i. \tag{19}$$

To calculate the expected profit from the project, the research costs must be deducted from this figure. The budgeted amount will be spent only if the project has not been completed. The probability of this event in year i is $(1 - C_{i-1})$. If the project is a success, some amount of money ϕ will have to be spent each year in defensive work to keep the product up to date and competitive. The probability that the project is successful by year i is P_{i-1}. The expected research expenditure in year i, R_i, is therefore

$$R_i = (1 - C_{i-1})x_i + P_{i-1}\phi \qquad i \leq 5 \tag{20}$$

$$R_i = P_5\phi \qquad i > 5. \tag{21}$$

Summing and discounting, the expected discounted research expenditure for the project R is

$$R = \sum_{i=1}^{N} R_i\left[\frac{1}{(1 + D)}\right]^{i-1} \tag{22}$$

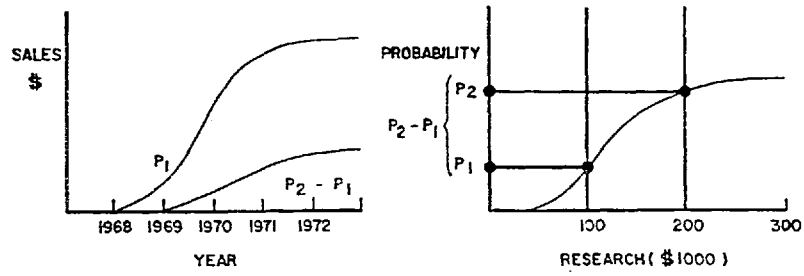

Fig. 2. Associating probability of success with levels of sales using the model.

The expected profit of the project in year 1 dollars, V, is the difference between the expected payoff and the research expenditure

$$V = G - R. \tag{23}$$

The value of V, the expected return from the project, depends, as we have shown, on the commercial information, the amount of money spent on the research project, and the way in which that money is spent over time.

III. Rate of Expenditure

The method by which probabilities are related to the various sales curves is illustrated in Fig. 2. Here, two sales curves, one starting in 1968 and the other in 1969, are shown on the left-hand side of the figure. The probabilities of success associated with different levels of spending are shown on the right-hand side. It is by expenditures of $100 000 in each of 1968 and 1969 that the probabilities P_1 and $(P_2 - P_1)$ of following the 1968 and 1969 curves result. It is assumed that, in this case, commercialization commences with success. The penalty for succeeding late is shown to be severe, and consideration should be given to accelerating the rate of research spending. A $200 000 expenditure in a single year could be considered but, although the total quantity is the same, this may not be as effective as spending the money over a two-year period. The probability of success associated with a single large expenditure will, of course, be greater than P_1 and cannot be greater than P_2. The value of this probability is the subject of this section.

Increasing the expenditure of research funds at the beginning of the project will result in an increased probability of success in the initial years and so lead to an increased value of V. However, too great an expenditure could result in the wastage inherent in a crash program. We have, therefore, defined the quantity y_i, the efficient research expenditure for year i.

If we let z_i be the effective research expenditure for year i, we have from (3)

$$C_i = f_1(z_i) \tag{24}$$

with

$$z_i = y_i\left(\frac{x_i}{y_i}\right)^{\epsilon_i} \qquad x_i > y_i, \quad \epsilon_i < 1$$
$$z_i = x_i \qquad \text{if} \quad x_i < y_i. \tag{25}$$

For expenditures greater than y_i, (25) guarantees that there will be a gradual decrease in the efficiency with which incremental expenditures above y_i are used. It remains to determine reasonable values for ϵ_i.

As the project proceeds, the value of ϵ_i should increase, for increased knowledge will lead to more efficient use of extra resources. When the total effective research expenditure equals the greatest cost of completing the project Rx_3, we have set C_i equal to 1. Initially, the ratio of the efficient expenditure for the first year y_1 to Rx_3 is a measure of how well defined the project is. The smaller this ratio, the less the amount of the program that will be completed in the first year. The value of ϵ_i is set equal to this ratio for the first year and increased in subsequent years with the total effective expenditure. We have

$$\epsilon_1 = \frac{y_1}{Rx_3} \tag{26}$$

and

$$\epsilon_i = \frac{(y_i + z_{i-1})}{Rx_3} \qquad i = 2, \cdots, 5. \tag{27}$$

Finally, the amount spent in one year will affect the amount that can be spent efficiently in the next. If no money is spent, some effort in the succeeding year will be devoted to training staff. If the rate of expenditure is increased above y_i, an increased number of trained people will be available for the succeeding year. Let y_i^0 be the original estimate of the efficient expenditure in year i. Then, if we write

$$y_i = y_i^0\left[1 + \frac{1}{2}\left(\left[\frac{z_{i-1}}{y_{i-1}^0}\right] - 1\right)\right] \tag{28}$$

we have a relationship that allows for increasing and decreasing y_i, depending upon the expenditure in the preceding year.

The mechanism just described penalizes expenditures above the efficient level to the greatest extent at the early stages of the project. It is felt that at this point the additional money is usually spent on parallel exploration and part of it must, by definition, be wasted. Accelerating the program in the latter stages is apt to be accomplished by carrying out definitive and necessary aspects of the research in parallel. In justification, the allocations suggested by this method, although different from those made by management, do seem reasonable.

IV. Optimization

We have described a mechanism for calculating the expected profit from any pattern of research expenditures. In determining an optimum allocation, we maximize the expected return V from a set of L projects, subject to some budgetary constraint T.

