APPLICATION OF A MEASURE OF INFORMATION TO THE DESIGN AND COMPARISON OF REGRESSION EXPERIMENTS

By M. Stone

British Medical Research Council Applied Psychology Unit

1. Introduction and Summary. A normal regression experiment can be represented by

$$Y_i = \sum_{j=1}^{k} X_{ij}\,\theta_j + \eta_i \qquad (i = 1, \cdots, n) \tag{1.1}$$

where $\{\eta_i/i = 1, \cdots, n\}$ is a set of normally distributed random variables with zero means and non-singular dispersion matrix C, $\theta = (\theta_1, \cdots, \theta_k)$ is the parameter-vector of interest and $X = (X_{ij})$ is a known $n \times k$ matrix which will be called the allocation matrix. The rows of X will be called the allocation vectors. We denote the experiment by $\mathcal{E}(X, C)$. We assume that C is known; generally it will be a function of X, $C(X)$. The particular realisation of Y will be denoted y. The matrix $F = X'C^{-1}X$ is the Fisher-information-matrix of $\mathcal{E}(X, C)$.

When F is non-singular, one answer to the question "What information does y give about θ?" is to quote F^{-1}, the dispersion matrix of the maximum-likelihood-estimates of θ. A strong argument in favour of this is that F^{-1} is independent of both θ and y. The fact that it is independent of θ means that the answer is not "local"; the fact that it is independent of y leads to simplicity. This approach is taken by Box and Hunter [1] in their work on rotatable designs. However, we must accept the fact that many experimenters wish to have a one-dimensional answer to the question i.e. we must associate with $\mathcal{E}(X, C)$ a single number which we call the "information". For instance Elfving [5] has developed the use of trace F^{-1}. In this paper we adopt the measure of information introduced by Lindley [7]. In Section 2 we generalise Lindley's treatment of the regression situation to include the singular case, explain the uses of the measure and compare it with that of Elfving. Section 3 deals with the analogue of Elfving's main theorem. Theorems 4.1 and 4.2 of Section 4 provide links with the traditional variance approach. In Section 5 we derive the asymptotic form of the measure as the n of (1.1) increases and show that this form can be derived also from Neyman-Pearsonian theory. In Section 6 the influence of nuisance parameters is discussed and an analogue of a theorem of Chernoff [2] is established.

2. The information measure is defined in the Bayesian framework. Generally, if before experimentation we express our knowledge of θ by the prior distribution $p(\theta)$ and, after the experiment defined by the set of probability density functions $p(y/\theta)$, we obtain a posterior distribution $p(\theta/y)$, then the gain of information is

Received November 4, 1957; revised March 15, 1958.

defined as the functional

$$\Delta I = Ip(\theta/y) - Ip(\theta) \tag{2.1}$$

where $Ip(u) = \int p(u) \log p(u)\, du$.

LEMMA 2.1. *If, before* $\mathcal{E}(X, C)$, θ *is normally distributed with mean* μ *and non-singular dispersion matrix* A *then*

$$\Delta I = \tfrac{1}{2} \log | I + AF | . \tag{2.2}$$

PROOF.

$$Ip(\theta) = \int p(\theta) \log [(2\pi)^{-\frac{1}{2}k} | A |^{-\frac{1}{2}} \exp \{ -\tfrac{1}{2}(\theta - \mu)'A^{-1}(\theta - \mu)\}]\, d\theta$$

$$= -\tfrac{1}{2} \log [(2\pi)^k | A |] - \tfrac{1}{2} \int p(\theta)(\theta - \mu)'A^{-1}(\theta - \mu)\, d\theta$$

$$= -\tfrac{1}{2} \log [(2\pi)^k | A |] - \tfrac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} A^{ij}A_{ij}$$

$$= -\tfrac{1}{2} \log [(2\pi)^k | A |] - \tfrac{1}{2}k.$$

Also

$$p(\theta/y) = p(y/\theta)p(\theta)/p(y)$$

$$\propto \exp [-\tfrac{1}{2}(y - X\theta)'C^{-1}(y - X\theta) - \tfrac{1}{2}(\theta - \mu)'A^{-1}(\theta - \mu)].$$

Therefore $p(\theta/y)$ is normal with dispersion matrix

$$(X'C^{-1}X + A^{-1})^{-1} = (F + A^{-1})^{-1}.$$

Hence

$$Ip(\theta/y) = -\tfrac{1}{2} \log [(2\pi)^k / | F + A^{-1} |] - \tfrac{1}{2}k$$

and

$$\Delta I = \tfrac{1}{2} \log (| F + A^{-1} | | A |) = \tfrac{1}{2} \log | I + AF | .$$

Among the class of regression experiments for which it is reasonable to take a normal prior distribution for θ with dispersion matrix A, the expression for ΔI just derived proves useful in three ways:

Use 1. If we decide to experiment until the gain of information reaches a certain level then the fact that ΔI is independent of y allows us to state in advance whether a particular experiment will give us the required gain. Among experiments which do, we may choose the one which is most economical in some sense. This is the case of fixed-sample-size-experimentation.

Use 2. Two experiments can be compared in the following sense:

"Any result of $\mathcal{E}(X_1, C_1)$ will give $\Delta I_1 - \Delta I_2$ more information than any result of $\mathcal{E}(X_2, C_2)$". We may note that we are not obliged to use the average gain of information (as defined by Lindley [7]) to compare $\mathcal{E}(X_1, C_1)$ and $\mathcal{E}(X_2, C_2)$ although the result of doing so would be the same.

Use 3. If we have a choice of performing any experiment from a given class, we may choose the $\mathcal{E}(X, C)$ which maximises ΔI. ΔI possesses two advantages over the measure trace F^{-1}:

(i) If $\phi = M\theta$ is a non-singular linear transformation then ΔI is the same whether we consider information about θ or about ϕ. For

$$\begin{aligned} \tfrac{1}{2} \log \mid I + AF \mid &= \tfrac{1}{2} \log [\mid M \mid \mid I + AF \mid \mid M^{-1} \mid] \\ &= \tfrac{1}{2} \log \mid I + (MAM')(XM^{-1})'C^{-1}(XM^{-1}) \mid . \end{aligned} \tag{2.3}$$

But MAM' is the prior dispersion matrix for ϕ and, under the transformation, $\mathcal{E}(X, C) \rightarrow \mathcal{E}(XM^{-1}, C)$ so that (2.3) is the ΔI for ϕ.

