LOCALLY OPTIMAL DESIGNS FOR ESTIMATING PARAMETERS

By Herman Chernoff

Stanford University

1. Summary. It is desired to estimate s parameters $\theta_1, \theta_2, \ldots, \theta_s$. There is available a set of experiments which may be performed. The probability distribution of the data obtained from any of these experiments may depend on $\theta_1, \theta_2, \cdots, \theta_k$, $k \geqq s$. One is permitted to select a design consisting of n of these experiments to be performed independently. The repetition of experiments is permitted in the design. We shall show that, under mild conditions, locally optimal designs for large n may be approximated by selecting a certain set of $r \leqq k + (k - 1) + \cdots + (k - s + 1)$ of the experiments available and by repeating each of these r experiments in certain specified proportions. Examples are given illustrating how this result simplifies considerably the problem of obtaining optimal designs. The criterion of optimality that is employed is one that involves the use of Fisher's information matrix. For the case where it is desired to estimate one of the k parameters, this criterion corresponds to minimizing the variance of the asymptotic distribution of the maximum likelihood estimate of that parameter.

The result of this paper constitutes a generalization of a result of Elfving [1]. As in Elfving's paper, the results extend to the case where the cost depends on the experiment and the amount of money to be allocated on experimentation is determined instead of the sample size.

2. Introduction. Before formulating the problem precisely we shall consider a simple special example which will illustrate many of the points involved. Consider the regression problem

$$y = \gamma + \delta x + u \qquad -1 \leqq x \leqq 1 \tag{1}$$

where u is an unobserved disturbance which is normally distributed with mean 0 and variance 1. The disturbances of successive observations are distributed independently of each other. Suppose that we are permitted to select a set of n values of x between -1 and $+1$ and to observe the corresponding values of y. If our objective were to estimate δ, it is well known that the best procedure consists of using $x = +1$ for half of the observations and $x = -1$ for the other half.

In this problem we may regard the observation of a y corresponding to a given value of x as an experiment E_x. The class of available experiments is the set $\{E_x: -1 \leqq x \leqq 1\}$. The parameter in which we are interested is δ, but the distribution of the data depends on γ also. In this case γ is a nuisance parameter. The optimal design consists of using each of the two experiments E_1 and E_{-1}

Received 10/28/52.

half the time (if n is even). It should be noted that if the set of experiments available were decreased so that E_x is available only for $-1 < x < 1$, no optimal design could be found. This is essentially due to the fact that given any design, a better one can be obtained by spreading out the values of x even more (i.e., by taking values of x closer to the end points -1 and $+1$).

A peculiarity of this particular problem is that no matter how many times a particular experiment E_x is repeated, no reasonable estimate of δ can be determined. At least two distinct experiments are required. Another peculiarity of this problem is that the variance of $\hat{\delta}$, the maximum likelihood estimate of δ does not depend on the value of γ and δ. In general, this latter property will not hold and we shall be restricted to obtaining locally optimal designs, that is, designs which are optimal if the parameters are known to be close to certain specified values.

We may consider a variation of the above problem. Suppose that it is desired to estimate γ and that δ is the nuisance parameter. Then it is well known that an optimal design consists in repeating the experiment E_0, n times. An equally optimal design may also be obtained by using any set of x's so that $\bar{x} = 0$.

3. Information matrices and mixed experiments. The formulation of our problem will involve the concepts of information matrices [2] and of randomized or mixed experiments. For the sake of notational convenience and in order to clear up some technicalities that arise, we shall discuss these concepts before proceeding to the formulation.

R. A. Fisher defined *the information matrix $X(\theta)$ for an experiment involving the parameter $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ by*

$$X(\theta) = -\left\| E\left\{\frac{\partial^2 L}{\partial\theta_i\,\partial\theta_j}\right\}\right\| = \|\, x_{ij}(\theta)\,\| \qquad i, j = 1, 2, \cdots, k \tag{2}$$

where L is the logarithm of the likelihood function. It should be noted that $X(\theta)$ ordinarily depends on θ. It is easily seen and well known that

$$X(\theta) = \left\| E\left\{\frac{\partial L}{\partial\theta_i}\frac{\partial L}{\partial\theta_j}\right\}\right\| \tag{3}$$

and hence that $X(\theta)$ is a nonnegative definite symmetric matrix.

Another well known property of information matrices is that of additivity. That is, if $E_1, E_2, \cdots, E_n$ are experiments yielding information matrices $X_1(\theta)$, $X_2(\theta)$, $\cdots$, $X_n(\theta)$, the combined experiment or design which consists in carrying out each of these experiments independently yields the information matrix $X_1(\theta) + X_2(\theta) + \cdots + X_n(\theta)$

The experiment which consists in carrying out one of the available experiments, this one to be determined by a random device, is called a *randomized or mixed experiment.* Hence if $p_1, p_2, \cdots, p_n$ are positive numbers adding up to one, the experiment which consists in carrying out E_i with probability p_i is mixed. It is easily seen that this experiment has information matrix $p_1X_1(\theta) + p_2X_2(\theta) + \cdots + p_nX_n(\theta)$.

Let an experiment E with positive definite information matrix $X(\theta)$ be carried out m times and let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)$ be the resulting maximum likelihood estimate of $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$. Under mild conditions [3], the covariance matrix of the asymptotic (as $m \to \infty$) distribution of $\sqrt{m}(\hat{\theta} - \theta)$ is given by

$$X^{-1}(\theta) = \| x^{ij}(\theta) \| \qquad i, j = 1, 2, \cdots, k \tag{4}$$

at all points of continuity of $X(\theta)$. This property suggests the usefulness of information matrices in comparing designs.

Unfortunately, it is possible for an information matrix to be singular and hence to fail to have an inverse. To allow for this situation, we extend the notion of inverse to the class of nonnegative definite symmetric matrices. *Let X be nonnegative definite symmetric and let Y be any other symmetric matrix so that $X + \lambda Y$ is positive definite for positive λ small enough. Then, let*

$$X^{-1} = \| x^{ij} \| = \lim_{\lambda \to 0+} (X + \lambda Y)^{-1}. \tag{5}$$[1]

In Appendix A it will be shown that this new definition is consistent with the usual definition and is statistically meaningful. Also, if x^{ii} and x^{jj} are finite, then x^{ij} is finite and x^{ii}, x^{jj} and x^{ij} are independent of the particular Y selected. It should be noted that X^{-1} is a continuous function of X on the set of positive definite symmetric matrices but that elements of X^{-1} may fail to be continuous for X singular.