The budget for a department may be represented by the L by 5 matrix **X**, where the kth row is the money budgeted for project k and the ith column is the money that will be spent on each project in year i if the project has not been completed. Formally, the optimization consists of finding

$$\sup_{\mathbf{X}\in T} V = \sup_{\mathbf{X}\in T} \sum_{k=1}^{L} V_k. \tag{29}$$

The constraint is that the total expected research expenditure for year i should be no greater than some amount T_i. Then from (20), the constraint is that

$$\sum_{k=1}^{L} R_{ki} = \sum_{k=1}^{L} [(1 - C_{k,i-1})x_{k,i} + P_{k,i-1}\phi_k] \leq T_i. \tag{30}$$

This is a formidable optimization problem, not only because of the number of variables, but because the linear constraint on the budgets in year i depends upon and varies with the expenditures in preceding years. We have attacked this problem using an iterative scheme in which the annual budget across all projects is optimized year by year and the process repeated until no further improvement is noticed. Apart from the reduction in the number of variables in each optimization from 5 L to L, the coefficients of the $x_{k,i}$ in (30) are constant for the optimization, depending as they do only on expenditures in earlier years. In this iterative scheme, the constraint may now be written as

$$\sum_{k=1}^{L} (1 - C_{k,i-1})x_{k,1} \leq T_i - \sum_{k=1}^{L} P_{k,i-1}\phi_k. \tag{31}$$

The latter term on the right-hand side is the amount of money that is already committed to fixed research costs as a result of success in preceding years. For a particular year, the quantity on the right-hand side depends only on expenditure in preceding years and, in the iterative scheme, is therefore constant. The constraint is thus of the form

$$\sum_{k=1}^{L} b_i x_i \leq T' \tag{32}$$

with, since the x_i cannot be negative, the L additional constraints

$$x_i \geq 0 \qquad i = 1, 2, \cdots, L. \tag{33}$$

These constraints define a simplex in L dimensions within which all possible allocations must lie. The constraints (33) can be dispensed with and (32) made a strict equality by making the transformation $w_i^2 = x_i$ and by introducing the slack variable x_{L+1}. In W space the constraint is

$$\sum_{i=1}^{L+1} b_i w_i^2 = T'. \tag{34}$$

This defines an ellipsoid in $L + 1$ dimensions on the surface of which all possible allocations must lie. Making the further substitution

$$\eta_i^2 = \frac{b_i w_i^2}{T'} \tag{35}$$

the constraint becomes

$$\sum \eta_i^2 = 1, \tag{36}$$

where each η_i^2 must therefore lie between 0 and 1. Writing

$$\begin{aligned} \eta_1^2 &= \cos^2 \theta_1 \\ \eta_2^2 &= \sin^2 \theta_1 \cos^2 \theta_2 \\ &\vdots \\ \eta_L^2 &= \cos^2 \theta_L \prod_{i=1}^{L-1} \sin^2 \theta_i \\ \eta_{L+1}^2 &= \prod_{i=1}^{L} \sin^2 \theta_i \end{aligned} \tag{37}$$

the constraint (36) will be satisfied for any values of the θ_i. The optimal budget may, therefore, be found by an unconstrained optimization in θ space.

The results of these optimizations, using the mechanism developed in Section III for the effect of the rate of expenditure, give, for the research programs studied, significantly increased expected returns. This is usually achieved by increasing expenditure on the more profitable projects at the expense of dropping some of the less profitable. The usual research policy seems to be to spread limited resources over too many alternatives, with the result that only a very few projects are completed early enough to make a real impact on the market.

V. Simulation

The numbers used in the calculation of the payoff are estimates and are subject to error. It is, therefore, meaningful to obtain ranges about these estimates. The numbers so obtained define the distribution of the variable as the cost of completing the project is defined by the Rx_i.

With point estimates replaced by distributions, the values of the payoff used in calculating the expected return are obtained by simulation. The distribution of each variable is randomly and independently sampled to establish a particular case and the payoff calculated for that case. The average value of the payoff obtained by repeating this process many times is an unbiased estimate of the expected value of the payoff.

The predicted sales in year 11 are E_3. Let the range about this value be E_5 and E_6. The distribution of year 11 sales will be given by (1) with the values of the constants calculated from E_5, E_3, and E_6. To sample

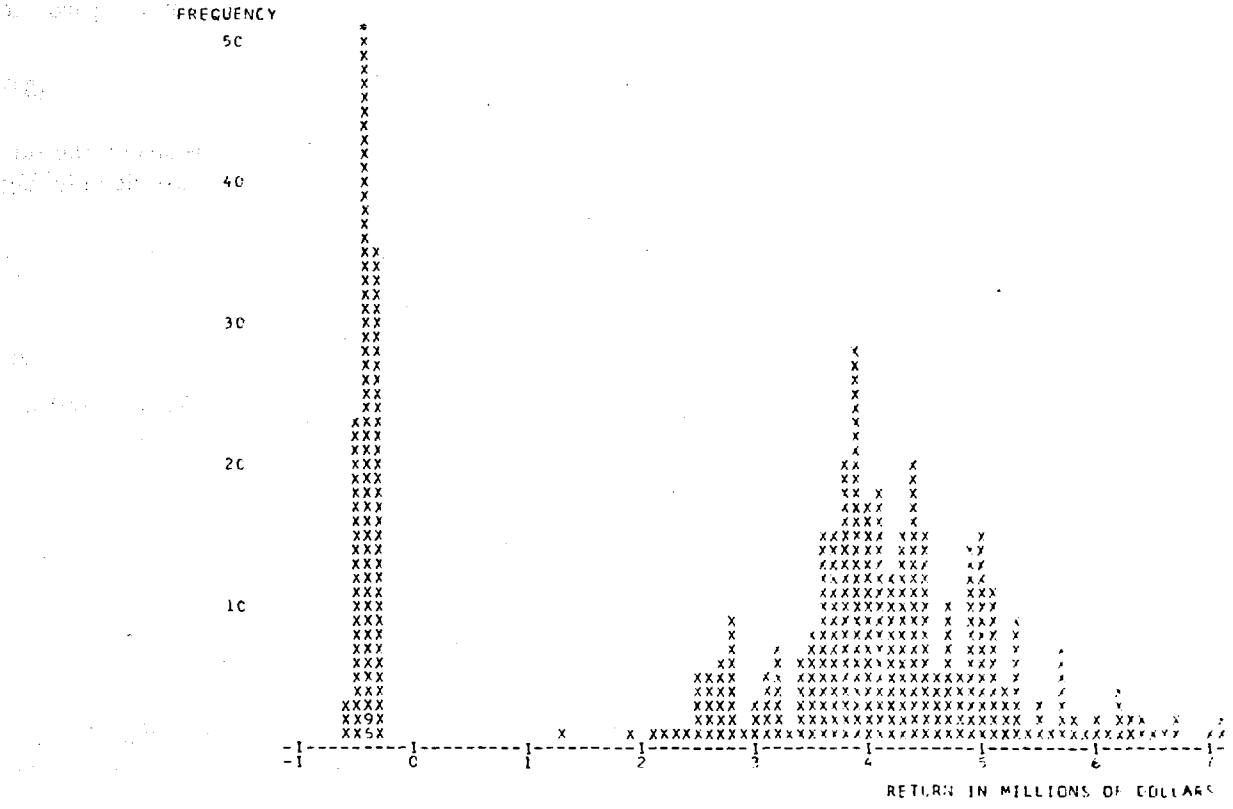

Fig. 3. The distribution of return for a typical project.