(ii) ΔI can be used even when not all the θ_i are estimable i.e. when F is singular, whereas in this case trace F^{-1} becomes infinite. For $\mid I + AF \mid = \mid A \mid \mid A^{-1} + F \mid$, A^{-1} is positive-definite and, although F is singular, it remains positive-semi-definite. Hence $\mid A^{-1} + F \mid$ and therefore $\mid I + AF \mid$ are non-zero.

3. In connection with Use 3 we proceed to show that a theorem proved by Elfving using trace F^{-1} still holds if we adopt ΔI. The theorem is concerned with the following problem: "Given g possible allocation vectors $x(1), \cdots, x(g)$ which are linearly independent, we are to make n independent observations where n is large and each observation can be made at any of the allocation vectors. How are the observations to be allocated to maximise ΔI?". To answer this we need a lemma.

LEMMA 3.1. *If $F(n_1, \cdots, n_g)$ is the Fisher-information-matrix of the experiment consisting of n_i observations at $x(i)$ $(i = 1, \cdots, g)$ with the errors (η) uncorrelated and if we replace n_i in this matrix by np_i, $p_i \geqq 0$, $\sum p_i = 1$, to obtain the matrix $F^*[np_1, \cdots, np_g]$ then in general $\mid I + AF^* \mid$ is maximum when no more than $\frac{1}{2}k(k + 1)$ of the p_i's are non-zero.*

PROOF. We have assumed that the non-diagonal elements of $C(X)$ are zero. There is no further loss of generality in assuming that $C(X) = I$ for we can always write $\lambda_i\, x(i)$ for $x(i)$ so that this is so. We find that the (r, s)th element of $F(n_1, \cdots, n_g)$ is $\sum_{i=1}^{g} n_i\, x_r(i)x_s(i)$. Then

$$F^*_{rs} = \sum_{i=1}^{g} np_i\, x_r(i)x_s(i).$$

There are two possibilities:

(a) If there exists an i such that $\mid I + AF^* \mid$ is maximum when $p_i = 1$ then since $1 \leqq \frac{1}{2}k(k + 1)$ the lemma holds.

(b) If $\mid I + AF^* \mid$ is maximum at $p = p(m)$ when more than one $p_i(m)$ is non-zero, we proceed as follows.

$$\frac{\partial}{\partial p_i} \mid I + AF^* \mid = \mid A \mid \frac{\partial}{\partial p_i} \mid A^{-1} + F^* \mid$$

$$= \mid A \mid \sum \begin{vmatrix} nx_1^2(i) & nx_1(i)x_2(i) & \cdots \\ A^{21} + F^*_{21} & A^{22} + F^*_{22} & \cdots \\ \vdots & \vdots & \end{vmatrix}$$

for $i = 1, \cdots, g$ after row-by-row differentiation. Expanding each of these determinants about the row which has been differentiated we have

$$\frac{\partial}{\partial p_i} | I + AF^* | = n | A | x'(i)Q[p]x(i) \qquad (i = 1, \cdots, g)$$

where $Q[p] = \text{adj}\ (A^{-1} + F^*)$. Now $\sum p_i = 1$ so that, applying the method of undetermined multipliers to variations of the p_i's for i's for which $p_i(m) \neq 0$, we see that there exists a λ such that

$$\left\{ \frac{\partial}{\partial p_i} | I + AF^* | \right\}_{p(m)} = \lambda$$

for i such that $p_i(m) \neq 0$. Hence for such i

(3.1) $$x'(i)\, Q[p(m)]x(i) = \lambda' \quad \text{where} \quad \lambda' = \lambda/n \mid A \mid .$$

Now $x'Q[p(m)]x = \lambda'$ is the equation of a central quadric. In general not more than $\frac{1}{2}k(k + 1)$ of the given allocation vectors can lie on a central quadric and hence the lemma is established. To make the lemma fully rigorous, we would need to consider the possibility that more than $\frac{1}{2}k(k + 1)$ of the vectors do lie on a central quadric. We omit this consideration since it is rather tedious.

The lemma relates directly to the problem stated. From (2.2)

$$\Delta I = \tfrac{1}{2} \log |I + AF(n_1, \cdots, n_g)| = \tfrac{1}{2} \log |I + AF^*[np_1, \cdots, np_g]|$$

evaluated at $p_i = n_i/n$. So that if we do the experiment consisting of $[np_i(m)]$ observations at $x(i)$ $(i = 1, \cdots, g)$, which involves using at most $\frac{1}{2}k(k + 1)$ of $x(1), \cdots, x(g)$, for n large we will be making $n - f$ observations in all where $f \leqq g$. The allocation indicated will provide (asymptotically) the maximum ΔI. Thus:

THEOREM 3.1. *To achieve maximum ΔI for the above problem it is not necessary to use more than $\frac{1}{2}k(k + 1)$ of the given allocation vectors.*

For n not large the theorem is not necessarily true. Although it does not specify $p(m)$, it is nevertheless useful in providing a rule for rejecting some experiments for using too many allocation vectors. Generally the calculation of $p(m)$ is not feasible. However when $k = 2$ and the elements of $(AF^*[np(m)])^{-1}$ are small, we may proceed to obtain $p(m)$ approximately. By Theorem 3.1 only three allocation vectors need be used. Consider them as the only vectors given: $x(1), x(2), x(3)$. Approximately

$$Q[p] = \begin{pmatrix} \sum np_j x_2^2(j) & -\sum np_j x_1(j)x_2(j) \\ -\sum np_j x_2(j)x_1(j) & \sum np_j x_1^2(j) \end{pmatrix}$$

and hence

$$x'(i)Q[p]x(i)/n = x_1^2(i) \left[\sum p_j x_2^2(j)\right] - 2x_1(i)x_2(i) \left[\sum p_j x_1(j)x_2(j)\right] + x_2^2(i) \left[\sum p_j x_1^2(j)\right] = \sum_{j=1}^{3} p_j E_{ij} \quad (i = 1, 2, 3)$$

