Statistical Science
1999, Vol. 14, No. 2, 201–205

Gustav Elfving's Impact on Experimental Design

Herman Chernoff

1. INTRODUCTION

During my visit at Stanford University in the summer and fall of 1951, some problems proposed by the National Security Agency (NSA) for an Office of Naval Research (ONR) applied research grant led to two of my publications [1, 2] which had a profound effect on my future research. Both papers had relevance to issues in experimental design. One of these concerned optimal design for estimation. Among other results, it demonstrated that, asymptotically, locally optimal designs for estimating one parameter require the use of no more than k of the available experiments, when the distribution of the data from these experiments involves k unknown parameters. A trivial example would be that to estimate the slope of a straight line regression with constant variance, where the explanatory variable x is confined to the interval $[-1, 1]$, an optimal design requires observations concentrated at the two ends, $x = 1$ and $x = -1$.

Shortly after I derived this result, I discovered a related publication by Gustav Elfving [3]. While Elfving's result is restricted to k-dimensional regression experiments, it gives an elegant geometrical representation of the optimal design accompanied by an equally elegant derivation, which I still find pleasure in presenting to audiences who are less acquainted with this paper than they should be.

In some problems, practical considerations make it impossible to apply *optimal* designs. One beauty of the Elfving result is that the graphical representation of his result makes it rather clear how much is lost by applying some restricted suboptimal methods, and gives some guidance to wise compromises between optimality and practicality.

By 1950, experimental design was a well-established field of statistics. Major sources of application were in agriculture and chemistry, and the analysis of variance played an essential role. Combinatorial and number theoretic approaches, including that of finite geometry, tended to be efficient statistically and computationally for estimating many parameters because of the implicit tendencies to have balance, symmetry and orthogonality. One important consequence of the theory that has been slow in penetrating other sciences is that standard approaches of varying one causal variable at a time is inefficient compared to techniques where several variables are manipulated simultaneously. The weighing schemes of Yates [10] and Hotelling [5] made this point very clearly.

In spite of the activity in experimental design, a general theory for optimal design for estimation was lacking. The revolutionary impact of Elfving's contribution was due to the confluence of several factors. The problem he formulated was general in that it applied to much of the known literature, but was not too general. The results had a simple geometric interpretation and were computationally easy before computer technology was highly developed. Finally they were illustrated in terms of two unknown coefficients which made the results easy to comprehend.

2. THE ELFVING PROBLEM

Consider the regression

$$Y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + u_i, \quad i = 1, 2, \ldots, n,$$

where $\mathbf{x}_i = (x_{i1}, x_{i2})^T$ may be selected by the experimenter from a set S, β_1 and β_2 are unknown parameters, the residuals u_i are independent with mean 0 and constant, possibly unknown, variance σ^2 and the Y_i are observed. A particular *level* $\mathbf{x} \in S$ may be selected several times, yielding independent values of Y. It is desired to estimate

$$\theta = a_1\beta_1 + a_2\beta_2$$

for a specified value of $\mathbf{a} = (a_1, a_2)^T$. How should one allocate the n choices of $\mathbf{x}$ so as to yield a most informative estimate $\hat{\theta}$ of θ?

Herman Chernoff is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138 (e-mail: chernoff@hustat.harvard.edu).

Assuming that n_i observations are selected at $\mathbf{x}_i \in S$ for $i = 1, 2, \ldots, r$, we wish to minimize the variance of the linear unbiased estimate

$$\hat{\theta} = \sum_{i=1}^{r} b_i \overline{Y}_i$$

where $\overline{Y}_i$ is the average of the n_i observations at $\mathbf{x}_i$ subject to

$$\sum_{i=1}^{r} n_i = n, \quad n_i > 0,$$

and the unbiasedness condition

$$\sum_{i=1}^{r} b_i \mathbf{x}_i = \mathbf{a}.$$

For large values of n, we obtain an approximate solution by disregarding the integer nature of the n_i. An elegant argument where we first assume the b_i and the $\mathbf{x}_i$ are given, and optimize the variance

$$\mathrm{Var}(\hat{\theta}) = \sigma^2 \sum_{i=1}^{r} b_i^2 / n_i$$

with respect to the n_i, provides an equally elegant solution.

Let $\bar{S}$ be the convex hull of the set $S \cup (-S)$. The optimal design is represented by the point $\mathbf{z}$ where the ray from the origin through $\mathbf{a}$ penetrates $\bar{S}$. If $\mathbf{z}$ is a weighted average of points $\mathbf{x} \in S$ or $\mathbf{x} \in -S$, then an optimal design consists of using those points in frequencies proportional to the weights. Thus if S is a compact set, we need at most two levels of $\mathbf{x} \in S$. Moreover, for this optimal design

$$\mathrm{Var}(\hat{\theta}) = \frac{\sigma^2}{n} \left(\frac{|\mathbf{a}|}{|\mathbf{z}|} \right)^2.$$

The variance does not depend on $\boldsymbol{\beta}$ and the optimality of this design does not depend on the value of σ^2. To distinguish between the parameters β_1 and β_2 on one hand and σ^2 on the other, we shall refer to the former as the *coefficients*.