from this distribution let U be a random variable uniformly distributed between 0 and 1. Then E', the corresponding value of year 11 sales, is given by

$$E' = \left[\frac{(\alpha + \log U - \log [1 - U])}{\beta}\right]^{1/\alpha}. \tag{38}$$

The percentage uncertainty of the sales estimate will increase with time. Let

$$Fs = \frac{(E' - E_3)}{10E_3}. \tag{39}$$

Then the value of the simulated sales in year j will be

$$E_j' = E_j[1 + (j - 1)Fs]. \tag{40}$$

Similar methods are used to establish values of the other variables for the particular case. The calculation of the payoff is then as described in Section II.

The values so obtained lead to the calculation of the expected value of the project. In order to describe the project fully, we need to know the probability of success and failure in each year and the monetary value of each event. This information is presented as a histogram, see Fig. 3, showing the frequency distribution of the return. This is built up by a two-stage simulation.

In the first stage, we sample to find out whether the project succeeded or failed, and in which year. Recalling that the overall probability of success is Ps, letting U again be a random number between 0 and 1, and remembering that if the project is not completed by the end of year 5 it is counted as a failure, we have the following distribution of events:

$0 < U < P_1$	success in year 1
$P_{i-1} < U < P_i$	success in year i
$Ps < U < Ps$	failure after five years
$Ps + C_{i-1}(1 - Ps) < U$ $< Ps + C_i(1 - Ps)$	failure in year i
$Ps + C_5(1 - Ps) < U < 1$	failure after five years.

(41)

If the project fails, the return is a loss of the research money. Let R_i' be the simulated return for this case. Then

$$R_i' = \sum_{j=1}^{i} x_j\left[\frac{1}{(1 + D)}\right]^{j-1}. \tag{42}$$

If the project suceeds, the second stage of the simulation is to calculate the payoff for success in year i in the manner described at the beginning of this section. Let this payoff be G_i'. Then

$$R_i' = G_i' - \sum_{j=1}^{i} x_j\left[\frac{1}{(1 + D)}\right]^{j-1} - \sum_{j=i+1}^{N}\left[\frac{1}{(1 + D)}\right]^{j-1}. \tag{43}$$

The distribution of the R_i', usually 500 in number, is presented as a histogram. Typically, as is shown in Fig. 3,

the histogram is bimodal, the left-hand peak representing failure and the right-hand success. The greater the probability of success, the greater is the area of the right-hand peak. For more profitable projects, this peak is further to the right. Projects that may succeed in several years and those in which the estimates of the variables are imprecise result in a more diffused right-hand peak. The combination of graphical output and optimization results in a powerful tool, both for describing the outcome of a given pattern of expenditure and for determining what that pattern should be.

We do, however, have to admit that there is one question we have not been able to answer; that is, what is an optimal allocation of research funds? The approach we have taken is to maximize the expected profit from the program. In the absence of any other criterion, this is not an unreasonable choice. Indeed, if the program were repeated infinitely, it would be the best choice of strategy. The program is performed only once, however. If a project fails, a loss will be sustained. It is possible that conditions could arise under which, however great the expected value of a program, the possible losses are too large and not likely to be tolerated. It is allocation under this kind of constraint that we have not considered.

Appendix
The Logistic Distribution

The logistic distribution is a continuous probability function with cdf

$$F(x) = \frac{1}{[1 + \exp(\alpha - \beta x)]} \tag{44}$$

and pdf

$$f(x) = \frac{\beta \exp(\alpha - \beta x)}{[1 + \exp(\alpha - \beta x)]^2} \tag{45}$$

for all x.

It is symmetrical about α/β, with variance $\pi^2/3\beta^2$. In shape it is similar to the normal distribution, but with more probability in the tails.

To represent a skewed distribution, suppose $u = x^a$ has the logistic distribution. The distribution of x has the cdf

$$F(x) = \frac{1}{[1 + \exp(\alpha - \beta x^a)]}. \tag{46}$$

When a is greater than 1, the distribution is skewed to the left and, for a less than 1, to the right. As a approaches 0, the logarithmic transformation is appropriate. When a is a noninteger, the distribution is not defined for negative values of x. Since all the variables with which we are dealing are definitely positive, this restriction is not harmful.

The estimates to which the distribution is to be fitted are the mode x_2 and some lower and upper percentile points x_1 and x_3. Taking these as $2\frac{1}{2}$ percent and $97\frac{1}{2}$ percent, we have

$$0.025 = \frac{1}{[1 + \exp(\alpha - \beta x_1^a)]} \tag{47}$$

$$0.975 = \frac{1}{[1 + \exp(\alpha - \beta x_3^a)]}. \tag{48}$$

Rearranging leads to the relationships

$$\alpha - \beta x_1^a = \ln 39 \tag{49}$$

$$\alpha - \beta x_3^a = -\ln 39. \tag{50}$$

For a given value of a, these relationships can be solved for α and β. The problem is to find that value of a for which x_2 is the mode of the distribution. At this point, the slope of the pdf is 0.

Differentiating the cdf (46) twice with respect to x gives the slope of the pdf. Solving (49) and (50) leads to values of α and β. From these the slope at x_2 can be calculated. As a function of a, this is similar in shape, although opposite in sign, to the slope of the pdf as a function of x. Too great a value of a results in a positive slope, too small in a negative one. Whether a is greater or less than 1 depends on whether x_2 is to the right or left of the midrange. The method of finding a consists in calculating the slope at x_2 for a equal to 1, and then taking steps in a in the relevant direction until the sign of the slope changes. Linear interpolation gives an approximate value of a that is used as a starting value for Newton's method.

