ON THE EFFICIENT DESIGN OF STATISTICAL INVESTIGATIONS

By Abraham Wald

Columbia University

1. Introduction. A theory of efficient design of statistical investigations has been developed by R. A. Fisher[1] and his followers mainly in connection with agricultural experimentation. However, the same methods can be applied to other fields also. All statistical designs treated in the aforementioned theory refer to problems of testing linear hypotheses. By testing a linear hypothesis we mean the following problem: Let $y_1, \cdots, y_N$ be N independently and normally distributed variates with a common variance σ^2. It is assumed that the expected value of y_α is given by

$$E(y_\alpha) = \beta_1 x_{1\alpha} + \beta_2 x_{2\alpha} + \cdots + \beta_p x_{p\alpha} \qquad (\alpha = 1, \cdots, N) \tag{1}$$

where the quantities $x_{i\alpha} (i = 1, \cdots, p; \alpha = 1, \cdots, N)$ are known constants and $\beta_1, \cdots, \beta_p$ are unknown constants. The coefficients $\beta_1, \cdots, \beta_p$ are called the population regression coefficients of y on $x_1, x_2, \cdots$, and x_p, respectively. The hypothesis that the unknown regression coefficients $\beta_1, \cdots, \beta_p$ satisfy a set of linear equations

$$g_{i1}\beta_1 + \cdots + g_{ip}\beta_p = g_i \qquad (i = 1, \cdots, r; r \leqq p), \tag{2}$$

is called a linear hypothesis. The problem under consideration is that of testing the hypothesis (2) on the basis of the observed values $y_1, \cdots, y_N$.

In many cases the experimenter has a certain amount of freedom in the choice of the values $x_{i\alpha}$. The efficiency of the test is greatly affected by the values of $x_{i\alpha}$. The statistical investigation is efficiently designed if the values $x_{i\alpha}$ are chosen so that the sensitivity of the test is maximized. Let us illustrate this by a simple example. Suppose that x and y have a bivariate normal distribution and we want to test the hypothesis that the regression coefficient β of y on x has a particular value β_0. Suppose, furthermore, that the test has to be carried out on the basis of N pairs of observations $(x_1, y_1), \cdots, (x_N, y_N)$, where the experiments are performed in such a way that $x_1, \cdots, x_N$ are not random variables but have predetermined fixed values. It is known that the variance of the least square estimate b of β is inversely proportional to $\sum_{\alpha=1}^{N} (x_\alpha - \bar{x})^2$ where $\bar{x} = (x_1 + \cdots + x_N)/N$. Hence, if we can freely choose the values $x_1, \cdots, x_N$ in a certain domain D, the greatest sensitivity of the test will be achieved by choosing $x_1, \cdots, x_N$ so that $\Sigma(x_\alpha - \bar{x})^2$ becomes a maximum.

In the next section we will introduce a measure of the efficiency of the design

[1] See for instance R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, London, 1935.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

of a statistical investigation for testing a linear hypothesis. In sections 3 and 4 it will be shown that some well known experimental designs, used widely in agricultural experimentation, are most efficient in the sense of the definition given in section 2.

2. A measure of the efficiency of the design of a statistical investigation for testing a linear hypothesis. The hypothesis (2) can be reduced by a suitable linear transformation to the canonical form

$$\beta_1 = \beta_2 = \cdots = \beta_r = 0, \qquad (r \leq p). \tag{3}$$

Hence, we can restrict ourselves without loss of generality to the consideration of the hypothesis (3).

Denote $\sum_{\alpha=1}^{N} x_{i\alpha}x_{j\alpha}$ by a_{ij} and let the matrix $|| c_{ij} ||$ be the inverse of the matrix $|| a_{ij} ||$ $(i, j = 1, \cdots, p)$. Denote by b_i the least square estimate of β_i $(i = 1, \cdots, p)$. It is known that the estimates $b_1, \cdots, b_p$ have a joint normal distribution with mean values $\beta_1, \cdots, \beta_p$, respectively. It is furthermore known that the covariance of b_i and b_j is equal to $c_{ij}\sigma^2$. The statistic used for testing the hypothesis (3) is given by

$$F = \frac{N - p}{r} \frac{\sum_{l=1}^{r}\sum_{m=1}^{r} a^*_{lm} b_l b_m}{\sum_{\alpha=1}^{N} (y_\alpha - b_1 x_{1\alpha} \cdots - b_p x_{p\alpha})^2} \tag{4}$$

where $|| a^*_{lm} ||$ is the inverse of $|| c_{lm} ||$ $((l, m = 1, \cdots, r)$. The statistic F has the F-distribution with r and $N - p$ degrees of freedom. The critical region for testing the hypothesis (3) is given by the inequality

$$F \geqq F_0, \tag{5}$$

where the constant F_0 is determined so that the probability that $F \geq F_0$ (calculated under the assumption that (3) holds) is equal to the level of significance we wish to have.