where $E_{ij} = [x_1(i)x_2(j) - x_2(i)x_1(j)]^2$. We note that $E_{ij} > 0$ when $i \neq j$ since the vectors $x(i)$ are supposed independent. Then, if $p_i(m) \neq 0$ $(i = 1, 2, 3)$, equation (3.1) gives

$$\begin{pmatrix} 0 & E_{12} & E_{13} \\ E_{21} & 0 & E_{23} \\ E_{31} & E_{32} & 0 \end{pmatrix} \begin{pmatrix} p_1(m) \\ p_2(m) \\ p_3(m) \end{pmatrix} = \begin{pmatrix} \lambda'' \\ \lambda'' \\ \lambda'' \end{pmatrix}$$

where $\lambda'' = \lambda'/n = \lambda/n^2|A|$. Using $\sum p_i(m) = 1$ we get

$$p_i(m) = E_{jk}(E_{ij} + E_{ik} - E_{jk}) \Big/ \sum_{i=1}^{3} E_{jk}(E_{ij} + E_{ik} - E_{jk}) \tag{3.2}$$

for $i = 1, 2, 3$; $(j, k) = (1, 2, 3) - (i)$. Also

$$\lambda'' = 2E_{23}E_{13}E_{12} \Big/ \sum_{i=1}^{3} E_{jk}(E_{ij} + E_{ik} - E_{jk}). \tag{3.3}$$

Hence for $p_i(m) > 0$ $(i = 1, 2, 3)$ either

$$E_{ij} + E_{ik} > E_{jk} \qquad (i = 1, 2, 3) \tag{3.4}$$

or

$$E_{ij} + E_{ik} < E_{jk} \qquad (i = 1, 2, 3). \tag{3.5}$$

The possibility of (3.5) can be readily excluded. We see that (3.3) and (3.5) imply $\lambda'' < 0$ but since Q is positive-definite $0 < x(i)'Qx(i) = \lambda' = n\lambda''$ or $\lambda'' > 0$, which is a contradiction. Therefore only (3.4) is consistent with $p_i(m) > 0$ $(i = 1, 2, 3)$. Equation (3.4) is not always satisfied by three given vectors. For example if $x(2)$ lies between $x(1)$ and $x(3)$ (in their two-dimensional representation) and

$$|x(2)| < \min(|x(1)|, |x(3)|), \qquad E_{ij} = |x(i) \wedge x(j)|^2 = 4A_{ij}^2$$

where A_{ij} = "area between $x(i)$ and $x(j)$". Clearly $A_{13} > A_{12} + A_{23}$; therefore $A_{13}^2 > A_{12}^2 + A_{23}^2$; therefore $E_{13} > E_{12} + E_{23}$.

However if (3.4) is satisfied the $p(m)$ given by (3.2) is that which approximately maximises ΔI. Also since Q is positive-definite $x'Qx = \lambda'$ is an ellipse, so that we need consider only triples of allocation vectors which lie on central ellipses.

We now evaluate the aymptotic maximum of ΔI:

(a) When (3.4) holds

$$\max \Delta I = \tfrac{1}{2} \log |I + AF^*[np(m)]| \cong \tfrac{1}{2} \log |A| + \tfrac{1}{2} \log |F^*[np(m)]|.$$

Now $|F^*[np]|$ is a homogeneous polynomial of degree two in p_i $(i = 1, 2, 3)$. Therefore

$$|F^*[np]| = \tfrac{1}{2} \sum p_i \frac{\partial}{\partial p_i} |F^*[np]|.$$

But, for the $p(m)$ of (3.2),

$$\left(\frac{\partial}{\partial p_i} \mid F^*[np] \mid\right)_{p(m)} = \frac{\lambda}{\mid A \mid}$$

so that

$$\mid F^*[np(m)] \mid = \tfrac{1}{2}\lambda/\mid A \mid = n^2 E_{23} E_{13} E_{12} / \sum E_{jk} (E_{ij} + E_{ik} - E_{jk}).$$

Hence

$$\text{(3.6)} \quad \max \Delta I \cong \log n + \tfrac{1}{2} \log \mid A \mid + \tfrac{1}{2} \log [E_{23} E_{13} E_{12} / \textstyle\sum E_{jk}(E_{ij} + E_{ik} - E_{jk})].$$

(b) When (3.4) is not satisfied, one of the vectors must have its associated p_i zero. Suppose $p_3(m) = 0$. Then the equations (3.1) lead to $p_2(m)E_{12} = \lambda'/n$ and $p_1(m)E_{12} = \lambda'/n$ so that $p_1(m) = p_2(m) = \frac{1}{2}$ and $\lambda'/n = \frac{1}{2} E_{12}$ or $\lambda = \frac{1}{2} n^2 \mid A \mid E_{12}$. Hence

$$\text{(3.7)} \quad \begin{aligned} \max \Delta I &\cong \tfrac{1}{2} \log |A| + \tfrac{1}{2} \log |F^*| = \tfrac{1}{2} \log |A| + \tfrac{1}{2} \log (\tfrac{1}{2} \lambda/|A|) \\ &= \log n + \tfrac{1}{2} \log |A| + \tfrac{1}{2} \log (E_{12}/4). \end{aligned}$$

In conclusion for the case $k = 2$ we state the experimental rule as follows: "Given g allocation vectors select those triples which obey (3.4) and calculate

$$E_{23} E_{13} E_{12} / \textstyle\sum E_{jk}(E_{ij} + E_{ik} - E_{jk}).$$

Also select those pairs which are not members of triples obeying (3.4) and calculate $E_{12}/4$. Make the observations at the pair or triple which gives the greatest number, with $n/2$ at each vector for a pair and $np_i(m)$ at $x(i)$ $(i = 1, 2, 3)$ for a triple where $p(m)$ is given by (3.2)."

Example 1. $k = 2$; $x(1) = (1, 1)$; $x(2) = (0, 1)$; $x(3) = (1, 0)$. Here $E_{12} = E_{13} = E_{23}$; therefore (3.4) is satisfied and, by (3.2), $p_i(m) = 1/3$ $(i = 1, 2, 3)$.