4. Formulation. In this section we shall formulate our problem and then indicate the reasons behind this formulation. Using the special example previously mentioned, we shall examine conditions which we shall impose to obtain the desired results.

There is a set $\{E\}$ of experiments available. The distribution of the data from one of these experiments depends on $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$. The information matrix $X(\theta)$ may be characterized by the elements on and above the main diagonal. These elements arranged in some order may be considered as components of a vector in $k(k + 1)/2$ dimensional space. This vector may be identified with the matrix. Since we are interested in locally optimal designs, that is, designs that are optimal when θ is known to be close to some given value, say $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_k^{(0)})$, we confine our attention to $X(\theta^{(0)})$.

Let R_1 be the set of vectors corresponding to the $X(\theta^{(0)})$ for the experiments of $\{E\}$. Let R be the convex hull of R_1, that is, *a typical element of R is the convex linear combination $p_1X_1 + p_2X_2 + \cdots + p_nX_n$ where $X_1, X_2, \cdots, X_n$ are elements of R_1 and $p_1, p_2, \cdots, p_n$ are positive numbers adding up to* 1.

From the previous section, it follows that R represents the set of information matrices of the class of mixed experiments.

---

[1] The author is indebted to Max A. Woodbury and the referee who independently pointed out a close relationship existing between this definition of inverse and the concept of the pseudo inverse of a matrix.

We shall be interested in showing that under certain conditions, an element $\tilde{X}$ *of* R *which minimizes*

$$v_s(X) = x^{11} + x^{22} + \cdots + x^{ss} \qquad s \leqq k, \tag{6}$$

can be represented as a convex linear combination of

$$r \leqq k + (k - 1) + \cdots + (k - s + 1) \tag{7}$$

elements $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_r$ *of* R_1.

It is evident that $\tilde{X}$ corresponds to a mixed experiment which is "optimal" in the sense that if $\hat{\theta}$ were based on n repetitions of this experiment, the sum of the variances in the asymptotic, (as $n \to \infty$), distribution of $\sqrt{n}(\hat{\theta}_1 - \theta_1)$, $\sqrt{n}(\hat{\theta}_2 - \theta_2), \cdots, \sqrt{n}(\hat{\theta}_s - \theta_s)$, would be a minimum.

Certain questions naturally arise concerning the usefulness of this criterion. First, it may be asked whether this criterion is relevant if one desires to confine oneself to pure experiments, that is, elements of $\{E\}$. Here we note that as $n \to \infty$, $\tilde{X}$ may be approximated by $(n_1\tilde{X}_1 + n_2\tilde{X}_2 + \cdots + n_r\tilde{X}_r)/n$ where $n_1, n_2, \cdots, n_r$ are positive integers adding up to n. The latter expression represents $1/n$ times the information matrix corresponding to the design where E_i is carried out n_i times. The answer to the last question would be *yes* if it were shown that $v_s(X)$ is continuous at $X = \tilde{X}$ on the convex set generated by $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_r$.

One may also ask why our criterion should involve information matrices. Such a criterion has a certain aesthetic appeal. Furthermore, we shall discuss in Appendix B how the main result yields a justification of this criterion.

Finally, one may seriously inquire whether a "good" design must minimize the sum of the asymptotic variances. In fact, we shall see in Appendix C that very often when one is interested in s parameters, a sound criterion for a "good" design involves minimizing $tr(AV)$ where A is a nonnegative definite symmetric matrix of rank less than or equal to s and V is the covariance matrix of the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$. By a linear transformation of θ this criterion may be transformed to that of minimizing the sum of no more than s asymptotic variances.

Since certain conditions must be imposed to obtain our desired result, we shall explain these conditions by referring to the example considered in Section 2. In that example the experiment E_x yields a likelihood function with logarithm given by

$$L = -\tfrac{1}{2} \log 2\pi - \tfrac{1}{2}(y - \gamma - \delta x)^2.$$

Let $\theta_1 = \delta$ and $\theta_2 = \gamma$. The corresponding information matrix is given by

$$X_x = \left\| \begin{matrix} x^2 & x \\ x & 1 \end{matrix} \right\|.$$

For this example, R_1 is the set of all points in three dimensional space whose coordinates are $(x^2, x, 1)$, $-1 \leqq x \leqq 1$. This set represents a segment of a parabola lying in a plane of three dimensional space. The convex set R generated by R_1 is the set bounded by R_1 and the line segment connecting the end points of R_1. The optimal design consisting in using X_1 and X_{-1}, each half the time, corresponds to the mid-point of the above-mentioned line segment, that is, the point (1, 0, 1).

We mentioned previously that if x is restricted to $-1 < x < 1$, no optimal nth order design exists. Note that in this case R has been changed by deleting the boundary line segment on which the optimizing point (1, 0, 1) lies. Although we can get arbitrarily close to this point when $-1 < x < 1$, we cannot reach it. In general, to prevent this minor difficulty we shall impose the condition that R be closed. Then R will contain all of its boundary points.

A second condition that we shall impose is that R be bounded. That is, no element of X_x can be made arbitrarily large by selecting E_x properly. This condition is satisfied in our example, for there no element can exceed 1 in absolute value. If, however, the example were modified to permit all real values of x, the element of the first row and first column of X_x would be unbounded. Note in this modified example, that if the parameter γ were known, δ could be estimated with arbitrarily small variance from one experiment by taking x large enough. This interpretation of the effect of unbounded R applies to the general case, too. If some element of X is unbounded, the fact that X is non-negative definite implies that some element of the main diagonal of X is unbounded. If the ith element of the main diagonal of X is unbounded, θ_i can be estimated with arbitrarily small asymptotic variance if all the other parameters are known.

5. Main results. In this section we state our main results. The proofs will first be given for $s = 1$ and then extended to $s > 1$.