References

[1] N. R. Baker and W. H. Pound, "R & D project selection: "Where we stand," *IEEE Trans. Engineering Management*, vol. EM-11, pp. 124–134, December 1964.
[2] S. W. Hess, "A dynamic programming approach to R and D budgeting and project selection," *IRE Trans. Engineering Management*, vol. EM-9, pp. 170–179, December 1962.

Reprinted by permission from IEEE TRANSACTIONS ON *ENGINEERING MANAGEMENT*
Vol. EM-1, Number 1, February 1969
Pp. 2-8

Biometrika (1969), 56, 1, *p.* 33
Printed in Great Britain

The use of residuals as a concomitant variable

BY ANTHONY C. ATKINSON
Imperial College

SUMMARY

Suppose that there is correlation between the yields of successive or adjacent experimental units. An estimator of treatment effects using the residuals of adjacent plots as a concomitant variable is investigated. It is shown to be very close to the maximum likelihood estimator when the errors form a first-order autoregressive series.

1. INTRODUCTION

In the analysis of experimental data it is quite commonly assumed that observations on successive or adjacent units have independent errors. In randomized experiments this assumption is justified by the randomization. But sometimes an analysis making explicit allowance for correlation between adjacent units will be valuable, either because higher precision can be obtained, or because non-randomized data are under analysis.

One procedure for allowing for correlation is to introduce a specific model for the error variability, for example, a first-order autoregressive process where the units are arranged in time or along a line. Williams (1952) suggested appropriate designs for this case and developed the method of analysis.

Another, apparently quite different, method (Papadakis, 1937; Bartlett, 1938) proceeds as follows. The yield of each unit is corrected for the mean effect over all units receiving the same treatment. The average of the corrected yields of adjacent units is then used as a concomitant variable in the analysis of covariance. No explicit probability model is assumed in forming the adjusted estimate of the treatment effects. The conditions under which the estimates of precision so obtained are meaningful seem never to have been defined.

The object of the present paper is to consider Papadakis's procedure in more detail and, in particular, to show its connexion with the analysis based on an autoregressive process. Variation is considered in only one dimension, although the method was originally proposed for spatial variation in two dimensions.

2. MAXIMUM LIKELIHOOD SOLUTION

There are p treatments. Let the ith plot receive treatment s, the effect of which is α_s. If the observations are denoted by $y_0, y_1, \ldots, y_N$ then the model of the process is of the form

$$y_i = u_i + \alpha_s, \tag{2·1}$$

where u_i is the error component. We assume that the u_i's have geometrically decreasing correlations and can therefore be represented by the first-order autoregressive series

$$u_i = \rho u_{i-1} + \epsilon_i, \tag{2·2}$$

where $|\rho| < 1$ and the ϵ_i are normally and independently distributed with zero mean and

variance σ^2. Assuming that the process is stationary, we have that

$$\text{var}(y_i) = \sigma^2/(1-\rho^2) = \sigma_y^2. \tag{2·3}$$

Maximum likelihood estimates of the α_s are given by Williams, whose nomenclature is used in the present paper. In this section we give some of his results which will be used later.

The maximum likelihood equations obtained by Williams do not have an explicit solution. In order to proceed it is necessary to make some stipulations about the design.

Attention is confined to experiments consisting of m blocks each containing the p treatments once. This structure is appealing in that it guards against confounding the effects of the treatments with any long-range trends. The blocks have no significance either physically or for the analysis. They are solely an algebraic convenience in establishing the design.

Even within this class there are a large number of possible designs. Williams defines as a type II design one in which (*a*) no treatment occurs next to itself, and (*b*) each treatment occurs equally often next to every other treatment.

Such designs simplify the analysis. They also have the property that the variance of the estimate of treatment differences obtained by averaging over the relevant treatments does not vary too greatly over different pairs of treatments. An example of such a design for $p = 9$ and $m = 4$ is

(123456789) (246813579) (369471582) (591483726).

The parentheses divide the blocks. In order for the adjacency property to be satisfied, it is necessary to think of the design as circular, so that, in the present example, treatments 1 and 6 do occur together.

Given the structure of a type II design, Williams's maximum likelihood solutions simplify to some extent. Although there is not an explicit solution for the treatment effects, the equations reduce to one in only one variable $\hat{\rho}$, the value of which can readily be found by iteration on a computer. The resultant maximum likelihood estimator of α_s is

$$a_s^W = \frac{(1+\hat{\rho}^2)\sum\limits_{[i]=s} y_i - \hat{\rho}\sum\limits_{[i\pm1]=s} y_i + \dfrac{2\hat{\rho}}{p-1}\sum\limits_{i=1}^{N} y_i}{m\left(1+\hat{\rho}^2+\dfrac{2\hat{\rho}}{p-1}\right)}. \tag{2·4}$$

The square bracket round a subscript is to be read as 'the treatment applied to the ith plot'. Thus

$$\sum_{[i\pm1]=s} y_i$$

denotes the sum of those y_i adjacent to a plot receiving treatment s. The third sum in the numerator of (2·4) is over all observations. This third term cancels in estimating treatment differences.

Williams finds the variance of this estimator by differentiating the likelihood equation twice and inverting the resultant matrix to give, amongst other results,

$$\text{var}(a_s^W - a_t^W) = \frac{2\sigma^2}{m}\frac{1}{1+\rho^2+\dfrac{2\rho}{p-1}}. \tag{2·5}$$

Having presented these results and having discussed briefly the choice of a design, we are now in a position to apply Papadakis's method to the first-order autoregressive scheme.

