OPTIMAL DESIGN OF EXPERIMENTS

Herman Chernoff [1/]
Stanford University

1. INTRODUCTION. I would like to discuss some aspects of the theory of optimal design of experiments with particular emphasis on its relevance to the practice of statistics. There are two major branches of classical statistics, Estimation and Testing of Hypotheses, for which the theory of optimal design yields different results. Because of the time limitation, I shall confine my attention to certain results and examples in the theory of estimation.

2. SOME EXAMPLES. To illustrate the theory let us consider three examples. The first example is a well known one with a trivial solution. That is the one of estimating the slope of a regression (straight line). More specifically we have

Example 1.

The experimenter may choose any number y between -1 and +1. This number y designates an elementary experiment which corresponds to observing

$$z = \alpha + \beta y + u$$

where u is normally distributed with mean 0 and variance 1 and α and β are unknown parameters. The experimenter is permitted to select a design consisting of n values $y_1, y_2, \ldots, y_n$, with possible repetitions. The design corresponds to performing the n designated experiments independently. It is desired to select a design which will yield the best possible estimate of the slope β.

It is well known and it is intuitively obvious that the best design consists of selecting $y = -1$ and $y = +1$ each half the time (providing n is even).

[1/] This work was supported in part by Office of Naval Research Contract Nonr-225(52) at Stanford University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Another example which is of some current interest, having been discussed in yesterday's paper by Mr. Langlie [5] on a problem in reliability, and which is also relevant to the problem of Probit Analysis, may be expressed as follows:

Example 2.

A device, which may be used only once, can operate successfully under a stress s with probability

$$p = \int_{\frac{s-\mu}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt .$$

In other words one may say that the strength of the device, as measured by the maximum stress under which it will operate successfully, is normally distributed with unknown mean μ and variance σ^2. It is desired to select a design consisting of the choice of stress levels $s_1, s_2, \ldots, s_n$ which will yield an optimal estimate of $\mu - k\sigma$. The elementary experiment, designated s, consists of course of observing the success or failure of the device when used under stress s.

Finally a third problem which was discussed in detail in a recent paper of mine [2] deals with accelerated life testing. Here we wish to estimate the mean life time of a device when used under an environment of ordinary stress conditions. If, this mean lifetime is great and it is desired to have the estimate soon, then it is necessary to accelerate. The device is subjected to a much larger than ordinary stress. The results of such accelerated life testing can be relevant only if one assumes some form of relationship connecting the mean lifetime under various stresses. As an approximation we shall assume a quadratic relationship for some limited range. In addition since time is of the essence we shall assume that the cost of observing a device under stress s is proportional to the mean lifetime under that stress. Let us be more specific.

Example 3.

A device under stress environment s has lifetime T with an exponential distribution with failure rate (reciprocal of mean) given by

$$\varphi = \theta_1 s + \theta_2 s^2 \qquad \text{for } 0 \leq s \leq s^*$$

where θ_1 and θ_2 are unknown parameters. It is desired to estimate the failure rate under the ordinary stress s_o. This is

$$\varphi_o = \theta_1 s_o + \theta_2 s_o^2 .$$

An elementary experiment designated by s consists of observing the lifetime T of a device subjected to the environment s. The cost of the experiment s is

$$C(s) = c(\theta_1 s + \theta_2 s^2)^{-1} .$$

It is desired to select a design consisting of experiments $s_1, s_2, \ldots;$ $0 \leq s_i \leq s^*$, so as to obtain an optimal estimate of φ_o for a specified total cost.

Each of these examples has certain elements in common. Each may be regarded as a special case of the following general formulation. There is a set $\mathcal{E}$ of available elementary experiments e. In each case the distribution of the data of an experiment depends on the experiment and on k unknown parameters represented by $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. We wish to estimate some function $g(\theta_1, \theta_2, \ldots \theta_k)$ of the parameters. A design consists of the independent performance of experiments $e_1, e_2 \ldots$ with possible repetitions. It is desired to find a design which yields the best possible estimate of $g(\theta_1, \theta_2, \ldots, \theta_k)$ for a specified total cost or for a specified number of observations.