THEOREM 1. *If R is closed and bounded there is an element $\tilde{X}$ of R which minimizes $v_s(X) = x^{11} + x^{22} + \cdots + x^{ss}$ and which is a convex linear combination of $r \leqq k + (k - 1) + \cdots + (k - s + 1)$ elements $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_r$ of R_1. Furthermore $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_r$ may be chosen so that $v_s(X)$ is a continuous function at $X = \tilde{X}$ with respect to the topology of the convex set generated by $\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_r$.*

We treat the case $s = 1$ where we let

$$z(X) = v_1(X) = x^{11}. \tag{8}$$

In outline, our proof for $s = 1$ consists in obtaining an expression for

$$\delta(X, \Delta) = z(X + \Delta) - z(X)$$

which will be used to show the existence of an $X^{(0)} \varepsilon R$ which minimizes $z(X)$ and such that $X^{(0)}$ lies on a supporting hyperplane of R. It will also be evident that $z(X)$ is constant on a sub-hyperplane. The dimension of this sub-hyperplane leads to the existence of $\tilde{X}$ with the desired properties. The complexity of the details of the proof arise mainly from difficulties in the case that $X^{(0)}$ is singular since $z(X)$ is not continuous at singular X.

LEMMA 1. *If X and $X + \Delta$ are nonnegative definite symmetric matrices and $z(X) \neq \infty$, then*

$$\delta(X, \Delta) \equiv z(X + \Delta) - z(X) = -\epsilon(X, \Delta) + \eta(X, \Delta) \tag{9}$$

where

$$\epsilon(X, \Delta) = \lim_{\lambda \to 0+} \epsilon_\lambda(X, \Delta), \tag{10}$$

$$\epsilon_\lambda(X, \Delta) = [(X + \lambda I)^{-1}\Delta(X + \lambda I)^{-1}]_{11}, \tag{11}$$

$$\eta(X, \Delta) = \lim_{\lambda \to 0+} \eta_\lambda(X, \Delta), \tag{12}$$

$$\eta_\lambda(X, \Delta) = [(X + \lambda I)^{-1}\Delta(X + \lambda I + \Delta)^{-1}\Delta(X + \lambda I)^{-1}]_{11} \tag{13}$$

and $\epsilon(X, \Delta)$ is a linear function in Δ and $\eta(X, \Delta) \geqq 0$.

PROOF. Since the matrices $(X + \lambda I)$ and $(X + \Delta + \lambda I)$ are positive definite for $\lambda > 0$

$$\delta(X, \Delta) = \lim_{\lambda \to 0+} [z(X + \Delta + \lambda I) - z(X + \lambda I)], \tag{14}$$

$$(X + \lambda I + \Delta)^{-1} - (X + \lambda I)^{-1} = -(X + \lambda I)^{-1}\Delta(X + \lambda I)^{-1} + (X + \lambda I)^{-1}\Delta(X + \lambda I + \Delta)^{-1}\Delta(X + \lambda I)^{-1}. \tag{15}$$

Since $z(X) \neq \infty$ it follows (see Appendix A, property 1) that $\lim_{\lambda \to 0+} (X + \lambda I)^{1i}$ exists and is finite for each i. Let us denote this limit by $X^{1i} = X^{i1}$. Hence

$$\epsilon(X, \Delta) = \lim_{\lambda \to 0+} \epsilon_\lambda(X, \Delta) = \sum_{i,j=1}^{k} X^{1i}\Delta_{ij}X^{j1}. \tag{16}$$

It follows that as $\lambda \to o+$, $\eta_\lambda(X, \Delta)$ converges (possibly to $+\infty$). Since the matrix, of which $\eta_\lambda(X, \Delta)$ is the element of the first row and first column, is nonnegative definite it follows that $\eta(X, \Delta) \geqq 0$.

LEMMA 2. *If X and $X + \Delta$ are nonnegative definite symmetric matrices and $z(X) = \infty$, then $\lim_{\Delta \to 0} z(X + \Delta) = \infty$.* (We write $\Delta \to 0$ if each element of Δ approaches zero. Note that Δ converges to zero subject to the condition that $X + \Delta$ is nonnegative definite and symmetric.)

PROOF. If A and B are symmetric matrices we use the notation $A \prec B$ or $B \succ A$ if $p'Ap \leqq p'Bp$ for every vector p. If A and B are positive definite and $A \prec B$, it is easily seen that $B^{-1} \prec A^{-1}$ by diagonalizing B and A. Also,

$$z(X + \Delta) = \lim_{\lambda \to 0+} [(X + \Delta + \lambda I)^{-1}]_{11}.$$

Let d be the largest characteristic value of Δ. Then

$$(X + \Delta + \lambda I) \prec (X + (\lambda + d)I), \qquad (X + \Delta + \lambda I)^{-1} \succ (X + (\lambda + d)I)^{-1}$$

$$z(X + \Delta) \geqq (X + dI)^{11}.$$

As $\Delta \to 0$, $d \to 0$. Furthermore $z(X) = \infty$. Hence $\lim_{\Delta \to 0} (X + dI)^{11} = \infty$, and our lemma follows.

LEMMA 3. *If R is closed and bounded, $z(X)$ attains its minimum on R.*

PROOF. Let $w = \inf_{X \varepsilon R} z(X)$. Because R is bounded, $w > 0$. The case $w = \infty$ is trivial and hence we assume $0 < w < \infty$. Since R is closed and bounded there is a sequence $\{X^{(i)}\}$ such that $X^{(i)} \varepsilon R$, $z(X^{(i)}) \to w$ and $\{X^{(i)}\}$ has a limit point $X^{(0)} \varepsilon R$. Let $\Delta^{(i)} = X^{(i)} - X^{(0)}$. It suffices to show that $z(X^{(0)}) \leqq w$. By Lemma 2, $z(X^{(0)}) \neq \infty$. Hence

$$z(X^{(0)} + \Delta^{(i)}) - z(X^{(0)}) = -\epsilon(X^{(0)}, \Delta^{(i)}) + \eta(X^{(0)}, \Delta^{(i)}).$$

Since ϵ is linear in $\Delta^{(i)}$,

$$\lim_{i \to \infty} \epsilon(X^{(0)}, \Delta^{(i)}) = 0.$$

But $\eta(X^{(0)}, \Delta^{(i)}) \geqq 0$. Letting $i \to \infty$, we obtain $w - z(X^{(0)}) \geqq 0$.

