UNIVERSITY OF NORTH CAROLINA
Department of Statistics
Chapel Hill, N. C.

USE OF CONCOMITANT MEASUREMENTS IN DESIGN AND ANALYSIS OF EXPERIMENTS

by

P. K. Bhattacharya

November, 1961

Contract No. AF 49(638)-213

The problem of using a concomitant variable (known beforehand) to design an experiment and to analyze the outcome for estimating the differences between several treatment effects has been considered and an optimum procedure has been obtained in a certain class. The investigation has been guided by Fisher's principle of increasing precision within the framework of valid procedures.

This research was partially supported by the Mathematics Division of the Air Force Office of Scientific Research.

Institute of Statistics
Mimeo Series No. 309

USE OF CONCOMITANT MEASUREMENTS IN DESIGN AND ANALYSIS OF EXPERIMENTS.

P. K. Bhattacharya[1]
Indian Statistical Institute and University of North Carolina

1. Introduction.

While using concomitant measurements in the analysis of treatment effects on the final responses, there is always the danger that the assumed nature of dependence between the final responses and the concomitant measurements may only be an approximation to the true one or may even be utterly wrong. In order to be able to make any valid inference about the treatment effects on the basis of the experimental data irrespective of whether the assumed nature of dependence is true or not, we must introduce a proper randomization in designing the experiment. Such a randomization is only a necessary step in solving the actual inference problem at hand, but there still remain other questions to be settled. First, for any given method of proper randomization, we have to find a final decision rule which should be valid even if the assumed nature of dependence is not true, and secondly, we have to find a method of randomization among all methods of proper randomization . In settling both these problems, we shall try to increase the precision of experimentation and decision under the assumed nature of dependence.

In this note, the problem of using a concomitant variable (known beforehand) to design an experiment and to analyze the outcome for estimating the differences between several treatment effects has been considered and

---

[1]This research was partially supported by the Air Force Office of Scientific Research.

an optimum procedure has been obtained in a certain class. The guiding principle is the one mentioned above, which is no other than Fisher's [2] principle of increasing precision within the framework of valid procedures.

2. Formulation of the problem.

Suppose there are v treatments $T_1, \ldots , T_v$, vr experimental units $U_1, \ldots , U_{vr}$ and let y_{ij}(real) denote the observation arising out of the application of T_j on U_i. Suppose further that there exist $v + vr$ quantities $\Theta_1, \ldots , \Theta_v$ and $a_1, \ldots , a_{vr}$ such that

(1) $$y_{ij} = a_i + \Theta_j , \qquad 1 \leq i \leq vr , \quad 1 \leq j \leq v .$$

These quantities $\Theta_1, \ldots , \Theta_v$, $a_1, \ldots , a_{vr}$ are all unknown and we want to estimate

$$\delta_{jj'} = \Theta_j - \Theta_{j'}$$

for all $j \neq j'$.

For each i, exactly one of the quantities $y_{i1}, \ldots , y_{iv}$ can be observed and we put the further restriction of observing exactly r of the quantities $y_{1j}, \ldots , y_{vr, j}$. In all, there are $M = (vr)!/(r!)^v$ such observational schemes of which a typical one, say the s-th, can be described as a partitioning of the set of integers $\{1, \ldots , vr\}$ into v disjoint sets $G_1(s), \ldots , G_v(s)$, each containing r integers such that y_{ij} is observed in the s-th scheme if and only if $i \in G_j(s)$.

Our problem is to choose a design of experiments P which is a probability distribution on the set of integers $\{1, \ldots , M\}$, on which an observation will be made and if that happens to be s, then observations will be made on y_{ij}'s according to the s-th scheme, and to choose an estimation rule T which defines for each $s = 1, \ldots , M$ and for each

$j \neq j' = 1, \ldots, v$, functions $T_{jj'|s}$ of the y_{ij}'s observed in the s-th scheme, whose values will be used for estimating $\delta_{jj'}$. We restrict our selection of (P, T) by the following conditions:--

(i) Each of the functions $T_{jj'|s}$ must be a linear function of y_{ij}'s observed in the s-th scheme.