3. THE LINEAR REGRESSION MODEL. In 1952, Elfving [4] derived an elegant geometric solution to the optimal design problem for a special but important case of the above general formulation. As we shall see this result is applicable to a large variety of problems. Let ξ be a set of experiments e denoted by (y_1, y_2). The experiment e consists of observing

$$Z = \theta_1 y_1 + \theta_2 y_2 + u$$

where u is normally distributed with mean 0 and variance 1. It is desired to obtain an optimal estimate of $a_1 \theta_1 + a_2 \theta_2$ using a design consisting of n observations. The first example of estimating the slope of a straight line is a special case of Elfving's linear regression model where ξ is the set of points (1, y) with $-1 \leq y \leq 1$, and $(a_1, a_2) = (0, 1)$.

Elfving's solution consists of constructing a set S which is the smallest convex set containing the points (y_1, y_2) of ξ and their negatives $(-y_1, -y_2)$. Then extend the vector from (0, 0) to (a_1, a_2) until it penetrates the set S. The point of penetration (w_1, w_2) represents the optimal design. If this point is one of the original points (y_1, y_2) or $(-y_1, -y_2)$ the optimal design consists of repeating (y_1, y_2) n times. Otherwise the point of penetration is on a line segment connecting points corresponding to two of the original experiments (or their negatives). Then the optimal design consists of repeating these two experiments in proportions given by the distances from (w_1, w_2) to the two points. The greater proportion corresponds to the experiment closer to (w_1, w_2). Finally the variance of the least squares estimate based on this design is

$$\sigma^2_{\hat{\phi}} = \left[n(w_1^2 + w_2^2)\right]^{-1}(a_1^2 + a_2^2) = a_1^2/nw_1^2 = a_2^2/nw_2^2$$

This solution can be illustrated with example 1. Here S is the square whose corners are (1, 1) and (-1, -1) corresponding to y = 1 and (1, -1) and (-1, 1) corresponding to y = -1. The line from (0,0) through $(a_1, a_2) = (0, 1)$ penetrates S at (0, 1) which is halfway between (1, 1) and (-1, 1). Thus the optimal design consists of repeating the experi-

ments corresponding to $y = 1$ and $y = -1$ each half the time (as was well known). Furthermore the variance of the estimate of β should be $1/n$.

Elfving's result applies in the obvious fashion to experiments involving k parameters. Here we need repeat at most k of the available experiments in certain proportions to obtain the optimal estimate.

4. RESULTS FOR THE MORE GENERAL PROBLEM. As mentioned in the preceeding section the problem treated by Elfving is a special case of the more general one formulated in section 2. For this more general problem, related results have been obtained [1]. These results concern designs which are asymptotically locally optimal. We shall defer the interpretation of these adjectives until the discussion of Example 2 in section 5.

It was shown that asymptotically locally optimal designs depend on the form of the matrix $J(e)$ which is defined as Fisher's information matrix divided by the cost of the experiment e. In other words if experiment e has cost $C(e)$ and yields data X with probability distribution $f(x, \theta, e)$, Fisher's information matrix is

$$I(e) = \left\| E\left\{ \frac{\partial \log f(X, \theta, e)}{\partial \theta_i} \quad \frac{\partial \log f(X, \theta, e)}{\partial \theta_j} \right\} \right\|$$

and the information per unit cost is

$$J(e) = I(e)/C(e).$$

Clearly if the cost of experimentation is constant one need concern oneself only with $I(e)$. The relevance of Fisher's Information derives from its well known additive properties and the fact that the maximum-likelihood estimate $\hat{\theta}_n$ based on the outcome of n independent repetitions of e, has an approximately normal distribution with mean θ and covariance matrix $[nI(e)]^{-1}$ for large n.

When it is desired to estimate one function of the k parameters, there are asymptotically locally optimal designs which involve at most k of the experiments of ξ in certain proportions. This result which corresponds to one of Elfving's results, together with the use of Fisher's Information, permits one to reduce the calculation of optimal designs to the maximization of a function of a fixed number of variables.

In the linear regression problem of Elfving, the information matrix for $e = (y_1, y_2)$ is

$$I = \| y_i y_j \| = J.$$

Since asymptotically optimal designs are determined by the information per unit cost it follows that for any problem where J(e) can be put in the above form, the solution is the same as Elfving's with a_i replaced by $\frac{\partial g}{\partial \theta_i}$.

The illustration of the next section will help clarify the meaning of these results. In the meantime it may be remarked that if for each experiment the distribution of the outcome depends on only one function of the parameters, J(e) can be put in the above form and Elfving's results are applicable. In particular they are applicable to both examples 2 and 3.