Example 2. k general. Suppose the given allocation vectors lie on the line $x = (1, x, \cdots, x^{k-1})$. This is the case of polynomial regression. $x'Qx = \lambda'$ is a polynomial of degree $2k - 2$ in x. Therefore in general at most $2k - 2$ of the vectors lie on a central quadric and therefore need be used.

In his discussion Elfving considers in detail the case $k = 2$. His solution, i.e. the "best allocation", is rather complicated when three vectors are used. For just two vectors he obtains

$$p_1(m) = |x(2)| / (|x(1)| + |x(2)|)$$

and

$$p_2(m) = |x(1)| / (|x(1)| + |x(2)|)$$

which clearly conflicts with $p_1(m) = p_2(m) = \frac{1}{2}$ using ΔI. The reason for the difference becomes clear in the case $x(1) = (c_1, 0)$, $x(2) = (0, c_2)$ for which

$$F^*[np] = \begin{pmatrix} np_1 c_1^2 & 0 \\ 0 & np_2 c_2^2 \end{pmatrix}.$$

The different answers arise because, effectively, we minimise the product of the variances of the maximum-likelihood-estimates of θ_1 and θ_2 while Elfving minimises their sum.

4. In this section we consider pairs of experiments $\mathcal{E}(X_1, C_1)$, $\mathcal{E}(X_2, C_2)$ and prove some theorems relating to the cases when $\mathcal{E}_1$ is always to be preferred to $\mathcal{E}_2$. Write $\Delta I_i(A) = \frac{1}{2} \log |I + AF_i|$ $(i = 1, 2)$.

THEOREM 4.1. *A necessary and sufficient condition that* $\Delta I_1(A) \geqq \Delta I_2(A)$ *for all positive-definite* A *is that* $F_1 - F_2$ *be positive-semi-definite.*

PROOF. Sufficiency. We use the fact that if L and M are positive-definite and $L - M$ is positive-semi-definite then $|L| \geqq |M|$. (Proved by diagonalising L and M.) Put $L = A^{-1} + F_1$ and $M = A^{-1} + F_2$; then if $F_1 - F_2$ is positive-semi-definite $|A^{-1} + F_1| \geqq |A^{-1} + F_2|$ which gives $\Delta I_1(A) \geqq \Delta I_2(A)$.

Necessity. $\Delta I_1(A) \geqq \Delta I_2(A)$ for all positive-definite A implies that $|A^{-1} + F_1| \geqq |A^{-1} + F_2|$ for all positive-definite A^{-1}. Now F_1 and F_2 are positive-semi-definite so that there exists a non-singular P such that $P'F_1P$ and $P'F_2P$ are diagonal.

$$P'F_iP = \begin{pmatrix} d_1(i) & & \\ & \ddots & \\ & & d_k(i) \end{pmatrix} \qquad (i = 1, 2) \tag{4.1}$$

Therefore

$$\left| P'A^{-1}P + \begin{pmatrix} d_1(1) & & \\ & \ddots & \\ & & d_k(1) \end{pmatrix} \right| \geqq \left| P'A^{-1}P + \begin{pmatrix} d_1(2) & & \\ & \ddots & \\ & & d_k(2) \end{pmatrix} \right|$$

where A^{-1}, and hence $P'A^{-1}P$, is arbitrary positive-definite. Taking $P'A^{-1}P$ diagonal with all diagonal elements large except the r'th we deduce $d_r(1) \geqq d_r(2)$. Therefore from equations (4.1)

$$P'(F_1 - F_2)P = \begin{pmatrix} d_1(1) - d_1(2) & & \\ & \ddots & \\ & & d_k(1) - d_k(2) \end{pmatrix}$$

is positive-semi-definite. Therefore $F_1 - F_2$ is positive-semi-definite.

We now give simpler proofs of theorems due to Ehrenfeld [4].

THEOREM 4.2. *If* $v_i(t)$ *is the variance of the maximum-likelihood estimate of* $t'\theta$ *from* $\mathcal{E}_i$ *and if* $F_1 - F_2$ *is positive-semi-definite then* $v_1(t) \leqq v_2(t)$ *for all* t *for which* $t'\theta$ *is estimable from both* $\mathcal{E}_1$ and $\mathcal{E}_2$.

PROOF. We know that $v_1 = \eta_1'F_1\eta_1$ and $v_2 = \eta_2'F_2\eta_2$ where η_1 and η_2 are any solutions of

$$F_1\eta_1 = t, \qquad F_2\eta_2 = t. \tag{4.2}$$

We have $\eta_1'(F_1 - F_2)\eta_1 \geqq 0$ or

$$v_1 - \eta_1'F_2\eta_1 \geqq 0 \tag{4.3}$$

while

$$(\eta_1 - \eta_2)'F_2(\eta_1 - \eta_2) = \eta_1'F_2\eta_1 - 2\eta_2'F_2\eta_1 + \eta_2'F_2\eta_2$$
$$\geqq 0.$$

From (4.2) $\eta_2'F_2\eta_1 = \eta_1'F_1\eta_1 = v_1$ therefore

$$\eta_1'F_2\eta_1 - 2v_1 + v_2 \geqq 0. \tag{4.4}$$

Adding (4.3) and (4.4) we obtain $v_1 \leqq v_2$.

THEOREM 4.3. *Given g allocation vectors $x(1), \cdots, x(g)$ and their convex hull $\mathfrak{C} = \{\sum \lambda_i x(i) / \sum \lambda_i = 1, \lambda_i \geqq 0\}$, if $F(x_1, \cdots, x_n)$ is the Fisher-information-matrix of the experiment consisting of n independent observations at $x_1, \cdots, x_n$ (with unit error variance) where $x_i \varepsilon \mathfrak{C}$ then we may take less than $n + g + 1$ observations at the vertices of $\mathfrak{C}$ so that their Fisher-information-matrix, F_V, is such that $F_V - F(x_1, \cdots, x_n)$ is positive-semi-definite.*