It is known that the power function[2] of the critical region (5) depends only on the single parameter

$$\lambda = \frac{1}{\sigma^2} \sum_{l=1}^{r}\sum_{m=1}^{r} a^*_{lm}\beta_l\beta_m. \tag{6}$$

Furthermore this power function is a monotonically increasing function of λ. The coefficients a^*_{lm} are functions of the quantities $x_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$. The choice of the values $x_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$ is the better the greater the corresponding value of λ. If $r = 1$, the expression λ

[2] See for instance P. C. TANG, "The power function of the analysis of variance tests," *Stat. Res. Mem.*, Vol. II, 1938.

reduces to $\frac{1}{\sigma^2} a_{11}^* \beta_1^2$. Hence, if $r = 1$, we maximize λ by maximizing a_{11}^*. Since $a_{11}^* = 1/c_{11}$, we maximize λ by minimizing c_{11}. Thus, if $r = 1$, we can say that we obtain the most powerful test by minimizing c_{11}, i.e. by minimizing the variance of b_1. If $r > 1$, the difficulty arises that no set of values $x_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$ can be found for which λ becomes a maximum irrespective of the values of the unknown parameters $\beta_1, \cdots, \beta_r$. Hence, if $r > 1$, we have to be satisfied with some compromise solution. For this purpose let us consider the unit sphere

$$\beta_1^2 + \cdots + \beta_r^2 = 1, \tag{7}$$

in the space of the parameters $\beta_1, \cdots, \beta_r$. It is known that the smallest root in ρ of the determinantal equation

$$\begin{vmatrix} a_{11}^* - \rho & a_{12}^* & \cdots & a_{1r}^* \\ a_{21}^* & a_{22}^* - \rho & \cdots & a_{2r}^* \\ \cdot & \cdot & \cdots & \cdot \\ a_{r1}^* & a_{r2}^* & \cdots & a_{rr}^* - \rho \end{vmatrix} = 0, \tag{8}$$

is equal to the minimum value of $\sigma^2\lambda$ on the unit sphere (7). Similarly the greatest root of (8) is equal to the maximum value of $\sigma^2\lambda$ on the sphere (7). The compromise solution of maximizing the smallest root of (8) seems to be a very reasonable one. However, for the sake of certain mathematical simplifications, we propose to maximize the product of the r roots of (8). Since the product of the roots of (8) is equal to the determinant

$$\begin{vmatrix} a_{11}^* & \cdots & a_{1r}^* \\ \cdot & \cdots & \cdot \\ a_{r1}^* & \cdots & a_{rr}^* \end{vmatrix}, \tag{9}$$

we have to maximize the determinant (9). The value of the determinant $|c_{lm}|$ $(l, m = 1, \cdots, r)$ is the reciprocal of that of (9). Hence we maximize (9) by minimizing the determinant $|c_{lm}|$. The generalized variance of the set of variates $b_1, \cdots, b_r$ is equal to the product of σ^{2r} and the determinant $|c_{lm}|$. Thus, our result can be expressed as follows: The optimum choice of the values of $x_{i\alpha}$ is that for which the generalized variance of the variates $b_1, \cdots, b_r$ becomes a minimum.

Any set of pN values $x_{i\alpha}$ $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$ can be represented by a point in the pN-dimensional Cartesian space. Denote by D the set of all points in the pN-dimensional space which we are free to choose. If N is fixed and if any point of D can be equally well chosen, the following two definitions seem to be appropriate:

DEFINITION 1. *Denote by c the minimum value of the determinant $|c_{lm}|$ $(l, m = 1, \cdots, r)$ in the domain D. Then the ratio $c/|c_{lm}|$ is called the efficiency of the design of the statistical investigation for testing the hypothesis* (3).

DEFINITION 2. *The design of the statistical investigation for testing the hypothesis* (3) *is said to be most efficient if its efficiency is equal to* 1.