(ii) For each $j \neq j' = 1, \ldots, v$ and for each $\Theta_1, \ldots, \Theta_v$, $a_1, \ldots, a_{vr}$ for which (1) holds,

$$\sum_{s=1}^{M} P(s) \; T_{jj'|s} = \Theta_j - \Theta_{j'}$$

must be satisfied.

(iii) If further, $a_1, \ldots, a_{vr}$ are stochastically related to known constants $x_1, \ldots, x_{vr}$ in the following way,

$$a_i = \beta \, x_i + \epsilon_i \quad , \tag{2}$$

where β is an unknown constant and $\epsilon_1, \ldots, \epsilon_{vr}$ are uncorrelated randon variables with common mean zero and a common unknown variance σ^2, then the average of the variances of the estimators of all $\delta_{jj'}$, which we denote by $V_p(T)$, must be a bounded function of $\beta, \Theta_1, \ldots, \Theta_k$ for any given σ^2, all expectations being taken with respect to $\epsilon_1, \ldots, \epsilon_{vr}$ as well as with respect to the distribution P over the observational schemes.

Let $\mathcal{C}$ denote the class of design and estimation rules (P, T) which satisfy conditions (i), (ii) and (iii). We want to find $(P^*, T^*) \epsilon \mathcal{C}$ such that

$$V_{P*}(T^*) \leq V_P(T)$$

for all $\beta, \Theta_1, \ldots, \Theta_k, \sigma^2$ and for arbitrary $(P, T) \; \epsilon \; \mathcal{C}$.

Consider the following condition:--

(ii)' If (1) holds for some $\Theta_1, \ldots, \Theta_v, a_1, \ldots, a_{vr}$ where the a_i's are random variables satisfying (2), then

$$\sum_{s=1}^{M} P(s) \; E\, T_{jj',\,s} = \Theta_j - \Theta_{j'}$$

for each $j \neq j' = 1, \ldots, v$ and for each $\Theta_1, \ldots, \Theta_v, \beta$, where E denotes expectation with respect to random variables $\epsilon_1, \ldots, \epsilon_{vr}$ occurring in (2).

Let $\mathcal{C}'$ denote the class of design and estimation rules (P, T) which satisfy conditions (i), (ii)' and (iii) .

Obviously, $\mathcal{C} \subset \mathcal{C}'$.

3. Optimum (P, T) in $\mathcal{C}'$ when P is fixed.

We shall first restrict our attention to a fixed P for which there exists some $(P, T) \in \mathcal{C}'$ to show that (P, T*) is optimum in $\mathcal{C}'$ where T* is the least squares estimation rule. For this we require the following lemmas.

Lemma 1. For each $s = 1, \ldots, M$,

$$\underset{(n(s)\times 1)}{y(s)} = \underset{(n(s)\times k)}{A(s)} \; \underset{(k\times 1)}{\Theta} + \underset{(n(s)\times 1)}{\epsilon(s)} \qquad (3)$$

where $\epsilon(s)' = (\epsilon_1(s), \ldots, \epsilon_{n(s)}(s))$ is a random vector whose elements are mutually uncorrelated, each with mean zero and a common unknown variance σ^2, A(s) is a known matrix and $\Theta' = (\Theta_1, \ldots, \Theta_k)$ is an unknown vector. P(s) is a known probability distribution over the set of integers $\{1, \ldots, M\}$ on which one observation is made and if that happens to be s, then an observation is made on the random vector y(s). If $c_1, \ldots, c_u$ are known k-dimensional vectors and $\ell_i(s)$, $i = 1, \ldots, u$; $s = 1, \ldots, M$, are n(s) - dimensional vectors satisfying

$$\sum_{s=1}^{M} P(s) \; E \; \ell_i(s)' \; y(s) = c_i' \; \Theta \; , \quad i = 1, \ldots , u \; ,$$

then a necessary condition for

$$\sum_{i=1}^{u} \sum_{s=1}^{M} P(s) \; E \left[\ell_i(s)' \; y(s) - c_i' \; \Theta \right]^2$$

to be a bounded function of $\underline{\Theta}$ is that

$$E \; \ell_i(s)' \; y(s) = c_i' \; \Theta \; , \qquad i = 1, \ldots , u \; .$$

for each s with $P(s) > 0$. (E denotes expectation over $\epsilon_1(s), \ldots ,$ $\epsilon_{n(s)}(s))$.