3. The Papadakis estimator in the first-order case

The corrected yield of the ith plot receiving treatment s is y'_i, where

$$y'_i = y_i - (1/m) \sum_{[i]=s} y_i. \tag{3·1}$$

The concomitant variable x_i is the average of the adjacent corrected yields, i.e.

$$x_i = \tfrac{1}{2}(y'_{i-1} + y'_{i+1}). \tag{3·2}$$

If we denote by a_s^P the Papadakis estimator of the effect of treatment s, then

$$a_s^P = \frac{1}{m}\left\{\sum_{[i]=s} y_i - \frac{\hat{\beta}}{2} \sum_{[i]=s} \left(y_{i\pm1} - \frac{1}{m} \sum_{[i\pm1]=s} y_i\right)\right\}, \tag{3·3}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{N} x_i y'_i\right) \Big/ \left(\sum_{i=1}^{N} x_i^2\right). \tag{3·4}$$

For a type II design each treatment except s occurs c times next to treatment s, where $c = 2m/(p-1)$. Therefore

$$a_s^P = \frac{1}{m}\left\{\left(1 - \frac{\hat{\beta}}{p-1}\right) \sum_{[i]=s} y_i - \frac{\hat{\beta}}{2} \sum_{[i\pm1]=s} y_i + \frac{\hat{\beta}}{p-1} \sum_{i=1}^{N} y_i\right\}. \tag{3·5}$$

The expectation of $\hat{\beta}$ is found by using a large sample approximation to x_i. Also we take the expectation of the ratio to be the ratio of the expectations. Thus

$$E(\hat{\beta}) = 2\rho/(1+\rho^2). \tag{3·6}$$

Substitution of this value in (3·5) gives

$$a_s^P = \frac{1}{m}\left[\left\{1 + \frac{2\rho}{(1+\rho^2)(p-1)}\right\} \sum_{[i]=s} y_i - \frac{\rho}{1+\rho^2} \sum_{[i\pm1]=s} y_i + \frac{2\rho}{(1+\rho^2)(p-1)} \sum_{i=1}^{N} y_i\right]. \tag{3·7}$$

Comparison of (3·7) with the maximum likelihood estimator given in (2·4) shows that the two are very similar. If $\hat{\rho}$ has its expected value the first term of a_s^P is the first-order polynomial approximation to the first term of a_s^W. For small ρ the other two terms are nearly identical. For ρ near one, the denominator of the terms in the Papadakis estimator is too small by $1/(p-1)$. The differences in the two estimators will thus be greatest for a small number of treatments.

In estimating treatment differences the discrepancy between the estimators is reduced because the third term of (3·7) cancels to give

$$a_s^P - a_t^P = \frac{1}{m}\left\{\left(1 - \frac{\hat{\beta}}{p-1}\right)\left(\sum_{[i]=s} y_i - \sum_{[i]=t} y_i\right) - \frac{\hat{\beta}}{2}\left(\sum_{[i\pm1]=s} y_i - \sum_{[i\pm1]=t} y_i\right)\right\}. \tag{3·8}$$

We have shown that the two estimators are nearly identical. We now consider their variances.

4. VARIANCE OF THE PAPADAKIS ESTIMATOR

The estimator of treatment differences (3·8) contains $\hat{\beta}$ multiplied by a combination of the observations which is pure error. To find the variance of this product we expand in a Taylor's series and consider the variance of the expansion. Since the expectation of the error term is zero, we obtain only one term, the product of the variance of the error and the square of the expectation of $\hat{\beta}$. We thus proceed treating $\hat{\beta}$ as having its expected value.

Let

$$z_i = y_i - \tfrac{1}{2}\beta(y_{i+1}+y_{i-1}). \tag{4·1}$$

Then

$$\operatorname{var}(z_i) = \frac{\sigma^2}{1+\rho^2} \tag{4·2}$$

and

$$\operatorname{cov}(z_i, z_j) = \begin{cases} -\tfrac{1}{2}\beta \operatorname{var}(z_i) & (i = j \pm 1), \\ 0 & \text{otherwise.} \end{cases} \tag{4·3}$$

Also

$$\operatorname{cov}(z_i, y_j) = \begin{cases} \operatorname{var}(z_i) & (i = j), \\ 0 & \text{otherwise.} \end{cases} \tag{4·4}$$

Rewriting (3·8) we have

$$a_s^P - a_t^P = \frac{1}{m}\left\{ \sum_{[i]=s} z_i - \sum_{[i]=t} z_i - \frac{\beta}{p-1}\left(\sum_{[i]=s} y_i - \sum_{[i]=t} y_i \right) \right\}, \tag{4.5}$$

whence

$$\operatorname{var}(a_s^P - a_t^P) = \frac{2\sigma^2}{m(1+\rho^2)}\left(1 - \frac{\beta}{p-1}\right) + \frac{1}{m^2}\left(\frac{\beta}{p-1}\right)^2 E\left(\sum_{[i]=s} y_i - \sum_{[i]=t} y_i \right)^2. \tag{4·6}$$

The variance of the maximum likelihood estimator is given in (2·5). Comparison of these two variances shows that the first terms are related in the same way as the first terms of the estimators. The second term of (4·6) is a small multiplier times the variance of the crude estimator obtained by averaging over the relevant observations. The variances of the two esitmators are thus nearly equal, that of the Papadakis estimator being slightly the larger.

These results were checked by simulation, good agreement with theory being obtained. For the 9-treatment design of § 2 with a value for ρ of 0·8 the variance of the Papadakis estimate was 4% greater than that of the maximum likelihood estimate. Both variances were less than half the variance of the crude estimate.

5. THE RESIDUAL SUM OF SQUARES

An estimate of the error in the estimates obtained by Papadakis's method is based on the residual sum of squares after correction for the covariance. In this section we investigate the expectation of this quantity and its relationship with the calculated variance of the treatment differences. Only if the two are nearly the same will significance tests using the residual sum of squares as an estimate of error be meaningful.

We denote by R the residual sum of squares where

$$R = \sum_{s=1}^{p} \sum_{[i]=s} \left\{ z_i - \frac{1}{m}\sum_{[i]=s} z_i + \frac{\beta}{m}\left(\frac{1}{2}\sum_{[j]=[i\pm1]} y_j - \frac{1}{p-1}\sum_{i\neq s} y_i \right) \right\}^2. \tag{5·1}$$

In this

$$\sum_{[j]=[i\pm1]} y_j \quad \text{and} \quad \sum_{i\neq s} y_i$$

denote respectively sums over plots receiving the same treatment as $i+1$ and $i-1$, and a sum over observations not receiving treatment s.