3. Efficiency of the Latin square design. A widely used and important design in agricultural experimentation is the so-called Latin square. Suppose we wish to find out by experimentation whether there is any significant difference among the yields of m different varieties $v_1, \cdots, v_m$. For this purpose the experimental area is subdivided into m^2 plots lying in m rows and m columns and each plot is assigned to one of the varieties $v_1, \cdots, v_m$. If each variety appears exactly once in each row and exactly once in each column, we have a Latin square arrangement. Denote by y_{ijk} the yield of the variety v_k on the plot which lies in the i-th row and j-th column. The subscript k is, of course, a single valued function of the subscripts i and j, since to each plot only one variety is assigned. The following assumptions are made: the variates y_{ijk} are independently and normally distributed with a common variance σ^2 and the expected value of y_{ijk} is given by

$$E(y_{ijk}) = \mu_i + \nu_j + \rho_k. \tag{10}$$

The parameters σ^2, μ_i, ν_j and ρ_k are unknown. The hypothesis to be tested is the hypothesis that variety has no effect on yield, i.e.

$$\rho_1 = \rho_2 = \cdots = \rho_k. \tag{11}$$

We associate the positive integer $\alpha(i, j) = (i - 1)m + j$ with the plot which lies in the i-th row and j-th column. $(i, j = 1, \cdots, m)$. It is clear that for any positive integer $\alpha \leq m^2$ there exists exactly one plot, i.e. exactly one pair of values i and j, such that $\alpha = \alpha(i, j)$. In the following discussions the symbol y_α $(\alpha = 1, \cdots, m^2)$ will denote the yield y_{ijk} where the indices i and j are determined so that $\alpha(i, j) = \alpha$. The plot in the i-th row and j-th column will be called the α-th plot where $\alpha = \alpha(i, j)$.

We define the symbols $t_{i\alpha}$, $u_{j\alpha}$, $z_{k\alpha}$ $(i, j, k = 1, \cdots, m; \alpha = 1, \cdots, m^2)$, as follows: $t_{i\alpha} = 1$ if the α-th plot lies in the i-th row, and $t_{i\alpha} = 0$ otherwise. Similarly $u_{j\alpha} = 1$ if the α-th plot lies in the j-th column, and $u_{j\alpha} = 0$ otherwise. Finally $z_{k\alpha} = 1$ if the k-th variety is assigned to the α-th plot, and $z_{k\alpha} = 0$ otherwise. Then equation (10) can be written as

$$\begin{aligned} E(y_\alpha) = {} & \mu_1 t_{1\alpha} + \cdots + \mu_m t_{m\alpha} + \nu_1 u_{1\alpha} + \cdots \\ & + \nu_m u_{m\alpha} + \rho_1 z_{1\alpha} + \cdots + \rho_m z_{m\alpha}. \end{aligned} \tag{12}$$

Denote the arithmetic means $\frac{1}{m^2}\sum_{\alpha=1}^{m^2} t_{i\alpha}$, $\frac{1}{m^2}\sum_{\alpha=1}^{m^2} u_{i\alpha}$, and $\frac{1}{m^2}\sum_{\alpha=1}^{m^2} z_{i\alpha}$ by $\bar{t}_i$, $\bar{u}_i$ and $\bar{z}_i$ respectively. Let $t'_{i\alpha} = t_{i\alpha} - \bar{t}_i$, $u'_{i\alpha} = u_{i\alpha} - \bar{u}_i$, $z'_{i\alpha} = z_{i\alpha} - \bar{z}_i$, $\mu'_i = \mu_i - \mu_m$, $\nu'_i = \nu_i - \nu_m$ and $\rho'_i = \rho_i - \rho_m$ for $i = 1, \cdots, m - 1$. Let furthermore $w_\alpha = 1$ for $\alpha = 1, \cdots, m^2$. Then we have

$$\begin{cases} t_{i\alpha} = t'_{i\alpha} + \bar{t}_i w_\alpha; \quad u_{i\alpha} = u'_{i\alpha} + \bar{u}_i w_\alpha; \quad z_{i\alpha} = z'_{i\alpha} + \bar{z}_i w_\alpha; \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (i = 1, \cdots, m - 1) \\ t_{m\alpha} = (1 - \bar{t}_1 - \cdots - \bar{t}_{m-1})w_\alpha - t'_{1\alpha} - \cdots - t'_{m-1,\alpha}, \\ u_{m\alpha} = (1 - \bar{u}_1 - \cdots - \bar{u}_{m-1})w_\alpha - u'_{1\alpha} - \cdots - u'_{m-1,\alpha}, \\ z_{m\alpha} = (1 - \bar{z}_1 - \cdots - \bar{z}_{m-1})w_\alpha - z'_{1\alpha} - \cdots - z'_{m-1,\alpha}. \end{cases} \tag{13}$$