This first result is easily generalized in several ways. If the regression involves k coefficients linearly, that is, S is in k-dimensional space, the same geometric interpretation holds, and an optimal design requires at most k levels or values of $\mathbf{x} \in S$. If the variance of Y is $\sigma^2 h(\mathbf{x})$ and the cost of selecting $\mathbf{x}$ is $g(\mathbf{x})$, where g and h are known functions, and optimality requires minimum variance for a given total cost, a simple modification of the solution applies.

The variation of this problem where we are interested in estimating both coefficients rather than a single linear function of β_1 and β_2 raises a more complex issue. How should we evaluate a design for this problem? Elfving suggests the minimization of the expectation of a nonnegative quadratic function of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, that is,

$$q = E \sum_{i,j} a_{ij} (\hat{\beta}_i - \beta_i)(\hat{\beta}_j - \beta_j) = \mathrm{tr}(A\Sigma),$$

where $A = \| a_{ij} \|$ is nonnegative definite and Σ is the covariance matrix of the vector of estimated coefficients $\hat{\boldsymbol{\beta}}$. This criterion fits in naturally with a decision theoretic view of the estimation problem. If A is positive definite and $k = 2$, a linear transformation of $\boldsymbol{\beta}$ would lead to the use of

$$q = \mathrm{Var}(\hat{\beta}_1) + \mathrm{Var}(\hat{\beta}_2).$$

Here Elfving shows that the relevant levels of $\mathbf{x}$ must lie on an ellipsoid as far as possible from the origin. One consequence is that at most three values of $\mathbf{x} \in S$ are required if the number of coefficients $k = 2$. More generally we may need $k(k+1)/2$ levels of $\mathbf{x}$.

Given the primitive state of computing technology in 1952, Elfving suggested that the analytic and computational aspects could become nasty for values of k larger than 2, although he was aware that classical applications involving symmetry and orthogonality tended to be efficient.

3. RELATION BETWEEN MY RESULTS AND THOSE OF ELFVING

We recall that for the Elfving restriction to linear regression problems, the efficiency of a design is independent of the unknown coefficients. If we assume normal disturbances, the Fisher information matrix corresponding to a design D using $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ in proportions $w_1, w_2, \ldots, w_r$ is $nM(D)$, where n is the total number of observations, and

$$M(D) = \sum_{i=1}^{r} w_i \mathbf{x}_i \mathbf{x}_i^T.$$

The problem formulation which I proposed was based in part on a mistaken interpretation by M. A. Girshick and myself of the *practical* problem proposed by the NSA. For our interpretation it seemed natural to ask how many experimental levels were required to get good estimates, since each such level involved the construction of a measuring device. A regression formulation was inappropriate, but a large-sample asymptotic approach seemed reasonable. Thus, it was natural to think of a design consisting of a collection of independent *elementary experiments* $e_1, e_2, \ldots, e_n$ with $e_i \in \mathscr{E}$ and with information matrices $J(e_i)$. For such a design the total information is

$$nM(D) = \sum_{i=1}^{n} J(e_i).$$

Consider the cases where the distribution of the outcomes of the elementary experiment $e \in \mathscr{E}$ depends on $k = 2$ unknown parameters β_1 and β_2, and we wish to minimize the asymptotic variance of the maximum likelihood estimate $\hat{\beta}_1$ of β_1. Then we are led to minimize

$$n^{-1}M^{11} = n^{-1}(M_{11} - M_{12}M_{22}^{-1}M_{21})^{-1}$$

if M is positive definite. Discarding the integer condition on the frequency with which each experiment e may be performed, the set $\mathscr{M}$ of possible $M(D)$ is the convex hull of $T = \{J(e): e \in \mathscr{E}\}$ which can be represented as a set of points $(M_{11}(e), M_{12}(e), M_{22}(e))$ in three-dimensional space. Thus we wish to maximize

$$(M^{11})^{-1} = M_{11} - M_{12}M_{22}^{-1}M_{21}$$

defined on a convex set $\mathscr{M}$ in three-dimensional space. But then it follows that every point of $\mathscr{M}$ can be expressed as a convex combination of at most four points of T. However, M^{11} is clearly monotone in M_{11} and any optimal design must correspond to a boundary point of $\mathscr{M}$ and be expressible as a convex combination of at most three points of T. This implies immediately, with no analysis, that an optimal design can be constructed using at most three of the available experiments in $\mathscr{E}$ in appropriate proportions. With some analysis of the effective information for estimating β_1 when β_2 is unknown, that is, $M_{11} - M_{12}M_{22}^{-1}M_{21}$, it can be shown that two experiments will suffice for an optimal design. Moreover, this result can be generalized so that for the quadratic criterion of minimizing

$$q = \operatorname{tr}(AM^{-1}),$$

where A has rank r, we need at most $k + (k-1) + \cdots + (k-r+1) = r(2k-r+1)/2$ of the elements $e \in \mathscr{E}$.