The terms in round brackets in (5·1) contain no plots which have received treatment s. From (4·4) the expectation of R is therefore the sum of the expectations of the square of the terms in z_i and of the terms in round brackets. This second square involves treatments received by plots next but one to each other.

Williams defines as a type III design one in which every treatment occurs not only adjacent to but also next but one to every other treatment c times. He gives examples for up to five treatments. There is no reason why such designs should not exist for larger numbers of factors, although the one given here for 9 treatments does not quite meet this requirement.

If we assume the design does meet this requirement, we obtain

$$E(R) = \frac{p(m-1)\sigma^2}{1+\rho^2} + \frac{p(p-3)\beta^2}{2m(p-1)^2} E\left\{\sum_{s=1}^{p} \left(\sum_{[i]=s} y_i\right)^2 - \frac{1}{p}\left(\sum_{i=1}^{N} y_i\right)^2\right\}. \tag{5.2}$$

The expression inside the expectation operator is the sum of squares of the treatment means. Its value depends not only on the parameters of the system but also on the particular design.

For an analysis without a concomitant variable when the observations are independent

$$E(R) = p(m-1)\sigma^2 \tag{5·3}$$

and

$$\text{var}\,(a_s^c - a_l^c) = 2\sigma^2/m, \tag{5·4}$$

where a_s^c is the crude estimate obtained by averaging.

Thus V, the estimated variance of the treatment differences, has expectation

$$E(V) = \frac{2\sigma^2}{m(1+\rho^2)} + \frac{p-3}{m-1}\left\{\frac{\beta}{m(p-1)}\right\}^2 E\left\{\sum_{s=1}^{p} \left(\sum_{[i]=s} y_i\right)^2 - \frac{1}{p}\left(\sum_{i=1}^{N} y_i\right)^2\right\}, \tag{5·5}$$

whereas the value of the variance of treatment differences, averaged over all pairs of treatments, is

$$\text{ave}\{\text{var}\,(a_s^P - a_l^P)\} = \frac{2\sigma^2}{m(1+\rho^2)}\left(1 - \frac{\beta}{p-1}\right) + \frac{2}{p-1}\left\{\frac{\beta}{m(p-1)}\right\}^2 E\left\{\sum_{s=1}^{p} \left(\sum_{[i]=s} y_i\right)^2 - \frac{1}{p}\left(\sum_{i=1}^{N} y_i\right)^2\right\}. \tag{5·6}$$

The first term of the estimated variance is greater by a small amount than the corresponding term of the theoretical variance. The second term is also greater if, approximately, p^2 is greater than $2m$. Even for small p (for example, 4), the number of replicates when the inequality is not satisfied will be so large that the effect of the second term in each expression is negligible. Thus any confidence statement about the treatment differences using the residual sum of squares as an estimate of error will always be conservative.

The amount by which the estimated variance is an overestimate is shown by the results of 500 simulations given in Table 1. The 9-parameter model was used with ρ equal to 0·8.

Table 1. *Comparison of the variance of a treatment contrast with the variance estimated from the residual sum of squares*

	Variance of estimate	Estimated variance
Theoretical*	0·2858	0·3818
Simulation	0·3000	0·3689

* Calculated from (5·6) and (5·5).

The estimated variance from the residual sum of squares is about 20 % too high. No way of correcting for this suggests itself. The estimated variance provided by standard covariance analysis is even greater than that estimated from (5·5).

6. THE NULL CASE

When there is correlation between adjacent units, the adjusted estimate of the earlier sections will have smaller variance than the crude unadjusted estimate. If, however, there is no real correlation present, the unadjusted estimates will be better. We therefore need to consider the loss of precision arising from the inappropriate use of the adjusted estimates.

Let $\rho = 0$ and let adjustments be made using some value β_0. Then from (3·8) we require the expectation of

$$\frac{1}{m^2}\left\{\left(\sum_{[i]=s} y_i - \sum_{[i]=t} y_i\right)^2\left(1-\frac{\beta_0}{p-1}\right)^2 + \frac{\beta_0^2}{4}\left(\sum_{[i\pm1]=s} y_i - \sum_{[i\pm1]=t} y_i\right)^2 - \beta_0\left(1-\frac{\beta_0}{p-1}\right)\left(\sum_{[i]=s} y_i - \sum_{[i]=t} y_i\right)\left(\sum_{[i\pm1]=s} y_i - \sum_{[i\pm1]=t} y_i\right)\right\}. \tag{6·1}$$

Since the y_i are independent we obtain

$$\operatorname{var}(a_s^P - a_t^P) = \frac{2\sigma^2}{m}\left\{1+\left(\frac{\beta_0}{p-1}\right)^2\frac{p(p-3)}{2}\right\}. \tag{6.2}$$

We now calculate the variance of the estimate of β. From (3·4) we have, approximately, that

$$\operatorname{var}(\hat{\beta}) \simeq \frac{4E\left(\sum_{i=1}^{N} y_i' y_{i+1}'\right)^2}{E\left(\sum_{i=1}^{N} y_i'^2 + \sum_{i=1}^{N} y_i' y_{i+2}'\right)}. \tag{6·3}$$

For a type III design the second term in the denominator is zero. The numerator contains the fourth moments of normally distributed random variables with zero mean, expressions for which are given by Parzen (1962, p. 93). Simplification yields

$$\operatorname{var}(\hat{\beta}) \simeq \frac{4}{N}\left(\frac{m}{m-1}\right)^2\left\{1-\frac{2(p-2)}{m(p-1)}\right\}, \tag{6·4}$$

whence

$$\operatorname{var}(a_s^P - a_t^P) \simeq \frac{2\sigma^2}{m}\left[1+\frac{2}{N}\left(\frac{m}{m-1}\right)^2\frac{p(p-3)}{(p-1)^2}\left\{1-\frac{2(p-2)}{m(p-1)}\right\}\right]. \tag{6·5}$$

Five hundred simulations were performed with ρ equal to zero when $p = 4$ and $m = 6$. The results are given in Table 2.

Table 2. *Comparison of variance of adjusted and unadjusted estimates in the absence of correlation*

	Variance of estimate		Percentage increase
	Unadjusted	Adjusted	
Theoretical	0·3333	0·3472	4·1
Simulated	0·3402	0·3636	6·9

With fewer replications of each treatment, the increase in variance would be greater. But these results do strongly suggest that little increase in variance results from using the adjusted estimates in the absence of correlation between plot yields.