5. ILLUSTRATION. We shall find it informative to illustrate the method with example 2. Here the outcome of the experiment s is success or failure where the probability of success is

$$p(s, \mu, \sigma) = \int_{\frac{s-\mu}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt = 1 - \Phi\left(\frac{s-\mu}{\sigma}\right)$$

where Φ is the normal cdf. In other words the role of the density $f(X, \theta, e)$ is played by

$$f = p^{X}(1 - p)^{1-X}$$

where $X = 1$ for success and 0 for failure.

$$\log f = X \log p + (1-X) \log(1 - p)$$

$$\frac{\partial \log f}{\partial \mu} = \frac{X - p}{p(1 - p)} \frac{\partial p}{\partial \mu}$$

$$\frac{\partial \log f}{\partial \sigma} = \frac{X - p}{p(1 - p)} \frac{\partial p}{\partial \sigma} .$$

Since $E\{(X-p)^2\} = p(1-p)$,

$$J(s) = I(s) = \left[p(1 - p) \right]^{-1} \left\| \begin{matrix} \left(\frac{\partial p}{\partial \mu}\right)^2 & \frac{\partial p}{\partial \mu} \frac{\partial p}{\partial \sigma} \\ \frac{\partial p}{\partial \mu} \frac{\partial p}{\partial \sigma} & \left(\frac{\partial p}{\partial \sigma}\right)^2 \end{matrix} \right\|$$

$$J(s) = \left\| y_i y_j \right\|$$

where

$$y_1(s) = \left[p(1-p) \right]^{-1/2} \frac{\partial p}{\partial \mu} = \left[2 \pi p(1-p) \right]^{-1/2} \sigma^{-1} \exp\left[-(s-\mu)^2 / 2\sigma^2 \right]$$

and

$$y_2(s) = \left[p(1-p) \right]^{-1/2} \frac{\partial p}{\partial \mu} = \left[2 \pi p(1-p) \right]^{-1/2} (s-\mu) \sigma^{-2} \exp\left[-(s-\mu)^2 / 2\sigma^2 \right].$$

Next we plot the set of points $\left[y_1(s), y_2(s) \right]$ in Figure 1. We add the negatives of these points and construct S the smallest convex set containing them. We note that for $s = \mu + t\sigma$, $y_2(s)/y_1(s) = t$. We also note the curve of $\left[y_1(s), y_2(s) \right]$ reaches its maximum and minimum at $s = \mu \pm k_o \sigma$ where $k_o = 1.57$. Finally, since we wish to estimate $\mu - k\sigma$,

we draw the vector from (0, 0) through (1, -k), i.e. the line through the origin with slope -k, and note where it penetrates the convex set S.

Clearly there are two cases.

Case 1. $|k| \leq k_o$. Here the vector penetrates S at one of the original $[y_1(s), y_2(s)]$ points. In fact this point corresponds to $s = \mu - k\sigma$ and hence the optimal design consists of using $s = \mu - k\sigma$ for all observations.

Case 2. $|k| > k_o$. Here the vector penetrates S at the straight line section of the boundary. The optimal design consists of applying the stress levels $\mu - k_o\sigma$ and $\mu + k_o\sigma$ in proportions $k+k_o$ to $k-k_o$. In cases 1 and 2 the formal application of the formula for the variance of the maximum likelihood estimate of $\mu - k\sigma$ based on the optimal design is given by

$$2\pi\sigma^2 \Phi(k)\left[1-\Phi(k)\right] \; e^{k^2} \; n^{-1}$$

in case 1, and

$$2\pi\sigma^2 \Phi(k_o)\left[1-\Phi(k_o)\right] \; e^{k_o^2} \; k_o^{-2} \; k^2 \; n^{-1} = 1.64\,\sigma^2 k^2 n^{-1}$$

in case 2.

6. THE RELEVANCE OF OPTIMAL DESIGN. Now we shall find the illustrative example helpful in interpreting the results of the theory of optimal design of experiments and in understanding its relevance in practical applications. For simplicity let us confine our attention to case 2 at first.