Hereafter we shall assume that R is closed and bounded. Then let ρ be the lowest rank associated with those elements of R which minimize $z(X)$. Now we assume that $X^{(0)}$ minimizes $z(X)$ on R, $z(X^{(0)}) \neq \infty$ and $X^{(0)}$ has rank ρ. We shall now reduce the set under consideration from R to $R \cap H_1$ where H_1 is a hyperplane containing $X^{(0)}$ and H_1 has dimension $\rho(\rho + 1)/2$. In the event that $X^{(0)}$ is nonsingular, no reduction from R has been effected. We shall not consider the trivial case $X^{(0)} = 0$ for then $w = \infty$.

We construct H_1 as follows. Corresponding to $X^{(0)}$, there is an orthogonal matrix $P = \| p_{ij} \|$ such that

$$X^{(0)} = P'EP \tag{17}$$

where

$$E = \begin{pmatrix} E_I & 0 \\ 0 & 0 \end{pmatrix} \tag{18}$$

and E_I is a diagonal $\rho \times \rho$ matrix where all the elements on the main diagonal are positive. *We define H_1 as the set of X for which*

$$P(X - X^{(0)})P' = \begin{pmatrix} D_I & 0 \\ 0 & 0 \end{pmatrix} \tag{19}$$

where D_I is a symmetric $\rho \times \rho$ matrix. It is evident that H_1 is a $\rho(\rho + 1)/2$ dimensional hyperplane containing $X^{(0)}$. We note that the nonnull set $R \cap H_1$ is the convex hull of $R_1 \cap H_1$ and is also closed and bounded.

LEMMA 4. *If $X^{(1)} \varepsilon R \cap H_1$, $X^{(1)}$ has rank ρ, $z(X^{(1)}) \neq \infty$, and $X^{(1)} + \Delta \varepsilon H_1$, then $\eta(X^{(1)}, \nu\Delta)$ approaches zero at least quadratically as $\nu \to 0$ and $z(X)$ is continuous at $X = X^{(1)}$ (in the topology of H_1).*

PROOF. Since $X^{(1)} \varepsilon R \cap H_1$ and $X^{(1)}$ has rank ρ,

$$PX^{(1)}P' = \begin{pmatrix} F_I & 0 \\ 0 & 0 \end{pmatrix}$$

where F_I is a positive definite symmetric matrix. If $X^{(1)} + \Delta \varepsilon H_1$,

$$P\Delta P' = \begin{pmatrix} D_I & 0 \\ 0 & 0 \end{pmatrix}$$

where D_I is symmetric. Let p_I be the vector consisting of the first ρ elements of the first column of P. If F_I and $F_I + \nu D_I$ are nonnegative definite we have,

$$\eta_\lambda(X^{(1)}, \nu\Delta) = p_I'(F_I + \lambda I)^{-1}(\nu D_I)(F_I + \lambda I + \nu D_I)^{-1}(\nu D_I)(F_I + \lambda I)^{-1}p_I. \tag{20}$$

But F_I is positive definite and for ν small enough $F_I + \nu D_I$ is also positive definite. Therefore,

$$\eta(X^{(1)}, \nu\Delta) = p_I'F_I^{-1}(\nu D_I)(F_I + \nu D_I)^{-1}(\nu D_I)F_I^{-1}p_I, \tag{21}$$

and

$$\lim_{\nu \to 0} \eta(X^{(1)}, \nu\Delta)/\nu^2 = p_I'F_I^{-1}D_IF_I^{-1}D_IF_I^{-1}p_I < \infty.$$

Similarly one may obtain

$$\epsilon(X^{(1)}, \nu\Delta) = p_I'F_I^{-1}(\nu D_I)F_I^{-1}p_I. \tag{22}$$

The continuity of $z(X)$ at $X = X^{(1)}$ follows immediately from equations (21) and (22).

LEMMA 5. *There is a sub-hyperplane H_2 of H_1 which is a supporting hyperplane of $R \cap H_1$ at $X^{(0)}$. H_2 has dimension $\frac{1}{2}(\rho(\rho + 1)) - 1$.*

PROOF. Suppose $X = X^{(0)} + \Delta \varepsilon R \cap H_1$. By convexity

$$X^{(0)} + \nu\Delta \varepsilon R \cap H_1 \qquad \text{for } 0 \leqq \nu \leqq 1.$$

If $\epsilon(X^{(0)}, \Delta) > 0$, it follows from Lemma 4 and the linearity of ϵ that $z(X^{(0)} + \nu\Delta) - z(X^{(0)}) < 0$ for small enough positive ν. This contradicts the fact that $X^{(0)}$ minimizes $z(X)$ on R. Hence

$$\epsilon(X^{(0)}, X - X^{(0)}) \leqq 0 \qquad \text{for } X \varepsilon R \cap H_1.$$

The sub-hyperplane H_2 of H_1 defined by the restriction

$$\epsilon(X^{(0)}, X - X^{(0)}) = p_I'E_I^{-1}D_IE_I^{-1}p_I = 0 \tag{23}$$

is a supporting hyperplane of $R \cap H_1$ at $X^{(0)}$. The fact that equation (23) actually constitutes a restriction on X depends on the fact that $p_I \neq 0$, and this in turn is easily established from $z(X^{(0)}) \neq \infty$, which implies that the last $k - \rho$ elements of the first column of P are all zero.

LEMMA 6. *There is a sub-hyperplane H_3 of H_2 so that $z(X) = z(X^{(0)})$ for $X \varepsilon R \cap H_3$. The dimension of H_2 minus that of H_3 is no more than $\rho - 1$.*

PROOF. For $X \varepsilon H_2$, $\epsilon(X^{(0)}, X - X^{(0)}) = 0$. From equation (21) it follows that if $E_I + D_I$ is nonsingular, the restriction

$$p_I'E_I^{-1}D_I = 0 \tag{24}$$

implies $\eta(X^{(0)}, \Delta) = 0$ and hence $z(X^{(0)} + \Delta) = z(X^{(0)})$. This implication holds even if $E_I + D_I$ is singular and nonnegative definite. For then we may apply equation (20) with F_I replaced by E_I and $\nu = 1$. We note that subject to restriction (24) $p_I'(E_I + \lambda I)^{-1}D_I = O(\lambda)$. Furthermore $(E_I + \lambda I + D_I)^{-1} < (1/\lambda)I$ whence $\eta_\lambda(X^{(0)}, \Delta) = O(\lambda)$ and $z(X^{(0)} + \Delta) = z(X^{(0)})$.