7. An assumed value of β

Instead of estimating β from the data we could assume *a priori* that it has the value β_0 and form the adjusted estimates using this value. The variance of this estimator is the expectation of (6·1).

The results of calculations of this quantity for treatments 5 and 8 of the 9-factor design are shown in Table 3. The quantity tabulated is the ratio of the variance using β_0 to make the adjustments to the variance using the β appropriate to the particular value of ρ.

Table 3. *Ratio of the variance of the Papadakis estimator using a preassigned value β_0 for the regression coefficient to the variance using the expected value*

β_0 \ ρ	0	0·2	0·4	0·6	0·8	0·9	1·0
0	1	1·065	1·258	1·619	2·252	2·666	2·996
0·3	1·038	1·002	1·071	1·235	1·526	1·713	1·356
0·5	1·106	1·008	1·010	1·079	1·211	1·294	1·182
0·7	1·207	1·053	1·001	1·001	1·029	1·048	1·060
0·8	1·270	1·090	1·016	0·992	0·989	0·989	0·989
0·9	1·342	1·136	1·043	1·003	0·983	0·974	0·969
0·99	1·414	1·186	1·080	1·030	1·006	0·998	0·995

For some values of β_0, the ratio of variances is slightly less than unity. This arises because the Papadakis estimator is not the minimum variance estimator.

As the value of β_0 increases, the ratio of variances for ρ equal to zero increases, whereas the ratio for ρ equals one decreases. For a value of β_0 of 0·6156 these ratios are equal with an increase in variance of just under 16 %. Thus the variance of the estimator is insensitive to the value used for the regression coefficient in calculating the adjusted estimates. Even if β_0 is far from the true value of β, the method will provide estimates with variance less than that of the crude estimate, provided some correlation is present.

8. Design and analysis

To compare the maximum likelihood and Papadakis estimators, we have assumed that the design has the property that each treatment occurs equally often next to every other treatment. We now compare the two estimators for other designs.

The maximum likelihood equation given by Williams for the estimate of a treatment effect is

$$-\hat{\sigma}^2\frac{\partial L}{\partial \alpha_s} = (1+\hat{\rho}^2)\sum_{[i]=s}(y_i - a_s^W) - \hat{\rho}\sum_{[i\pm 1]=s}(y_i - a_{[i]}^W) = 0. \qquad (8\cdot1)$$

For a type II design the last term simplifies as it does in the derivation of equation (3·5). For other designs a set of p simultaneous equations similar to (8·1) have to be solved for each value of $\hat{\rho}$, the correct value of which is found by iteration.

We replace $\hat{\rho}$ by its expected value. Then (8·1) may be rewritten as

$$a_s^W = \frac{1}{m}\left\{\sum_{[i]=s} y_i - \frac{\rho}{1+\rho^2}\sum_{[i\pm1]=s}(y_i - a_{[i]}^W)\right\}. \tag{8·2}$$

This set of equations could be solved either by matrix inversion or by iteration using as starting values the crude estimates obtained by averaging. Denote these estimates by a_s^c. Then the expression for the Papadakis estimator given in (3·3) may be rewritten as

$$a_s^P = \frac{1}{m}\left\{\sum_{[i]=s} y_i - \frac{\rho}{1+\rho^2}\sum_{[i\pm1]=s}(y_i - a_{[i]}^c)\right\}. \tag{8·3}$$

Thus the Papadakis estimator is a first approximation to the maximum likelihood estimator, regardless of the design. The two estimators will have very similar values.

The design furthest from Williams's type II design is the systematic one in which each treatment always occurs next to the same two treatments. An exact solution of the maximum likelihood equations is that the estimator is equal to the crude estimator. Similarly for the Papadakis estimator no correction by covariance is possible and equation (8·3) again reduces to the crude estimate.

For type II designs we have already shown that the variances of the two estimators are nearly the same. For a systematic design they are identical. Since the estimators are so close in form, it is reasonable to assume that the variances are similar for all designs. No analytical expressions have been obtained for these variances. To investigate the dependence of variance on design we concentrate on the Papadakis estimator which, unlike the maximum likelihood estimator, is readily calculable.

Intermediate between the type II and systematic designs is the randomized block design subject to the restriction that no treatment occurs next to itself. Five hundred simulations with $\rho = 0{\cdot}8$ for a 9-treatment design in four blocks gave an empirical variance for the treatment contrast of 0·3831. For a comparable type II design the Papadakis estimator had a variance of 0·3000 and the crude estimator a variance of 0·6168.

These results suggest that, of all randomized block designs, type II designs yield minimum variance estimates. They are therefore the preferred designs of this class. For other randomized block designs a significant increase in precision can be obtained by using either the maximum likelihood or Papadakis estimators. Appreciable reduction in computation results from employing the latter method of analysis.

9. CONCLUSIONS

The method suggested by Papadakis of using residuals of adjacent plots as concomitant variables has been applied to the analysis of results from a time series, when the underlying process is first-order autoregressive. We have shown that the properties of this estimator are very close to those of the maximum likelihood estimator. When the observations are in fact independent, using this method does not lead to a great increase in the variance of the estimator, whereas failure to take account of the structure of the errors results in considerable loss of precision. For other designs the Papadakis estimator, unlike the maximum likelihood estimator, is readily calculable and yields estimates of greater precision than those obtained by averaging over the relevant treatments. It is thus reasonable to suggest

that the method be considered whenever ordering of the observations in time or space has some meaning.

The method was originally suggested for the two-dimensional case of the results from field trials. Unfortunately, the extension of our results to this case is not obvious.

I am grateful to Professor D. R. Cox for his guidance of my work on this topic. This research was supported by an IBM Fellowship.

References

Bartlett, M. S. (1938). The approximate recovery of information from replicated field experiments with large blocks. *J. Agric. Sci.* **28**, 418–27.

Papadakis, J. S. (1937). Méthode statistique pour des expériences sur champ. *Bull. Inst. Amél. Plantes à Salonique*, no. 23.

Parzen, E. (1962). *Stochastic Processes*. San Francisco: Holden-Day.