First we note one very peculiar aspect of the optimal design. Since it involves using stress levels $\mu - k_o\sigma$ and $\mu + k_o\sigma$, to apply it one must know μ and σ. But if one knew μ and σ, there would be no need to experiment. While this seems to be ridiculous, a glance at figure 1 indicates that if one used an approximation to $\mu \pm k_o\sigma$, one would have a rather good approximation to the optimal design. Thus there is surprisingly little loss of efficiency when one is not certain about μ and σ. It is this property that the word local is used to describe. In other words our design would be efficient if we knew the parameters and

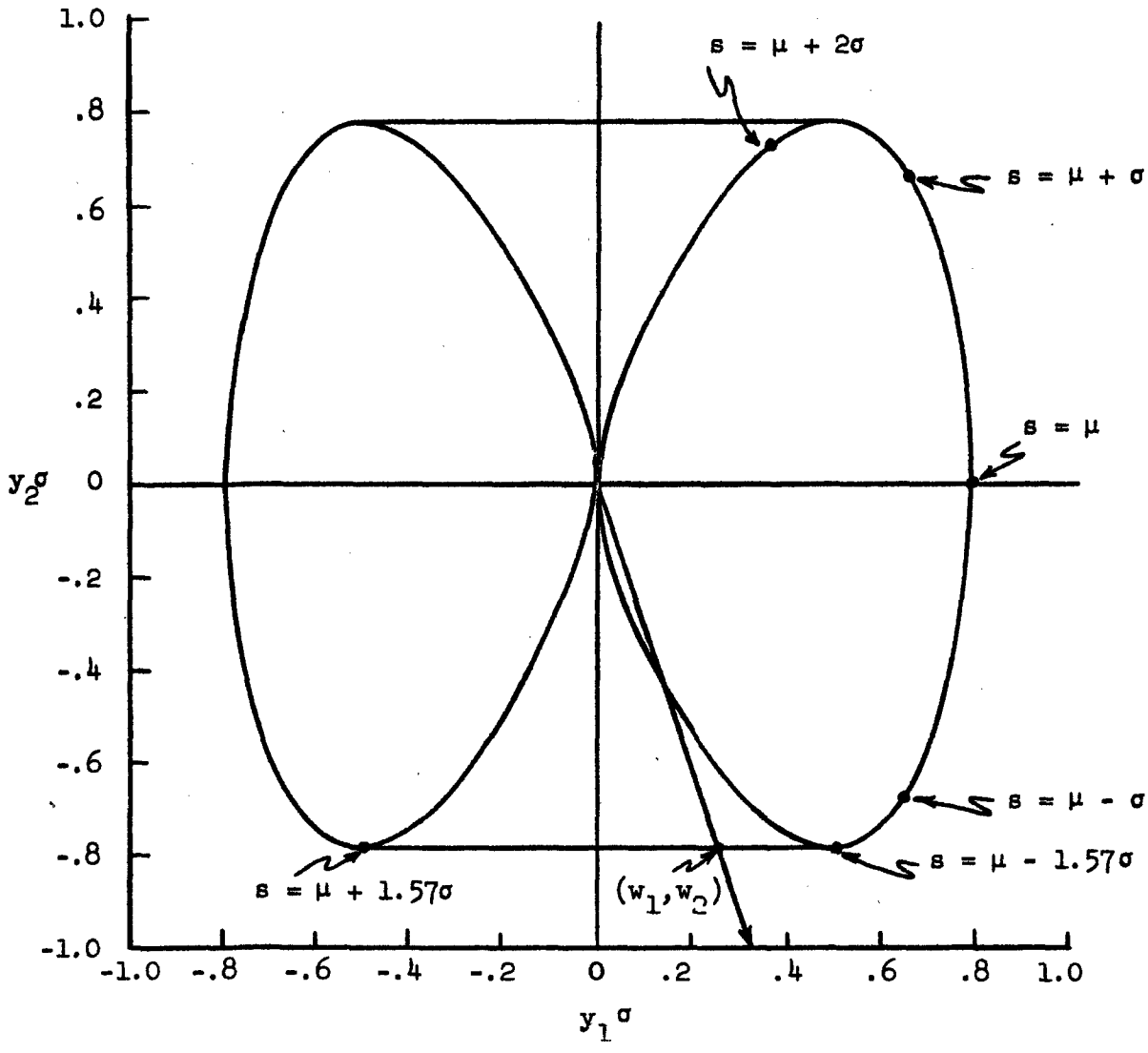
1.0
.8
.6
.4
.2
0
-.2
-.4
-.6
-.8
-1.0
$y_2\sigma$
-1.0 -.8 -.6 -.4 -.2 0 .2 .4 .6 .8 1.0
$y_1\sigma$
$s = \mu + 2\sigma$
$s = \mu + \sigma$
$s = \mu$
$s = \mu - \sigma$
$s = \mu - 1.57\sigma$
$s = \mu + 1.57\sigma$
(w_1, w_2)

Figure 1

is approximately efficient if we use an approximation to the unknown parameters.