PROOF. Suppose $x_i = \sum_{j=1}^{g} \lambda_{ij} x(j)$. For each i we have

$$\sum_{j=1}^{g} \lambda_{ij} x(j)x'(j) - x_i x_i' = \sum_{j=1}^{g} \lambda_{ij} x(j)x'(j) - \sum_{j=1}^{g} \sum_{k=1}^{g} \lambda_{ij} \lambda_{ik}\, x(j)x'(k)$$

and for any t

$$t'\left[\sum_{j=1}^{g} \lambda_{ij} x(j)x'(j) - x_i x_i'\right] t = \sum_{j=1}^{g} \lambda_{ij} t'x(j)x'(j)t - t'x_i x_i' t$$
$$= \sum_{j=1}^{g} \lambda_{ij} \left[t'x(j) - \sum_{j=1}^{g} \lambda_{ij} t'x(j)\right]^2$$
$$\geqq 0.$$

Hence $\sum_{j=1}^{g} \lambda_{ij} x(j)x'(j) - x_i x_i'$ is positive-semi-definite for $i = 1, \cdots, n$. Therefore

$$\sum_{i=1}^{n} \sum_{j=1}^{g} \lambda_{ij} x(j)x'(j) - \sum_{i=1}^{n} x_i x_i'$$

is positive-semi-definite. Therefore

$$\sum_{j=1}^{g} \left(\left[\sum_{i=1}^{n} \lambda_{ij}\right] + 1\right) x(j)x'(j) - F(x_1, \cdots, x_n)$$

is positive-semi-definite where $[a]$ is the integral part of a. We can now identify $F_V = \sum_{j=1}^{g} ([\sum_{i=1}^{n} \lambda_{ij}] + 1)\, x(j)x'(j)$ for it is the Fisher-information-matrix of the experiment in which $[\sum_{i=1}^{n} \lambda_{ij}] + 1$ independent observations are made at $x(j)$ $(j = 1, \cdots, g)$ with error variances unity and also

$$\sum_{j=1}^{g} \left(\left[\sum_{i=1}^{n} \lambda_{ij}\right] + 1\right) \leqq n + g.$$

5. From Section 4 we see that the only case in which the ordering of two experiments by the criterion $\Delta I(A)$ is the same for all A is when $F_1 - F_2$ is positive-semi-definite. Clearly since $F_1 - F_2$ may be neither positive- nor negative-semi-definite, not all pairs of experiments can be compared in this clear-cut manner. However when A and F are non-singular so is AF and we may write $|I + AF| = |A|\,|F|\,|I + (AF)^{-1}|$ and

$$\Delta I(A) = \tfrac{1}{2} \log |A| + \tfrac{1}{2} \log |F| + \tfrac{1}{2} \log |I + (AF)^{-1}|.$$

If the elements of $(AF)^{-1}$ are small we have

$$\Delta I(A) \cong \tfrac{1}{2} \log |A| + \tfrac{1}{2} \log |F| \text{ and } \Delta I_1(A) - \Delta I_2(A) \cong \tfrac{1}{2} \log (|F_1| / |F_2|).$$

So we obtain the criterion $|F|$. The conditions under which it is valid are when, roughly speaking, either

(i) all the diagonal elements of A are large representing large prior uncertainity for all the parameters or

(ii) all the diagonal elements of F are large which is usually so if the n of (1.1) is large.

We now introduce a criterion based on the Neyman-Pearson theory of tests and show that a particular case of it leads to the $|F|$ criterion. Lehmann has given [6] a proof that for the experiment (1.1) the uniformly-most-powerful invariant test of the hypothesis $H_0: \theta = 0$ is provided by the usual $\mathfrak{F}$-test based on

$$\mathfrak{F} = \frac{\hat{\theta}' F \hat{\theta}/k}{(y - X\hat{\theta})C^{-1}(y - X\hat{\theta})/(n - k)}$$

where $\hat{\theta}$ are the maximum-likelihood estimates of θ. Taking as critical region $\mathfrak{F} > \mathfrak{F}_0$, denote by $P_{\text{II}}(\theta)$ the probability of error of the second kind under the alternative hypothesis $H: \theta$. A criterion for design can be stated as follows: "Take a probability density function for θ, $p(\theta)$, and choose the experiment to minimise $\int p(\theta) P_{\text{II}}(\theta)\, d\theta$." The choice of $p(\theta)$ is arbitrary but in a situation where we are initially very uncertain about θ it would be sensible to take $p(\theta)$ to be normal with mean 0 and diagonal dispersion matrix with variances all equal to E and consider what happens as $E \to \infty$. This we now do and state the theorem:

THEOREM 5.1. *If $p(\theta/E)$ is the probability density function just described then choosing the experiment to minimise $\int p(\theta/E) P_{\text{II}}(\theta)\, d\theta$ is equivalent to choosing it to maximise $|F|$.*

PROOF. Tang has shown [8] that $P_{\text{II}}(\theta)$ depends solely on the function $\lambda = \frac{1}{2}\theta' F \theta$. In fact

$$P_{II}(\theta) = \sum_{i=0}^{\infty} c_i \lambda^i e^{-\lambda}/i!$$

where the c_i are functions of i, $\mathfrak{F}_0$, k, n but are independent of X and C and also $c_i \to 0$ as $i \to \infty$ thus making the series uniformly convergent in $0 \leqq \lambda < \infty$.

$$\int p(\theta/E) P_{\text{II}}(\theta)\, d\theta = (2\pi E)^{-\frac{1}{2}k} \int \exp\left(-\tfrac{1}{2}\theta' E^{-1} \theta\right) P_{\text{II}}(\theta)\, d\theta$$

and

$$\lim_{E\to\infty} \int \exp\left(-\tfrac{1}{2}\theta' E^{-1}\theta\right) P_{\mathrm{II}}(\theta)\, d\theta = \int P_{\mathrm{II}}(\theta)\, d\theta.$$