Williams, R. M. (1952). Experimental designs for serially correlated observations. *Biometrika* **39**, 151–67.

[*Received May* 1968. *Revised October* 1968]

TECHNOMETRICS VOL. 11, NO. 3 AUGUST 1969

Constrained Maximisation and the Design of Experiments

ANTHONY C. ATKINSON

Imperial College, London

SUMMARY

M. J. Box has shown how transformations may be used to eliminate constraints in maximisation problems. This technique is described with reference to the design of experiments.

The February 1968 issue of Technometrics contained two papers in which the problem of maximising (or minimising) a function of several variables arose, where the values of the variables were subject to constraints. It is the purpose of this note to call attention to the use of transformations whereby certain forms of such constraints can be eliminated.

The method was suggested by M. J. Box [1]. His paper does not seem to be well known to statisticians. It is hoped that what follows will serve to publicise this useful technique. With one exception, all the transformations in the present paper are given by Box.

The paper by Hill, Hunter and Wichern [5] is an example of this situation in the context of the design of experiments. The constraints on the process variables time and temperature define a region of operability. The design problem is to find the maximum in this region of a function of the process variables called the design criterion.

For one or two variables the maximum may easily be found by searching over a grid. For more variables, however, this process rapidly becomes inefficient. In an unconstrained problem recourse would be made to one of the well established hill climbing programmes such as that of Powell [8]. Use of such a procedure when constraints are present may well result in the location of a maximum outside the region of operability. In such a situation the use of transformations may allow us to employ the power of the hill climbing technique whilst ensuring that the solution obtained does not violate any of the constraints.

As the simplest example, suppose we have a variable x_i which is required to be non negative. Then if we write $x_i = y_i^2$, an unconstrained search over all values of y_i is equivalent to a constrained search for x_i .

The next simplest problem arises in the design of experiments in regression situations. See, for example, Clark [2]. Here each variable is subject to the constraint $-1 \leq x_i \leq 1$, and the design criterion is to maximise the value of the

Received May 1968.

determinant $|X'X|$, where X is the design matrix. In this case we let

$$x_i = \sin y_i . \tag{1}$$

For all values of y_i, the constraint will be satisfied. The existence of multiple solutions in Y space does not cause any problems, provided the maximising routine does not take such large steps that it jumps from peak to peak of the function.

The commonest form of constraint is where x_i lies between two values, i.e. $X_{\min} \leq x_i \leq X_{\max}$. Here we write

$$x_i = X_{\min} + (X_{\max} - X_{\min}) \sin^2 y_i . \tag{2}$$

This kind of constraint was considered by Jennrich and Sampson [6] in their paper on non-linear least squares. One method of handling the boundary restriction is to find the overall minimum using the Gauss–Newton method. If this minimum lies outside the constraints, the foot of the perpendicular from the minimum to the constraint surface is taken as the least squares estimate of the parameters. In their figure 1 they show how this procedure may be inappropriate. This problem could be overcome by using the transformation given in equation (2). Although it could lead to increased computation in the calculation of the partial differentials of the sum of squares surface, it would enable the Gauss–Newton method to be used in constrained problems.

A different system of constraints arises in the problem of designing experiments for mixtures as discussed, amongst others, by Gorman and Hinman [4]. In this case we have $n + 1$ components. Let the amount of the ith component, for example the weight fraction, be x_i. Then the problem is to maximise some design criterion subject to the constraints

$$x_i \geq 0 \qquad i = 1, 2, \cdots, n + 1$$

$$\sum_{i=1}^{n+1} x_i = 1. \tag{3}$$

Here, since the weight fractions must sum to unity, there are only n independent variables. If we write

$$\begin{aligned} x_1 &= \sin^2 y_1 \\ x_i &= \sin^2 y_i \prod_{j=1}^{i-1} \cos^2 y_j \qquad i = 2, 3, \cdots, n \\ x_{n+1} &= \prod_{j=1}^{n} \cos^2 y_j \end{aligned} \tag{4}$$

an unconstrained search in n dimensional Y space will result in a search in that part of X space in which the constraints are satisfied.

Even when it is not possible to eliminate all the constraints, it is still advantageous to use transformations to eliminate as many as possible. In the design problem for mixtures considered by McLean and Anderson [7] the amounts of the components are not only subject to constraints of the form given in (3) but also have to lie between upper and lower limits. It has been suggested by Gorman [3] that their resultant extreme vertices designs can, under unfavourable

conditions, lead to the clustering of design points in the factor space. It does not seem possible to find any combination of transformations which will allow the investigation of the optimality of these designs in an unconstrained space. One approach would be to use the transformation given in equation (4) to ensure that all combinations considered do form a mixture, and then to penalise the design criterion if any of the constraints on the quantities of individual components are violated. The penalty function should not be such as to introduce any discontinuities into the surface. It might, for example, be proportional to the square of the amount by which the constraint is violated.

In this note we have drawn attention to the potential use of transformations coupled with a hill climbing technique to solve problems in the design of experiments. For a more detailed discussion of optimization methods and examples of the use of transformations the reader is referred to the paper by Box.

References

[1] Box, M. J., (1966). A comparison of several current optimization methods, and the use of transformations in constrained problems. The Computer Journal, *9*, 67–77.

[2] Clark, V., (1965). Choice of levels in polynomial regression with one or two variables. Technometrics, *7*, 325–333.

[3] Gorman, J. W., (1966). Discussion of 'Extreme vertices design of mixture experiments' by R. A. McLean and V. L. Anderson. Technometrics, *8*, 455–456.

[4] Gorman, J. W. and Hinman, J. E., (1962). Simplex lattice designs for multicomponent systems. Technometrics *4*, 463–487.

[5] 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.

[6] Jennrich, R. I. and Sampson, P. F., (1968). Application of stepwise regression to non-linear estimation. Technometrics, *10*, 63–72.

[7] McLean, R. A. and Anderson, V. L., (1966). Extreme vertices design of mixture experiments. Technometrics, *8*, 447–454.

[8] Powell, M. J. D., (1964). An efficient method of finding the minimum of a function of several variables without calculating derivatives. The Computer Journal, *7*, 155–162.