From (12) and (13) we obtain

$$E(y_\alpha) = \xi w_\alpha + \sum_{i=1}^{m-1} \mu_i' t_{i\alpha}' + \sum_{i=1}^{m-1} \nu_i' u_{i\alpha}' + \sum_{i=1}^{m-1} \rho_i' z_{i\alpha}' \tag{14}$$

where

$$\xi = \sum_{i=1}^{m-1} \mu_i' \bar{t}_i + \sum_{i=1}^{m-1} \nu_i' \bar{u}_i + \sum_{i=1}^{m-1} \rho_i' \bar{z}_i + \mu_m + \nu_m + \rho_m .$$

The hypothesis (11) can be written as

$$\rho_1' = \rho_2' = \cdots = \rho_{m-1}' = 0. \tag{15}$$

This is a linear hypothesis in canonical form as given in (3). The values $z_{i\alpha}'$ ($i = 1, \cdots, m-1$; $\alpha = 1, \cdots, m^2$) depend on the way in which the varieties $v_1, \cdots, v_m$ are assigned to the m^2 plots. We will show that we obtain a most efficient design if we distribute the varieties over the m^2 plots in a Latin square arrangement, i.e. if each variety appears exactly once in each row and exactly once in each column.

Let $q_{1\alpha} = w_\alpha$, $q_{i+1,\alpha} = t_{i\alpha}'$ ($i = 1, \cdots, m-1$), $q_{m+j,\alpha} = u_{j\alpha}'$ ($j = 1, \cdots, m-1$) and $q_{2m-1+k,\alpha} = z_{k\alpha}'$ ($k = 1, \cdots, m-1$). Denote $\sum_{\alpha=1}^{m^2} q_{i\alpha} q_{j\alpha}$ by a_{ij} ($i, j = 1, 2, \cdots, 3m-2$) and let the matrix $\| c_{ij} \|$ be the inverse of the matrix $\| a_{ij} \|$ ($i, j = 1, \cdots, 3m-2$). Let us denote by Δ the determinant $| a_{ij} |$ ($i, j = 1, \cdots, 3m-2$), by Δ_1 the determinant $| a_{ij} |$ ($i, j = 1, \cdots, 2m-1$), by Δ_2 the determinant $| a_{ij} |$ ($i, j = 2m, \cdots, 3m-2$) and Δ_2' the determinant $| c_{ij} |$ ($i, j = 2m, \cdots, 3m-2$). We have to show that for the Latin square arrangement Δ_2' becomes a minimum. From a known theorem[3] about determinants it follows that

$$\Delta_2' = \Delta_1/\Delta. \tag{16}$$

Hence, we have merely to show that Δ/Δ_1 becomes a maximum for the Latin square arrangement. Denote by $\bar{\Delta}$, $\bar{\Delta}_1$ and $\bar{\Delta}_2$ the values taken by Δ, Δ_1 and Δ_2, respectively, in the case of a Latin square arrangement. Since, for the Latin square arrangement, as is known,

$$\sum_{\alpha=1}^{m^2} z_{k\alpha}' u_{j\alpha}' = \sum_{\alpha=1}^{m^2} z_{k\alpha}' t_{i\alpha}' = \sum_{\alpha=1}^{m^2} z_{k\alpha}' w_\alpha = 0 \qquad (i, j, k = 1, \cdots, m-1)$$

we have

$$\frac{\bar{\Delta}}{\bar{\Delta}_1} = \bar{\Delta}_2 . \tag{17}$$

Since the matrix $\| a_{ij} \|$ ($i, j = 1, \cdots, 3m-2$) is positive definite we have

$$\frac{\Delta}{\Delta_1} \leqq \Delta_2 . \tag{18}$$

[3] See M. Bôcher, *Introduction to Higher Algebra*, 1931, pp. 31.

Because of (17) and (18) the Latin square design is proved to be most efficient if we show that $\Delta_2 \leq \bar{\Delta}_2$.