Therefore

$$\int p(\theta/E) P_{\mathrm{II}}(\theta)\, d\theta \sim (2\pi E)^{-\frac{1}{2}k} \int P_{\mathrm{II}}(\theta)\, d\theta = (2\pi E)^{-\frac{1}{2}k} \int \left(\sum_{i=0}^{\infty} c_i \lambda^i e^{-\lambda}/i!\right) d\theta$$

$$= (2\pi E)^{-\frac{1}{2}k} \sum_{i=0}^{\infty} c_i \int \lambda^i e^{-\lambda}\, d\theta/i!$$

since the series is uniformly convergent in $0 \leqq \lambda < \infty$. Now

$$\int \lambda^i e^{-\lambda}\, d\theta = 2^{-i} \int (\theta' F\theta)^i \exp\left(-\tfrac{1}{2}\theta' F\theta\right) d\theta$$

$$= \frac{(2\pi)^{\frac{1}{2}k}}{2^i \mid F \mid^{\frac{1}{2}}} \int \frac{\mid F \mid^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}k}} \exp\left(-\tfrac{1}{2}\theta' F\theta\right)(\theta' F\theta)^i\, d\theta.$$

Now under the probability density function $(2\pi)^{-\frac{1}{2}k}|F|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\theta' F\theta\right)$, $\theta' F\theta$ is distributed as chi-square with k degrees of freedom. Hence

$$\int (2\pi)^{-\frac{1}{2}k} \mid F \mid^{\frac{1}{2}} \exp\left(-\tfrac{1}{2}\theta' F\theta\right)(\theta' F\theta)^i\, d\theta = g(i, k)$$

say where $g(i, k)$ is a function of i and k only. Therefore

$$\int p(\theta/E) P_{\mathrm{II}}(\theta)\, d\theta \sim E^{-\frac{1}{2}k}\left(\sum c_i g(i, k)/i!\, 2^i\right) \mid F \mid^{-\frac{1}{2}} \propto \mid F \mid^{-\frac{1}{2}}.$$

Hence minimising $\int p(\theta/E) P_{\mathrm{II}}(\theta)\, d\theta$ is equivalent to maximising $|F|$.

We give now an example of the use of the $|F|$ criterion, which has been treated by Tocher [9] from another viewpoint. If $C = I$ then $F = X'X$. Suppose the allocation vectors $x_i = (X_{i1}, \cdots, X_{ik})$ $(i = 1, \cdots, n)$ can be varied subject to the restriction $\sum_{i=1}^{n} X_{ij}^2 = a_j$. Then:

Theorem 5.2. *If $\sum_{i=1}^{n} X_{ij}^2 = a_j$ $(j = 1, \cdots, k)$ where $a_1, \cdots, a_k$ are positive constants then $|F|$ is maximum when $x_1, \cdots, x_n$ are chosen so that $F_{rs} = 0$, $r \neq s$, i.e. when the design is orthogonal.*

Proof.

$$F_{rs} = \sum_{i=1}^{n} X_{ir} X_{is}$$

Therefore, by a well-known property of positive-definite matrices

$$\mid F \mid \leqq \left(\sum X_{i1}^2\right)\left(\sum X_{i2}^2\right) \cdots \left(\sum X_{ik}^2\right) = a_1 a_2 \cdots a_k.$$

But when $F_{rs} = 0$, $r \neq s$, $|F| = a_1 a_2 \cdots a_k$. Therefore $|F|$ is maximum when the design is orthogonal.

6. We now consider the modifications in the $|F|$ criterion imposed by the presence of nuisance-parameters, ϕ, which enter linearly into the expressions for the expectations of our random variables, Y, just as the parameters of interest, θ, do.

Let there be q nuisance parameters and suppose that Y is now normal with mean $X\theta + Z\phi$ and dispersion matrix I. For simplicity take the case where

$$F_1 = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix}$$

is non-singular. Then θ and ϕ are estimable by the maximum-likelihood estimates $\hat{\theta}$ and $\hat{\phi}$. Write

$$\omega = \begin{pmatrix} \theta \\ \phi \end{pmatrix} \quad \text{and} \quad \hat{\omega} = \begin{pmatrix} \hat{\theta} \\ \hat{\phi} \end{pmatrix}.$$

Then

$$p(\hat{\omega}/\omega) = (2\pi)^{-\frac{1}{2}(k+q)} \, |F_1|^{\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\hat{\omega} - \omega)'F_1(\hat{\omega} - \omega)\right].$$

Suppose that the prior distribution for ω is

$$p(\omega) = (2\pi)^{-\frac{1}{2}(k+q)} \, |D|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\omega - \omega_0)D^{-1}(\omega - \omega_0)\right]$$

where

$$D = \begin{pmatrix} A & E \\ E' & B \end{pmatrix}$$

(A and B are the prior dispersion matrices for θ and ϕ respectively.) By Bayes' Theorem we find that $p(\omega/\hat{\omega})$ is normal with information matrix $(F_1 + D^{-1})$. To find the information about θ we must integrate out ϕ in (a) $p(\omega)$ and (b) $p(\omega/\hat{\omega})$ to obtain the marginal distributions of θ. We find:

(a) $p(\theta)$ is normal with dispersion matrix A

(b) $p(\theta/\hat{\omega})$ is normal with dispersion matrix L where L is the leading $k \times k$ diagonal sub-matrix of $(F_1 + D^{-1})^{-1}$. Then

$$Ip(\theta) = -\tfrac{1}{2} \log\left[(2\pi)^{\frac{1}{2}k} \, |A|\right] - \tfrac{1}{2}k$$

and

$$Ip(\theta/\hat{\omega}) = -\tfrac{1}{2} \log\left[(2\pi)^{\frac{1}{2}k} \, |L|\right] - \tfrac{1}{2}k$$

and

$$\Delta I = Ip(\theta/\hat{\omega}) - Ip(\theta) = -\tfrac{1}{2} \log |L| + \tfrac{1}{2} \log |A|. \tag{6.1}$$

If the elements of $(DF_1)^{-1}$ are small

$$(F_1 + D^{-1})^{-1} = F_1^{-1}(I + (DF_1)^{-1})^{-1} \cong F_1^{-1}.$$

Write

$$F_1^{-1} = \begin{pmatrix} \mathcal{A} & \gamma \\ \gamma' & \beta \end{pmatrix} \tag{6.2}$$

where $\mathcal{A}$ is the $k \times k$ dispersion matrix of $\hat{\theta}$ in $p(\hat{\omega}/\omega)$. Then $L \cong \mathcal{A}$ and

$$\Delta I = -\tfrac{1}{2} \log |\mathcal{A}| + \tfrac{1}{2} \log |A|. \tag{6.3}$$

The conditions under which the elements of $(DF_1)^{-1}$ are small are when, roughly speaking, either

(i) all the diagonal elements of D are large corresponding to large prior ignorance of all the parameters or

(ii) all the diagonal elements of F_1 are large corresponding to a "strong" experiment.