Equation (24) constitutes a set of at most ρ linearly independent restrictions on $X = X^{(0)} + \Delta$. However, since the restriction $\epsilon(X^{(0)}, \Delta) = 0$ may be written $p_I'E_I^{-1}D_IE_I^{-1}p_I = 0$ it follows that on H_2, the restriction (24) constitutes a set of at most $\rho - 1$ linearly independent restrictions. Let H_3 be the sub-hyperplane of H_2 on which $p_IE_I^{-1}D_I = O$. Lemma 6 follows.

LEMMA 7. *There is an element $\tilde{X}$ of R which minimizes $z(X)$ and which is a convex combination of a set of $r \leqq \rho$ elements of $R_1 \cap H_2$.*

PROOF. The set $R \cap H_3$ is closed, convex and bounded. There exists at least one element $\tilde{X}$ of $R \cap H_3$ which is not a convex combination of any two distinct elements of $R \cap H_3$. By Lemma 6, $z(\tilde{X}) = z(X^{(0)})$, that is, $\tilde{X}$ minimizes $z(X)$ on R. The matrix $\tilde{X}$ is an element of H_2 which supports $R \cap H_1$. Hence $\tilde{X}$ is a convex combination of elements of $R_1 \cap H_2$. Let r be the least number of elements of $R_1 \cap H_2$ which are required to yield $\tilde{X}$ as a convex combination. Then $\tilde{X}$ is an interior point of $R \cap H_4$ where H_4 is an $r - 1$ dimensional sub-hyperplane of H_2. Since X was selected so that $\tilde{X}$ is not an interior point of any line segment of $R \cap H_3$, $H_4 \cap H_3$ must have dimension 0 and hence $r - 1 \leqq \rho - 1$.

LEMMA 8. *Theorem 1 is valid for $s = 1$.*

PROOF. Lemma shows the existence of $\tilde{X}$ and the continuity property is given by Lemma 4.

Now that Theorem 1 has been established for $s = 1$, we shall extend the proof for $s > 1$. In that case we change our notation slightly. We let

$$z(X) = v_s(X) = x^{11} + x^{22} + \cdots + x^{ss} \tag{25}$$

$$\delta(X, \Delta) = z(X + \Delta) - z(X) \tag{26}$$

$$\epsilon_\lambda(X, \Delta) = \epsilon_\lambda^{(1)}(X, \Delta) + \epsilon_\lambda^{(2)}(X, \Delta) + \cdots + \epsilon_\lambda^{(s)}(X, \Delta) \tag{27}$$

$$\epsilon(X, \Delta) = \lim_{\lambda \to 0+} \epsilon_\lambda(X, \Delta) \tag{28}$$

$$\eta_\lambda(X, \Delta) = \eta_\lambda^{(1)}(X, \Delta) + \eta_\lambda^{(2)}(X, \Delta) + \cdots + \eta_\lambda^{(s)}(X, \Delta) \tag{29}$$

$$\eta(X, \Delta) = \lim_{\lambda \to 0+} \eta_\lambda(X, \Delta) \tag{30}$$

where $\epsilon_\lambda^{(i)}(X, \Delta)$ and $\eta_\lambda^{(i)}(X, \Delta)$ are obtained from the ith diagonal terms of the matrices appearing in equations (11) and (13), respectively. Then Lemmas 1, 2, 3, and 4 may be established as in the case for $s + 1$. Equations (20), (21), and (22) are slightly modified. To illustrate, equation (20) becomes

$$\eta_\lambda(X^{(1)}, \nu\Delta) = \sum_{i=1}^{s} p_I^{(i)\prime}(F_I + \lambda I)^{-1}(\nu D_I)(F_I + \lambda I + \nu D_I)^{-1} \cdot (\nu D_I)(F_I + \lambda I)^{-1} p_I^{(i)} \tag{31}$$

where $p_I^{(i)}$ is the vector whose components are the first ρ elements of the ith column of P. It will be useful to note later that the condition $z(X^{(0)}) \neq \infty$ implies that the last $k - \rho$ elements of the first s columns of P are all zeros. In fact, $p_I^{(1)}, p_I^{(2)}, \cdots, p_I^{(s)}$ are then unit orthogonal vectors.

Lemma 5 follows as before with the restriction defining H_2 replaced by

$$\epsilon(X^{(0)}, X - X^{(0)}) = \sum_{i=1}^{s} p_I^{(i)\prime} E_I^{-1} D_I E_I^{-1} p_I^{(i)} = 0. \tag{32}$$

In Lemma 6, the wording must be modified so that the dimension of H_2 minus that of H_3 is no more than $\rho + (\rho - 1) + \cdots + (\rho - s + 1) - 1$. The change is due to the fact that restriction (24) is now replaced by

$$P_I' E_I^{-1} D_I = 0 \tag{33}$$

where P_I is the $(\rho \times s)$ matrix of rank s consisting of the first ρ rows and s columns of P. It is possible to rearrange the rows and columns of D_I (maintaining symmetry) so that equation (33) may be expressed by

$$(Q_{11} Q_{12}) \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} = 0 \tag{34}$$

where Q_{11} is nonsingular, Q_{11}, Q_{12}, D_{11}, $D_{12} = D_{21}'$ and D_{22} are of order $s \times s$, $s \times (\rho - s)$, $s \times s$, $s \times (\rho - s)$ and $(\rho - s) \times (\rho - s)$, respectively, and

$$\begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$$

is the result of rearranging the rows and columns of D_I. But then

$$D_{21} = -Q_{11}^{-1} Q_{12} D_{22}$$

$$D_{11} = -Q_{11}^{-1} Q_{12} D_{12} = -Q_{11}^{-1} Q_{12} D_{21}'.$$

Hence, after the restriction (33), D is determined by D_{22} and has only $(\rho - s)(\rho - s + 1)/2$ linearly independent elements. Hence, equation (33) imposes

$$\frac{\rho(\rho + 1)}{2} - \frac{(\rho - s)(\rho - s + 1)}{2} = \frac{s(2\rho - s + 1)}{2}$$
$$= \rho + (\rho - 1) + \cdots + (\rho - s + 1)$$

independent linear restrictions on the symmetric matrix D. But as before one of these restrictions is lost on H_2, for (33) implies $\epsilon(X^{(0)}, \Delta) = 0$. Lemma 7, with ρ replaced by $\rho + (\rho - 1) + \cdots + (\rho - s + 1)$ follows as before. Theorem 1 is once more an immediate consequence of Lemmas 4 and 7.