Denote by Δ_2^* the m-rowed determinant $|a_{ij}|$ $(i, j = 1, 2m, 2m + 1, \cdots, 3m - 2)$. Since $a_{1j} = 0$ for $j \neq 1$, we have

$$\Delta_2^* = a_{11}\Delta_2 = m^2\Delta_2. \tag{19}$$

Denote $\sum_{\alpha=1}^{m^2} z_{i\alpha}z_{j\alpha}$ by b_{ij} $(i, j = 1, \cdots, m)$. Then

$$\begin{cases} b_{ij} = 0, & \text{for } i \neq j \\ \text{and} \quad b_{ii} = N_i, \end{cases} \tag{20}$$

where N_i denotes the number of plots to which the variety v_i has been assigned. Because of (20) we have

$$\begin{vmatrix} b_{11} & \cdots & b_{1m} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ b_{m1} & & b_{mm} \end{vmatrix} = N_1N_2 \cdots N_m. \tag{21}$$

According to (13) we have

$$\begin{aligned} z'_{i\alpha} + \bar{z}_i w_\alpha &= z_{i\alpha}, \qquad (i = 1, \cdots, m - 1) \\ -z'_{1\alpha} - \cdots - z'_{m-1,\alpha} + w_\alpha(1 - \bar{z}_1 - \cdots - \bar{z}_{m-1}) &= z_{m\alpha}. \end{aligned} \tag{22}$$

The determinant of these equations is given by

$$\lambda = \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \bar{z}_1 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \bar{z}_2 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 0 & 1 & \bar{z}_{m-1} \\ -1 & -1 & -1 & \cdots & -1 & -1 & \delta \end{vmatrix} \tag{23}$$

where $\delta = 1 - \bar{z}_1 - \bar{z}_2 - \cdots - \bar{z}_{m-1}$. It is easy to verify that

$$\lambda = 1. \tag{24}$$

From (21), (22) and (24) it follows that

$$\Delta_2^* = N_1N_2 \cdots N_m. \tag{25}$$

Hence, from (19) we obtain

$$\Delta_2 = N_1N_2 \cdots N_m/m^2. \tag{26}$$

In the case of a Latin square design we have $N_1 = N_2 = \cdots = N_m = m$. Hence

$$\bar{\Delta}_2 = m^{m-2}. \tag{27}$$

Because of the condition $N_1 + N_2 + \cdots + N_m = m^2$, the right hand side of (26) becomes a maximum when $N_1 = N_2 = \cdots = N_m = m$. Thus $\Delta_2 \leq \bar{\Delta}_2$ and consequently the Latin square design is proved to be most efficient.

4. Efficiency of Graeco-Latin and higher squares. Consider m varieties $v_1, \cdots, v_m$ and m treatments $q_1, \cdots, q_m$ Suppose that we wish to find out by experimentation whether the yield is affected by varieties or treatments. For this purpose the experimental area is subdivided into m^2 plots lying in m rows and m columns and to each plot one of the varieties and one of the treatments is assigned. We call this arrangement a Graeco-Latin square if the following conditions are fulfilled: 1) each variety appears exactly once in each row and exactly once in each column; 2) each treatment appears exactly once in each row and exactly once in each column; 3) each variety is combined with each of the treatments exactly once.

The following general abstract scheme includes the Latin square and Graeco-Latin square as special cases: Consider an r-way classification with m classes in each classification. Denote by $y_{a_1 a_2 \cdots a_r}$ the value of a certain characteristic of an individual who is classified in the a_1-class of the first classification, in the a_2-class of the second classification, $\cdots$, and in the a_r-class of the r-th classification. Suppose that m^2 observations are made for the purpose of investigating the effect of the classes on the value of the characteristic under consideration. We will say that we have a generalized Latin square design if the following condition is fulfilled: *Let i, j, m' and m'' be an arbitrary set of four positive integers for which $i \neq j$, $i \leq r$, $j \leq r$, $m' \leq m$ and $m'' \leq m$. Then among the m^2 individuals observed there exists exactly one individual who belongs to the m'-class of the i-th classification and m''-class of the j-th classification.*

It is clear that if $r = 3$ the above scheme is a Latin square. If $r = 4$ we have a Graeco-Latin square.

Assume that the observations $y_{a_1 \cdots a_r} (a_1, a_2, \cdots, a_r = 1, \cdots, m)$ are normally and independently distributed with a common variance σ^2. Assume furthermore that the expected value of $y_{a_1 \cdots a_r}$ is given by

$$E(y_{a_1 a_2 \cdots a_r}) = \gamma_{1a_1} + \cdots + \gamma_{ra_r} .$$

The parameters σ^2 and γ_{ia} $(i = 1, \cdots, r;\ a = 1, \cdots, m)$ are unknown constants. Suppose that we wish to test the hypothesis that

$$\gamma_{i1} = \gamma_{i2} = \cdots = \gamma_{im} . \tag{28}$$

It can be shown that if the number of observations is limited to m^2, we obtain a most efficient design by constructing a generalized Latin square. The proof of this statement is similar to that of the efficiency of the Latin square and is therefore omitted.