In summary, my results were more general than those of Elfving in two respects. They dealt with experiments that were not necessarily of the regression type. Where Elfving's results applied to $r = 1$ and $r = k$, mine applied to the cases where $1 \le r \le k$.

In both approaches, the problems of singular information matrices had to be and were dealt with. Where Elfving's optimal designs were independent of the values of the coefficients, the optimality of the designs for the more general problem do often depend on the unknown values of the parameters. For this reason my results are asymptotically locally optimal, and a prior estimate of the values of the parameters is often required to derive good designs. As more information is accumulated, these designs can be improved.

Finally, Elfving presented an easily computed method of producing optimal designs. His approach is especially useful in those cases where practical considerations restrict the choice of designs. In such cases the geometric character of his solution clarifies how to deal with such restrictions.

There is a class of problems which are not of the regression type, but for which the Elfving solution applies. If the distribution of the outcome of an elementary experiment depends on one function of the k parameters, the information matrix $J(e)$ is singular, of rank 1, and therefore of the form $\mathbf{x}\mathbf{x}^T$. But then the problem is asymptotically equivalent to a regression problem, and the Elfving solution applies. In the probit problem of estimating $\mu - 2.87\sigma$ (0.002 dose response level) where there is a response at dose level d with probability $\Phi(\sigma^{-1}(d-\mu))$ and Φ is the standard normal c.d.f., the asymptotically optimal design is easily shown to involve dose levels $\mu \pm 1.57\sigma$. Clearly this design is local since it depends on the values μ and σ of the unknown parameters of the probit model.

4. LATER RESULTS

In 1959 Elfving [4] summarized the current state of research on the design of linear experiments in an article in the Cramér festschrift [4]. Here he discussed minimax designs and admissibility.

Two kinds of minimaxity were considered. In one, labeled s.p., one attempts to find the design for which the maximum of the variances of the estimates of each of the k coefficients is minimized. The second, labeled st.f., minimizes

$$\max_{\|\mathbf{c}=1\|} \operatorname{Var}(\mathbf{c}^T\hat{\boldsymbol{\beta}}).$$

In these cases theorems, due to Moriguti and Ehrenfeld and Gustafson, which present sufficiency conditions invoking symmetry and orthogonality, are described. Also there is some consideration of the case where minimax is applied to a limited subset of r of the k coefficients.

An experimental design D is regarded as *admissible* if there is no alternative design D^* which yields smaller variances for every linear function of the parameters. This implies that, for each D^*, $M(D^*) - M(D)$ is not nonnegative definite. It is pointed out that admissibility is a property of the *spectrum* of D, that is, the elementary experiments from which D is formed, and not to the proportions in which they are used. A characterization is given of the elementary experiments involved in an admissible

design. These must lie on a positive-definite quartic centered at the origin. In this section, Elfving refers frequently to unpublished notes of L. J. Savage. Elfving has been generous in his attribution to other workers, and it would require some effort to disentangle his contributions from those of the others from his presentation of the current state of the subject.

At the end of this paper, there is a note added in proof in which he mentions that he had recently had access to a very interesting paper by Kiefer and Wolfowitz, "Optimum designs in regression problems," which was to appear in the *Annals of Mathematical Statistics*. In fact he was a referee for that article, which represented the first step in the next revolutionary development stemming from his work.

5. THE KIEFER–WOLFOWITZ FORMULATION

Around 1959, based in part on Elfving's results, another revolutionary development appeared in experimental design for estimation in regression experiments. In a series of papers by Kiefer and Wofowitz, and then by Karlin and Studden, further insights and extensions of Elfving's results appeared using game theoretic and minimax ideas.

In cases where we are interested in estimating a linear function of the k unknown coefficients, the computational problem implied by the results of Elfving and Chernoff requires the minimization of a function of the $2k-1$ variables consisting of the k elementary experiments and of the $k-1$ frequencies allocated to these experiments. Kiefer and Wolfowitz [7] indicated how this problem can be divided into two lesser problems of k and $k-1$ dimensions. The choice of elementary experiments or levels was reduced to a Chebyshev approximation problem which was already solved in some special cases of interest.