6. Remarks. In many cases, the cost of experimentation depends on the experiment. Then the usual design problem is to maximize information, given a certain amount of money to spend on experimentation. Our results of Section

5 are easily seen to apply in this case, too. Here we identify with each experiment a matrix

$$Y(\theta) = X(\theta)/c \tag{35}$$

where c is the cost of performing the experiment. The matrix $Y(\theta)$ represents information per unit cost. The matrix which we associate with the mixed experiment when E_j is carried out with probability p_j, $i = 1, 2, \cdots, n$, is

$$Y(\theta) = \frac{\sum_{i=1}^{n} p_i X_i(\theta)}{\sum_{i=1}^{n} p_i c_i} = \frac{\sum_{i=1}^{n} c_i p_i Y_i(\theta)}{\sum_{i=1}^{n} c_i p_i}. \tag{36}$$

It is evident that a reasonable criterion for a good mixed experiment is that $v_s[Y(\theta)]$ be minimized.

In [1], Elfving obtained our result (Theorem 1) for $s = 1$ and $s = k$ in the case of linear regression. Elfving also indicated an elegant geometrical method of obtaining the optimal design. The methods used by Elfving depend only on the assumption that for any nonrandomized experiment, $X(\theta)$ may be represented in the form

$$X(\theta) = \| x_{ij}(\theta) \| = \| x_i(\theta)x_j(\theta) \|. \tag{37}$$

Hence, these methods may be applied in many examples which are not regression problems.

7. Examples. In this section we shall discuss some examples in order to show how the results of Section 5 may be used to reduce considerably the amount of work required to obtain optimal designs of experiments. The results of Elfving [1] make it unnecessary for us to consider the important and interesting examples from linear regression theory.

EXAMPLE 1. Suppose that

$$p_d = e^{-\theta d} \qquad \theta > 0, d \geqq 0, \tag{38}$$

represents the probability that an insect will survive a dose of d units of a certain insecticide. It is desired to select n dose levels to try on n insects to estimate θ in an optimal fashion.

Here the information matrix corresponding to a particular d is given by

$$X_d = d^2e^{-\theta d}/(1 - e^{-\theta d}), \qquad d > 0 \tag{39}$$

and

$$X_d = 0 \qquad d = 0. \tag{40}$$

The conditions of Theorem 1 are satisfied and hence it follows that a *locally optimal design consists of repeating one dose level n times*. Maximizing X_d we find that this dose level satisfies

$$2e^{-\theta d} + \theta d = 2$$

$$d \approx \frac{1.6}{\theta}. \tag{40}$$

For this locally optimal dose level, the probability of survival p_d is very close to .2. An interesting by-product is that for the general design the maximum likelihood estimates are not too simple to obtain or study. For the optimal design the estimation problem is that of estimating the probability associated with a binomial distribution.[2]

EXAMPLE 2. Let A and B represent two characteristics that members of a population may or may not have, for example, the habit of smoking and heart disease. Let $\bar{A}$ and $\bar{B}$ represent the complementary characteristics. It is desired to estimate the degree of dependence of the two characteristics A and B. Five experiments may be performed. These correspond to examining individuals either: (i) at random; (ii) with characteristic A; (iii) with characteristic $\bar{A}$; (iv) with characteristic B; or (v) with characteristic $\bar{B}$. The parameters involved are p_A, $p_{\bar{A}} = 1 - p_A$, p_B, $p_{\bar{B}} = 1 - p_B$, and $\theta = p_{AB} - p_A p_B$ where p with a subscript indicates the proportion of the population with the characteristics of the subscript. There are three independent parameters.

In the case where p_A and p_B are known, it has been shown by Blackwell [4] that to test for independence an optimal design involves repeating one experiment n times. This experiment is the one which corresponds to the smallest of the four probabilities p_A, $p_{\bar{A}}$, p_B, and $p_{\bar{B}}$. Here, Theorem 1 may be applied to yield the same result if it is desired to estimate θ when θ is assumed to be close to zero.

Suppose, now, that our problem is modified so that p_B is only approximately known. Here, Theorem 1 applies with $k = 2$, $s = 1$, and tells us that we should use at most two of the experiments. Furthermore, since selecting an individual at random is equivalent to a mixture of two of the other experiments we may confine our attention only to pairs of the other four experiments. Let us now evaluate the information matrix X_A corresponding to examining an individual with characteristic A, this information matrix to be evaluated at $\theta = 0$. In this experiment the probability of observing a smoker is $p_B + \theta/p_A$. If the individual observed has characteristic B

$$L = \log\left(p_B + \frac{\theta}{p_A}\right), \quad \frac{\partial L}{\partial \theta} = \frac{1}{\left(p_B + \frac{\theta}{p_A}\right) p_A},$$

and

$$\frac{\partial L}{\partial p_B} = \frac{1}{p_B + \frac{\theta}{p_A}}.$$

[2] The author wishes to express his thanks to Fred Andrews for his assistance on this example.