This raises the issue of the adjective asymptotic. If one had a large sample available, one could use some of the initial observations to derive an initial estimate of and on which to base an approximation to the optimal design. Furthermore the qualification asymptotic derives from a couple of other aspects. First, the properties relating the variance of the approximate distribution of the maximum likelihood to the information matrix and giving the efficiency of this estimate is based on asymptotic theory assuming large sample size. A second and relatively minor point, is illustrated by example 1 if an odd number of observations are available. The optimal design calls for putting half the observations at +1 and half at -1. This is impossible in a trivial way when n is odd. On the other hand the effect of this impossibility is negligible when n is large.

Having seen how we must qualify the tern optimal by the adjectives local and asymptotic, we can now consider a more fundamental issue. Briefly, our optimal design is simply impractical. Only in the rather unrealistic context where I had absolute faith in the model would I consider this as a solution. In fact, any reasonable statistician would insist on using several other stress levels at least to check on the model.

Another unreasonable aspect of our optimal design arises from its derivation based on the single minded purpose of obtaining a good estimate of one function $g(\theta_1, \theta_2, \ldots, \theta_k)$ of the parameters. In many practical problems, experimentation is used to serve several purposes simultaneously.

One may reasonably inquire about what function does the theory of optimal design serve, if (1) the optimality must be qualified as locally asymptotically optimal and (2) the designs it yields are unreasonable. Basically the functions are the following. First, the theory provides a yardstick for comparison purposes. If the designs proposed yesterday by Mr. Langlie, or the Up and Down Method [3, p. 319], or some other practical design turns out to be relatively efficient compared to our solution (as measured by asymptotic variance) then clearly there is no point in attempting to improve on this aspect of these methods. If, on the other hand, one of these methods were to have a low efficiency, then one is forced to delve deeper to see what, if anything, can be done to improve the design.

Second, theory not only presents an optimal design but indicates rather clearly how this design can be modified with relatively low loss of efficiency. The theory serves to direct the attention of the practical statistician toward designs which combine relatively high efficiency with practical utility when robustness and multi-purpose considerations are taken into account.

7. MISCELLANEOUS COMMENTS. I would like to conclude this paper with a few assorted comments. First, the proposed solution to example 2 in case 1 when $|k| \leq k_0$ consists of repeating one experiment n times. Not only is this solution impractical, but from a theoretical point of view it represents a degenerate situation. When a single level s is used, one can use the data to estimate only

$$p(s, \mu, \sigma) = \int_{\frac{s-\mu}{\sigma}}^{\infty} (2\pi)^{-1/2} e^{-t^2/2}\, dt$$

or functions of $p(s, \mu, \sigma)$. Then one can check whether $\frac{s-\mu}{\sigma}$ is in fact close to k (as it should be if the design were optimal). But not knowing σ, one can not estimate $\mu - k\sigma$. Thus the formula for the asymptotic variance presented at the end of section 5 is meaningful only as an approximation to the case where several levels of stress close to the optimal one were used. Alternatively one could regard $p(1-p)n^{-1}$ as the asymptotic variance of the estimate of p.

For a large sample sequential procedure, it seems clear that our theory is applicable. If one were to reestimate the parameters after each observation, and use these estimates to derive approximations to the optimal design, the resulting procedure should be asymptotically optimal in the sequential version and the adjective local need not be applied.

What is more interesting, perhaps, is the study of the "not so large" sample sequential case. Here even the following seemingly simple problem proposed by Harold Gumbel does not have a simple solution. Suppose that experiment e_i yields observation X_i which is normally distributed with unknown mean μ and unknown variance σ_i^2, i=1, 2, and it is desired to estimate μ. In other words, two measuring instruments of unknown accuracy are available. How should one select between the

two experiments sequentially so as to obtain a good estimate efficiently when the sample size is not necessarily very large?

BIBLIOGRAPHY

(1) Chernoff, H. (1953), Locally optimal designs for estimating parameters, Ann. Math. Statist. 24 586-620.

(2) Chernoff, H. (1962), Optimal accelerated life designs for estimation, Technometrics 4, 381-408.

(3) Dixon, W. J., and Massey, F. J. (1957), Introduction to Statistical Analysis, (2nd ed.) McGraw Hill, New York.

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

(5) Langlie, H. J. (1962), A Reliability Test Method for "One-Shot" Items; presented at the Eighth Conference on The Design of Experiments in Army Research, Development and Testing.