So, by (6.3), we see that under the conditions stated maximising ΔI is equivalent to minimising $|\mathcal{A}|$.

A. Wald [10] developed the use of $|\mathcal{A}|$ which he called the "generalised variance" but his justification of it was pragmatical rather than logical.

In most problems it is usually a simple matter to calculate F_1 from the allocation vectors. Then by Jacobi's Theorem we obtain $|\mathcal{A}| = |Z'Z| \,/\, |F_1|$.

EXAMPLE 1. A simple 2^2 factorial experiment without interaction with the base level (both factors absent) as the nuisance parameter. $k = 2$; $q = 1$.

$$\begin{array}{rl} & \quad\; n_0 \qquad\quad n_1 \qquad\quad n_2 \qquad\quad n_3 \\ X' = & \begin{pmatrix} 0 \cdots 0 & 1 \cdots 1 & 0 \cdots 0 & 1 \cdots 1 \\ 0 \cdots 0 & 0 \cdots 0 & 1 \cdots 1 & 1 \cdots 1 \end{pmatrix} \\ Z' = & (1 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots 1) \end{array}$$

Suppose $n = n_0 + n_1 + n_2 + n_3 = 4m$. Then

$$F_1 = \begin{pmatrix} n_1 + n_3 & n_3 & n_1 + n_3 \\ n_3 & n_2 + n_3 & n_2 + n_3 \\ n_1 + n_3 & n_2 + n_3 & n \end{pmatrix}$$

$$|\mathcal{A}| = n/|F_1| = n/(n_1n_2n_3 + n_0n_2n_3 + n_0n_1n_3 + n_0n_1n_2).$$

For minimum $|\mathcal{A}|$, $n_i = m$ $(i = 0, 1, 2, 3)$.

EXAMPLE 2. The addition of an interaction term θ_3 to Example 1.

$$\begin{array}{rl} & \quad\; n_0 \qquad\quad n_1 \qquad\quad n_2 \qquad\quad n_3 \\ X' = & \begin{pmatrix} 0 \cdots 0 & 1 \cdots 1 & 0 \cdots 0 & 1 \cdots 1 \\ 0 \cdots 0 & 0 \cdots 0 & 1 \cdots 1 & 1 \cdots 1 \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & 1 \cdots 1 \end{pmatrix} \\ Z' = & (1 \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots 1) \end{array}$$

Suppose $n = 4m$. Then

$$F_1 = \begin{pmatrix} n_1 + n_3 & n_3 & n_3 & n_1 + n_3 \\ n_3 & n_2 + n_3 & n_3 & n_2 + n_3 \\ n_3 & n_3 & n_3 & n_3 \\ n_1 + n_3 & n_2 + n_3 & n_3 & n \end{pmatrix}$$

$$|\mathfrak{A}| = n/n_0 n_1 n_2 n_3 .$$

For minimum $|\mathfrak{A}|$, $n_i = m$ $(i = 0, 1, 2, 3)$.

EXAMPLE 3. k treatments and a control [3]; $q = 1$.

$$X' = \begin{pmatrix} \overbrace{0 \cdots 0}^{n_0} & \overbrace{1 \cdots 1}^{n_1} & \overbrace{0 \cdots 0}^{n_2} & & \overbrace{0 \cdots 0}^{n_k} \\ 0 \cdots 0 & 0 \cdots 0 & 1 \cdots 1 & \cdots & 0 \cdots 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 \cdots 0 & 0 \cdots 0 & 0 \cdots 0 & & 1 \cdots 1 \end{pmatrix}$$

$$Z' = (1 \cdots\cdots\cdots\cdots\cdots 1)$$

Suppose n is divisible by $k + 1$. Then

$$F_1 = \begin{pmatrix} n_1 & & & n_1 \\ & \ddots & 0 & \vdots \\ 0 & & n_k & n_k \\ n_1 & \cdots & n_k & n \end{pmatrix}$$

$$| \mathfrak{A} | = n/|F_1| = n/n_0 n_1 \cdots n_k$$

which is minimised when $n_i = n/(k + 1)$ $(i = 0, 1, \cdots, k)$.

The calculation of $|\mathfrak{A}| = |Z'Z| \,/\, |F_1|$, though simple in the examples given, may be complicated. Then it may be possible to use a method we now elaborate.

DEFINITION. An experiment involving nuisance-parameters is "part-orthogonal" if $\gamma = 0$. [See (6.2).]

The customary definition of "orthogonal" requires that all non-diagonal elements of F_1^{-1} should be zero. However, if $\gamma = 0$, we could achieve this condition by two separate orthogonal transformations of θ and ϕ. From an informational point of view these are irrelevant.

By Wegner's Theorem $|F_1^{-1}| \leqq |\mathfrak{A}|\ |\beta|$ with equality when $\gamma = 0$. But by Jacobi's Theorem $|X'X| = |\beta| \,/\, |F_1^{-1}|$. Hence $|\mathfrak{A}| \geqq 1/|X'X|$ with equality when $\gamma = 0$. From this we derive the working principle: "Find the design which maximises $|X'X|$; if this is a part-orthogonal design (experiment) then it is the design which minimises $|\mathfrak{A}|$."

The allocation problem of Section 3 remains important in the presence of nuisance parameters. We prove a theorem which generalises Thorem 3.1. It is analogous to one due to Chernoff [2]. Now, the allocation vectors are the rows of the $n \times (k + q)$ matrix (XZ); denote them by u. Given g possible allocation vectors $u(i) = \begin{pmatrix} x(i) \\ z(i) \end{pmatrix}$ $(i = 1, \cdots, g)$ denote by $\Delta I(n_1, \cdots n_g)$ the information about θ in the experiment consisting of n_i observations at $u(i)$ $(i = 1, \cdots, g)$ with the errors (η) uncorrelated with unit variance.