If the individual observed has characteristic $\bar{B}$,

$$L = \log\left(1 - p_B - \frac{\theta}{p_A}\right), \qquad \frac{\partial L}{\partial \theta} = \frac{-1}{\left(1 - p_B - \frac{\theta}{p_A}\right) p_A}$$

and

$$\frac{\partial L}{\partial p_B} = \frac{-1}{\left(1 - p_B - \frac{\theta}{p_A}\right)}.$$

Hence

$$X_A = \frac{1}{p_B p_{\bar B}} \left\| \begin{matrix} \frac{1}{p_A^2} & \frac{1}{p_A} \\ \frac{1}{p_A} & 1 \end{matrix} \right\| = \frac{1}{p_A\, p_{\bar A}\, p_B\, p_{\bar B}} \left\| \begin{matrix} \frac{p_{\bar A}}{p_A} & p_{\bar A} \\ p_{\bar A} & p_{\bar A}\, p_A \end{matrix} \right\|. \tag{41}$$

Similarly

$$X_{\bar A} = \frac{1}{p_A\, p_{\bar A}\, p_B\, p_{\bar B}} \left\| \begin{matrix} \frac{p_A}{p_{\bar A}} & -p_A \\ -p_A & p_A\, p_{\bar A} \end{matrix} \right\|, \tag{42}$$

$$X_B = \frac{1}{p_A\, p_{\bar A}\, p_B\, p_{\bar B}} \left\| \begin{matrix} \frac{p_{\bar B}}{p_B} & 0 \\ 0 & 0 \end{matrix} \right\| \tag{43}$$

and

$$X_{\bar B} = \frac{1}{p_A\, p_{\bar A}\, p_B\, p_{\bar B}} \left\| \begin{matrix} \frac{p_B}{p_{\bar B}} & 0 \\ 0 & 0 \end{matrix} \right\|. \tag{44}$$

From the remarks of the previous section, it follows that Elfving's results [1] may be applied. The geometrical figure that is developed shows immediately that the optimal design consists of using either $\bar{B}$, or $\bar{B}$ or A and $\bar{A}$ each half the time, according as to which of the numbers

$$\sqrt{\frac{p_{\bar B}}{p_B}}, \quad \sqrt{\frac{p_B}{p_{\bar B}}}, \quad \frac{1}{2\sqrt{p_A\, p_{\bar A}}}$$

is greatest. This last result can also be obtained directly without computational difficulty.

8. Appendices.

Appendix A. Extension of the inverse to nonnegative definite symmetric matrices.

Here we extend the notion of the inverse of a matrix to nonnegative definite symmetric matrices and show how this extension has statistical significance. Suppose that X is nonnegative definite and symmetric. Let Y be a symmetric matrix so that $X + \lambda Y$ is positive definite for positive λ sufficiently small. *Then we define the inverse of X relative to Y by*

$$X_Y^{-1} = \lim_{\lambda \to 0+} (X + \lambda Y)^{-1}. \tag{45}$$

The usefulness of this definition arises mainly from the following property.

PROPERTY 1. *The diagonal elements of X_Y^{-1} are independent of Y. Furthermore, if the ith and jth diagonal elements of X_Y^{-1} are finite, the (i, j) element of X_Y^{-1} is finite and independent of Y. If the ith diagonal element of X_Y^{-1} is finite, all the elements of the ith row of X_Y^{-1} are finite.*

PROOF. Corresponding to X there is an orthogonal matrix P such that

$$P'XP = E = \begin{pmatrix} E_{11} & 0 \\ 0 & 0 \end{pmatrix} \tag{46}$$

where E_{11} is a diagonal $\rho \times \rho$ matrix whose diagonal elements $e_1, e_2, \cdots, e_\rho$ are positive. We define F by

$$P'YP = F = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}. \tag{47}$$

Then F_{22} is positive definite and

$$\begin{aligned} (X + \lambda Y)^{-1} &= P(E + \lambda F)^{-1}P' \\ &= P \begin{pmatrix} E_{11}^{-1} + O(\lambda) & -[E_{11}^{-1} + O(\lambda)]F_{12}F_{22}^{-1} \\ -F_{22}^{-1}F_{21}[E_{11}^{-1} + O(\lambda)] & \frac{1}{\lambda}[F_{22} - \lambda F_{21}[E_{11} + \lambda F_{11}]^{-1}F_{12}]^{-1} \end{pmatrix} P'. \end{aligned}$$

Let p_{i1} and p_{i2} represent the first ρ and the remaining $k - \rho$ elements of the ith column of P. Then

$$\begin{aligned} (X + \lambda Y)^{ij} = {} & p_{i1}'E_{11}^{-1}p_{j1} - p_{i2}'F_{22}^{-1}F_{21}E_{11}^{-1}p_{ji} - p_{i1}'E_{11}^{-1}F_{12}F_{22}^{-1}p_{j2} \\ & + \frac{1}{\lambda}p_{i2}'F_{22}^{-1}p_{j2} + p_{i2}'F_{22}^{-1}F_{21}E_{11}^{-1}F_{12}F_{22}^{-1}p_{j2} + O(\lambda). \end{aligned} \tag{48}$$

Suppose that $\lim_{\lambda \to 0+}(X + \lambda Y)^{ii}$ is finite. Then $p_{i2} = 0$ and

$$\lim_{\lambda \to 0+} (X + \lambda Y)^{ii} = p_{i1}'E_{11}^{-1}p_{i1} = \sum_{h=1}^{\rho} \frac{p_{hi}^2}{e_h} \tag{49}$$

which is independent of Y. Also

$$\lim_{\lambda \to 0+} (X + \lambda Y)^{ij} = p_{i1}'E_{11}^{-1}p_{ji} - p_{i1}'E_{11}^{-1}F_{12}F_{22}^{-1}p_{j2} \tag{50}$$

which is finite (though it depends on Y). Suppose that $(X + \lambda Y)^{ii}$ and $(X + \lambda Y)^{jj}$ are finite. Then $p_{i2} = 0$, $p_{j2} = 0$, and

$$\lim_{\lambda \to 0+} (X + \lambda Y)^{ij} = p'_{i1} E_{11}^{-1} p_{j1} = \sum_{h=1}^{p} \frac{p_{hi} p_{hj}}{e_h} \tag{51}$$

which is independent of Y.