<u>Proof</u>: Suppose for some $i = i_o$ and for some $s = s_o$ with $P(s_o) > 0$,

$$E \; \ell_{i_o}(s_o)' \; y(s_o) = c_{i_o}' \; \Theta + h_{i_o}(s_o)' \; \Theta \; ,$$

where at least one co-ordinate of the vector $h_{i_o}(s_o)$ is non-zero. Let that be the j_o-th co-ordinate of $h_{i_o}(s_o)$ and denote it by $h_{i_o j_o}(s_o)$. Then,

$$\sum_{i=1}^{u} \sum_{s=1}^{M} P(s) \; E\left[\ell_i(s)' \; y(s) - c_i' \; \underline{\Theta} \right]^2$$

$$\geq P(s_o) \; E \left[\ell_{i_o}(s_o)' \; y(s_o) - c_{i_o}' \; \Theta \right]^2$$

$$= P(s_o) \left[h_{i_o}(s_o)' \; \Theta \right]^2$$

$$= P(s_o) \; \{ h_{i_o j_o} (s_o) \}^2 \; \Theta_{j_o}^2$$

for all $\Theta' \in A_{j_o}$ where

$$A_{j_o} = \{ (\Theta_1, \ldots , \Theta_k) : \; \Theta_j = 0 \text{ for all } j \neq j_o \} \; .$$

Therefore, for any arbitrary B, $\sum_{i=1}^{u} \sum_{s=1}^{M} P(s) \; E\left[\ell_i(s)' \; y(s) - c_i' \; \Theta \right]^2$

can be made larger than B by choosing $\Theta' \in A_{j_o}$ with

$$|\Theta_{j_o}| > B/\sqrt{P(s_o)} \; | h_{i_o j_o}(s_o) | ,$$

and that completes the proof.

Lemma 2. In the set-up of Lemma 1, if there exists an estimate T of $(c_1' \Theta , \ldots , c_u' \Theta)$ of the form $T_s = (T_{1|s}, \ldots , T_{u|s}) = (\ell_1(s)' y(s), \ldots , \ell_u(s)' y(s))$, $s = 1, \ldots , M$ which satisfy

$$\sum_{s=1}^{M} P(s) \; E \; \ell_i(s)' \; y(s) = c_i' \Theta , \quad i = 1, \ldots , u ,$$

and for which

$$V_P(T) = \frac{1}{u} \sum_{i=1}^{u} \sum_{s=1}^{M} P(s) \; E \left[T_{i|s} - c_i' \; \Theta \right]^2$$

is bounded, then among all such estimates, $V_P(T)$ is minimized for

$$T_s^* = (T_{1|s}^*, \ldots , T_{u|s}^*) , \quad s = 1, \ldots , M$$

where $T_{i|s}^*$ is the least squares estimate of $c_i' \Theta$ in the linear estimation set-up in the s-th observational scheme given in (3).

Proof. By virtue of Lemma 1, all linear unbiased estimates T for which $V_P(T)$ is bounded, must satisfy

$$E \; T_{i|s} = c_i' \; \Theta , \qquad i = 1, \ldots , u ,$$

for each s with $P(s) > 0$, and it is immaterial how T_s is defined for those s which have probability zero under P. An application of Markoff's theorem on least squares estimates now completes the proof.

Now coming back to our problem let us define an estimation rule T^* such that for each s and for each $j \neq j' = 1, \ldots , v$,

$$(4) \quad T^*_{jj'|s} = [\bar{y}_j(s) - \bar{y}_{j'}(s)] - b(s)[\bar{x}_j(s) - \bar{x}_{j'}(s)]$$

where $\bar{x}_j(s) = \sum_{i \in G_j(s)} x_i/r$, $\bar{y}_j(s) = \sum_{i \in G_j(s)} y_{ij}/r$, $j = 1, \ldots, v$;