In cases where we are interested in estimating all of the k coefficients, a remarkable result by Kiefer and Wolfowitz [8], later elaborated by Karlin and Studden [6], establishes the equivalence of two optimality criteria. These are D-optimality, where we minimize the determinant of the covariance matrix of the least squares estimates of the coefficients, and A-optimality, where we minimize the maximum variance of the estimated regression. Note that, for each level, the variance of the estimated regression depends on the level $\mathbf{x} \in S$, and our criterion applies, for each design, the maximum of the variance of $\hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ for $\mathbf{x} \in S$.

This minimax criterion seems generally more natural and relevant for problems where extrapolation is not an issue, than the two minimax criteria discussed by Elfving in [4]. On the other hand, when we are interested in $r < k$ of the coefficients, the extension of the equivalence result equates the design which minimizes the determinant of the covariance of the estimates of the r parameters to a minimax variance which does not seem especially relevant as a natural criterion.

Incidentally, Studden [9] later described a natural, if rather complicated, extension of Elfving's geometric results to the selection of experiments which minimize the expectation of a quadratic function of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$.

These results were revolutionary and insightful and led to a rich literature and many applications using the developing computer technology. However, I have some criticisms of the direction in which these results pointed.

My main objection is to the concentration on D-optimality. I much prefer the quadratic criterion of minimizing $\operatorname{tr}(A\Sigma)$ for some nonnegative-definite symmetric A. The latter criterion makes sense from a decision theory point of view where the costs of error are approximated by a nonnegative quadratic function. The rank of A determines the number of linearly independent coefficients of concern. One of my objections to D-optimality is the sensitivity of the determinant to the variance of the estimates of the individual coefficients. More important is that the invariance property of this criterion, which is regarded by some as a desirable property, is undesirable. It is a surrender of the decision maker's ability to measure the cost of error to the mathematical structure of the problem. To be more specific, if $\boldsymbol{\beta}^* = C\boldsymbol{\beta}$ is a linear transformation of the coefficients, the D-optimal criterion applied to $\boldsymbol{\beta}^*$ leads to the same solution as that applied to $\boldsymbol{\beta}$. The experimenter's costs, considered in A above nowhere enter into the application of this criterion.

Interestingly enough the equivalence result shows that the D-optimal criterion makes sense for the minimaxer who is not concerned with extrapolation, that is, interested in estimating the regression only for $\mathbf{x} \in S$. But this equivalence result degenerates to something less natural for the experimenter if not all of k coefficients are of interest.

A minor quarrel that I have had with this stream of papers is a strange use of notation. Rather than take

$$Y = \mathbf{x}^T\boldsymbol{\beta} + u \quad \text{for } \mathbf{x} \in S,$$

where S is compact, the notation used is

$$Y = \mathbf{f}(x)^T\boldsymbol{\beta} + u \quad \text{for } x \in \mathscr{X},$$

where $\mathscr{X}$ is compact and **f** is a continuous vector-valued function of x. This notation is presumably derived from an interest in the application

$$Y = \beta_0 + \beta_1 x + \cdots + \beta_{k-1} x^{k-1} \quad \text{for } a \le x \le b,$$

where Chebyshev approximations are standard. Personally, I like the standard notation in regression and find this *new* use of **f** disconcerting and regret its prevalence.

6. SUMMARY

Experimental design is one of the major keystones of statistical theory and application. Its development is one of R. A. Fisher's major contributions to science. The work of Elfving introduced a fundamentally new direction in this field. It provided important insights into the efficiency of good designs and served as a stepping stone for later developments which have exploited computer technology.

REFERENCES

[1] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of hypotheses based on the sums of observation. *Ann. Math. Statist.* **23** 493–507.

[2] CHERNOFF, H. (1953). Locally optimal designs for estimating parameters. *Ann. Math. Statist.* **24** 586–602.

[3] ELFVING, G. (1952). Optimum allocation in linear regression theory. *Ann. Math. Statist.* **23** 255–262.

[4] ELFVING, G. (1959). Design of linear experiments. In *Probability and Statistics*: *The Harald Cramér Volume* (U. Grenander, ed.) 58–74. Wiley, New York.

[5] HOTELLING, H. (1944). Some improvements in weighing and other experimental techniques. *Ann. Math. Statist.* **15** 297–306.

[6] KARLIN, S. and STUDDEN, W. J. (1966). Optimal experimental designs. *Ann. Math. Statist.* **37** 783–815.

[7] KIEFER, J. and WOLFOWITZ, J. (1960). Optimum designs in regression problems. *Ann. Math. Statist.* **30** 271–294.

[8] KIEFER, J. and WOLFOWITZ, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.* **12** 363–366.

[9] STUDDEN, W. J. (1971). Elfving's theorem and designs for quadratic loss. *Ann. Math. Statist.* **42** 1613–1621.

[10] YATES, F. (1935). Complex experiments. *J. Roy. Statist. Soc.* **2** 181–247.