Let us now assume that the probability distribution of the data of an experiment depends on $\varphi = (\varphi_1, \varphi_2, \cdots, \varphi_a)$ and that the information matrix with respect to φ is positive definite. Let us assume that the above distribution is independent of $\psi = (\psi_1, \psi_2, \cdots, \psi_b)$. Suppose now, that the parameters in which we are interested are $\theta = (\theta_1, \theta_2, \cdots, \theta_c)$ and $\eta = (\eta_1, \eta_2, \cdots, \eta_d)$ where $a + b = c + d$ and there is a one to one relationship between (φ, ψ) and (θ, η). In fact, suppose that

$$\theta = g_1(\varphi) \tag{53}$$

$$\eta = g_2(\varphi, \psi) \tag{54}$$

where the Jacobian of the transformation is not zero and where for each component of η the partial derivative with respect to some component of ψ does not vanish. We also suppose that the likelihood may be expressed in terms of (θ, η), that is,

$$L = u(\varphi) = w(\theta, \eta). \tag{55}$$

We are interested in the following information matrices:

$$U = E\{u'_\varphi u_\varphi\} \tag{56}$$

$$W = E\left\{\begin{pmatrix} w'_\theta w_\theta & w'_\theta w_\eta \\ w'_\eta w_\theta & w'_\eta w_\eta \end{pmatrix}\right\} = \begin{pmatrix} W_{\theta\theta} & W_{\theta\eta} \\ W_{\eta\theta} & W_{\eta\eta} \end{pmatrix} \tag{57}$$

where u_φ is a row vector whose ith component is $\partial u/\partial \varphi_i$. We shall also use the notation φ_θ to denote an $a \times c$ matrix whose (i, j) element is $\partial\varphi_i/\partial\theta_j$. We assume that U is positive definite and U^{-1} represents a covariance matrix $\sum_{\varphi\varphi}$. For our extension of the notion of the inverse of a matrix to be suitable, it should yield for us the following property.

PROPERTY 2. *The matrix W^{-1} may be decomposed as follows:*

$$W^{-1} = \begin{pmatrix} W^{\theta\theta} & W^{\theta\eta} \\ W^{\eta\theta} & W^{\eta\eta} \end{pmatrix} \tag{58}$$

where $W^{\theta\theta}$ is uniquely defined and is given by

$$W^{\theta\theta} = \theta_\varphi \sum_{\varphi\varphi} \theta'_\varphi = \sum_{\theta\theta} \tag{59}$$

and where the diagonal elements of $W^{\eta\eta}$ are infinite.

PROOF:

$$W = A' \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix} A$$

where

$$A = \begin{pmatrix} \varphi_\theta & \varphi_\eta \\ \psi_\theta & \psi_\eta \end{pmatrix} = \begin{pmatrix} \theta_\varphi & 0 \\ \eta_\varphi & \eta_\psi \end{pmatrix}^{-1}$$

is nonsingular. Furthermore

$$\left[W + \lambda A' \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} A\right]^{-1} = A^{-1} \begin{pmatrix} \sum_{\varphi\varphi} & 0 \\ 0 & I/\lambda \end{pmatrix} A'^{-1}$$

$$W_{A'\left(\begin{smallmatrix} 0 & 0 \\ 0 & I \end{smallmatrix}\right) A} = \begin{pmatrix} \theta_\varphi \sum_{\varphi\varphi} \theta_\varphi' & \theta_\varphi \sum_{\varphi\varphi} \eta_\varphi' \\ \eta_\varphi \sum_{\varphi\varphi} \theta_\varphi' & \eta_\varphi \sum_{\varphi\varphi} \eta_\varphi' + \lim_{\lambda \to 0+} \dfrac{\eta_\psi \eta_\psi'}{\lambda} \end{pmatrix}.$$

Property 1, together with the fact that not all components of η_ψ vanish, yields our desired results.

APPENDIX B. *Justification for the use of information matrices.* We sketch here a brief justification for the use of information matrices in our formulation. This justification presupposes that we are interested in the variances of the asymptotic distribution of the estimate based on our design. Rubin has shown [5] that under mild conditions these variances are greater than or equal to the diagonal elements of the inverse of the information matrix. On the other hand, if the design involves repeating a fixed number of these experiments in certain proportions, one (again under mild conditions) obtains equality. Since the "optimal" design using the information criterion involves repeating a fixed number of experiments in certain proportions, the sum of the variances of the asymptotic distributions of the estimates with this "optimal" design is actually equal to the minimum v_s which is a lower bound for the sum of the variances of the asymptotic distributions of the estimates for all designs.

APPENDIX C. *The relevance of sums of variances.* If one is interested in the parameters $\theta_1, \theta_2, \cdots, \theta_s$, it may be assumed that for a given estimate $t_1, t_2, \cdots, t_s$ there is a loss represented by a function

$$g(t, \theta) = g(t_1, t_2, \cdots, t_s, \theta_1, \theta_2, \cdots, \theta_s) \tag{60}$$

which as a function of the t_i is a minimum at $t_i = \theta_i$. If we assume that g is sufficiently well-behaved and that the sample is large enough so that the t_i are close to θ_i with large probability

$$g(t, \theta) = g(\theta, \theta) + \sum_{i,j=1}^{s} \frac{\partial^2 g(\theta, \theta)}{\partial t_i \partial t_j} (t_i - \theta_i)(t_j - \theta_j) + O(t - \theta)^3.$$

The "value" of our statistic is measured by how small $E\{g(t, \theta)\}$ is. For large samples (size n) we have, under mild conditions,

$$E\{g(t, \theta)\} = g(\theta, \theta) + \frac{1}{n} \sum_{i,j=1}^{s} a_{ij} \sigma_{ij} + o\left(\frac{1}{n}\right)$$

where $\| \sigma_{ij} \|$ is the covariance matrix of the asymptotic distribution of t and $a_{ij} = \dfrac{\partial^2 g(\theta, \theta)}{\partial t_i \partial t_j}$. A reasonable criterion of a good statistic t should then be that it minimizes

$$\sum_{i,j=1}^{s} a_{ij} \sigma_{ij}. \tag{61}$$

We now note that since g is minimized at $t = \theta$, the matrix $A = \| a_{ij} \|$ should be nonnegative definite. If A has rank ρ, it is possible to reduce the above expression to $\sum_{i=1}^{\rho} \sigma_{ii}$ by a linear transformation on θ. Ordinarily one would expect $\rho = s$ if one is interested in s parameters.

REFERENCES

[1] G. Elfving, "Optimum allocation in linear regression theory," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 255–62.

[2] R. A. Fisher, *Contributions to Mathematical Statistics*, Papers 10, 11, and 38, John Wiley and Sons, 1950.

[3] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[4] D. Blackwell, "Comparison of experiments," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1951.

[5] H. Rubin, "The asymptotic analogue of the theorem of Cramér and Rao," (Abstract), *Ann. Math. Stat.*, Vol. 19 (1948), p. 121.