THEOREM 6.1.[1] *When n is large, $\Delta I(n_1, \cdots, n_g)$ is in general maximised when no more than $\frac{1}{2}k(k + 1 + 2q)$ of the allocation vectors are used.* (It is not necessary that F_1 be non-singular.)

PROOF. By (6.1), $\Delta I(n_1, \cdots, n_g) = -\frac{1}{2}\log|L(n_1, \cdots, n_g)| + \frac{1}{2}\log|A|$ where $L(n_1, \cdots, n_g)$ is the leading $k \times k$ diagonal sub-matrix of $(F_1 + D^{-1})^{-1}$ with $F_1 = F_1(n_1, \cdots, n_g)$ and $[F_1]_{rs} = \sum_{i=1}^{g} n_i u_r(i) u_s(i)$. Suppose

$$D^{-1} = \begin{pmatrix} P & S \\ S' & R \end{pmatrix}$$

where P is $k \times k$. Then by Jacobi's Theorem

$$|L(n_1, \cdots, n_g)| = |Z'Z + R| \,/\, |F_1 + D^{-1}|.$$

Replace n_i by np_i, $p_i \geqq 0$, $\sum p_i = 1$:

$$L(n_1, \cdots, n_g) \rightarrow L[np]$$

$$F_1 \rightarrow F_1[np]$$

where

$$F_1\,[np] = \begin{pmatrix} F[np] & G[np] \\ G'[np] & H[np] \end{pmatrix}$$

say

$$Z'Z \rightarrow H[np]$$

$$|L[np]| = |H[np] + R| \,/\, |F_1[np] + D^{-1}|$$

We show that $|L[np]|$ is minimised when no more than $\frac{1}{2}k(k + 1 + 2q)$ of the p_i's are non-zero:

(a) If $|L[np]|$ is minimum at $p = p(m)$ where $p_i(m) = 1$ then, since $1 \leqq \frac{1}{2}k(k + 1 + 2q)$, the statement holds.

(b) If $|L[np]|$ is minimum at $p = p(m)$ when more than one $p_i(m)$ 0, $\neq$ we proceed as follows.

$$\frac{\partial}{\partial p_i}\log|\,L[np]\,| = \frac{1}{|\,H[np] + R\,|}\frac{\partial}{\partial p_i}|\,H[np] + R\,| - \frac{1}{|\,F_1\,[np] + D^{-1}\,|}\frac{\partial}{\partial p_i}|\,F_1\,[np] + D^{-1}|.$$

But

$$H[np]_{rs} = \sum_{i=1}^{g} np_i\, z_r(i) z_s(i)$$

$$F_1[np]_{rs} = \sum_{i=1}^{g} np_i\, u_r(i) u_s(i)$$

[1] The author is indebted to a referee for suggesting this theorem.

and by row-by-row differentiation of the determinants we find

$$\frac{\partial}{\partial p_i} \mid H[np] + R \mid = nz'(i) \text{ adj } (H[np] + R)z(i)$$

$$\frac{\partial}{\partial p_i} \mid F_1[np] + D^{-1} \mid = nu'(i) \text{ adj } (F_1[np] + D^{-1})u(i).$$

Therefore

$$\begin{aligned} \frac{\partial}{\partial p_i} \log \mid L[np] \mid &= n\, [z'(i)(H[np] + R)^{-1}z(i) - u'(i)(F_1[np] + D^{-1})^{-1}u(i)] \\ &= -nu'(i)Q[np]u(i) \end{aligned} \tag{6.4}$$

where

$$\begin{aligned} Q[np] &= (F_1\, [np] + D^{-1})^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & (H[np] + R)^{-1} \end{pmatrix} \\ &= (F_1\, [np] + D^{-1})^{-1} \begin{pmatrix} I & -(G[np] + S)(H[np] + R)^{-1} \\ 0 & 0 \end{pmatrix} \end{aligned}$$

Therefore the rank of $Q[np]$ is k. But $\sum_1^g p_i = 1$, therefore by the method of undetermined multipliers applied to variations of the p_i's for i's for which $p_i(m) \neq 0$, we see that for such i there exists a λ such that

$$\left(\frac{\partial}{\partial p_i} \log \mid L[np] \mid\right)_{p(m)} = \lambda$$

or, by (6.4), $u'(i)Q[np(m)]\, u(i) = -\lambda/n$. Hence those allocation vectors which are used to minimise $|L[np]|$ must lie on a central quadric of rank k. Such a quadric is determined by $(k + q) + (k + q - 1) + \cdots + (q + 1) = \frac{1}{2}k(k + 1 + 2q)$ constants implying that in general no more than $\frac{1}{2}k(k + 1 + 2q)$ of the vectors can have their associated $p_i(m)$'s non-zero. Now

$$\Delta I(n_1, \cdots, n_g) = -\tfrac{1}{2} (\log |L[np]|)_{p_i = n_i/n} + \tfrac{1}{2} \log |A|.$$

Therefore, when n is large, we can say that in general $\Delta I(n_1, \cdots, n_g)$ may be maximised when no more than $\frac{1}{2}k(k + 1 + 2q)$ of the n_i are non-zero and the theorem is proved.

REFERENCES

[1] G. E. P. Box and J. S. Hunter, "Multifactor experimental designs for exploring response surfaces," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 195–241.

[2] H. Chernoff, "Locally optimum designs for estimating parameters," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 586–602.

[3] C. W. Dunnett, "Multiple comparison procedures for comparing several treatments with control," *J. Amer. Stat. Assn.*, Vol. 50 (1955), pp. 1096–1121.

[4] S. Ehrenfeld, "Complete class theorems in experimental designs," *Third Berkeley Symposium*, Vol. 1, pp. 57–67.

[5] G. Elfving, "Optimum allocation in regression theory," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 255–262.

[6] E. L. LEHMANN, "Lecture notes," University of California.
[7] D. V. LINDLEY, "On a measure of the information provided by an experiment," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 986–1005.
[8] P. C. TANG, "The power function of the analysis of variance," *Stat. Res. Memoirs*, Vol. 2, pp. 126–149.
[9] K. D. TOCHER, "A note on the design problem," *Biometrika*, Vol. 39 (1952), p. 189.
[10] A. WALD, "On the efficient design of statistical investigations," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 134–140.