If

$$W_{xx}(s) = \sum_{j=1}^{v} \sum_{i \in G_j(s)} [x_i - \bar{x}_j(s)]^2 ,$$

$$W_{yx}(s) = \sum_{j=1}^{v} \sum_{i \in G_j(s)} y_{ij}[x_i - \bar{x}_j(s)] ,$$

and

$$b(s) = \begin{cases} W_{yx}(s)/W_{xx}(s) & \text{if } W_{xx}(s) > 0 \\ 0 & \text{if } W_{xx}(s) = 0 . \end{cases}$$

If $x_1 = \ldots = x_{vr}$, $W_{xx}(s) = 0$ and hence $b(s) = 0$ for each s and $T^*_{jj'|s} = \bar{y}_j(s) - \bar{y}_{j'}(s)$ is the least squares estimate of $\theta_j - \theta_{j'}$ under the s-th observational scheme. Also in this case $V_P(T^*)$ is bounded for each P.

Excluding the case $x_1 = \ldots = x_{vr}$ we see that $W_{xx}(s) > 0$ for some s and in the observational scheme for each of these s, $T^*_{jj'|s}$ is the least squares estimate of $\theta_j - \theta_{j'}$, whereas if for some s, $W_{xx}(s) = 0$ and if there is a design P with positive probability on any of these s, then it can be easily seen by using lemma 1 that such a design with every accompanying estimation rule is outside the class $\mathcal{C}'$ and therefore need not be taken into consideration. For any other P, $V_P(T^*)$ is bounded.

An application of lemma 2 now gives

<u>Theorem 1</u>. If for a fixed design P_o there exists some $(P_o, T) \in \mathcal{C}'$ and if T^* is as defined in (4), then $(P_o, T^*) \in \mathcal{C}'$ and for any T such that $(P_o, T) \in \mathcal{C}'$,

$$V_{P_o}(T) \geq V_{P_o}(T^*) .$$

The above theorem makes our search for an optimum (P , T) in $\mathcal{C}'$ much easier. What we have to do now is to consider all P for which $V_P(T^*)$ is bounded and then choose the one among them for which $V_P(T^*)$ is smallest.

It is easy to see that for all P having $V_P(T^*)$ bounded, the following expression holds.

$$V_P(T^*) = \begin{cases} \frac{2\sigma^2}{r} & \text{if } x_1 = \dots = x_{vr} \\ \frac{2\sigma^2}{r} \left[1 + \frac{1}{v-1} \sum_{s=1}^{M} P(s) \cdot \frac{T_{xx} - W_{xx}(s)}{W_{xx}(s)} \right] & \text{otherwise} \end{cases}$$

where $T_{xx} = \sum_{i=1}^{vr} x_i^2 - (\sum_{i=1}^{vr} x_i)^2/vr$.

4. <u>Optimum (P, T) in $\mathcal{C}$.</u>

We shall now find the optimum (P, T) in $\mathcal{C}'$ and will show that this optimum in $\mathcal{C}'$ is actually contained in $\mathcal{C}$ and since $\mathcal{C} \subset \mathcal{C}'$, it will prove that this is optimum in $\mathcal{C}$.

Let s^* be such that

(5) $W_{xx}(s^*) \geq W_{xx}(s)$, $s = 1, \dots , M$,

and let S^* be the set of all permutations of s^* i.e. the set of all s' such that for each $j = 1, \dots , v$ we can find some j' (depending on s') such that $G_j(s^*) = G_{j'}(s')$. Obviously, $W_{xx}(s) = W_{xx}(s^*)$ for all $s \in S^*$. (If $W_{xx}(s)$ attains its minimum value for more than one s which are not permutations of each other in the above sense, then we shall choose any one of them, call it s^* and find the set S^* of all its permutations.)

Define a design P^* as follows ,

(6) $$P^*(s) = \begin{cases} 1/v! & \text{if } s \in S^* \\ 0 & \text{otherwise.} \end{cases}$$

Theorems 2 and 3 will establish that (P^*, T^*) is the desired design and estimation rule.

Theorem 2. For all $(P, T) \in \mathcal{C}'$,

$$V_P(T) \geq V_{P^*}(T^*)$$

where P^* is defined in (6) and T^* is defined in (4).

Proof. Since $(P, T) \in \mathcal{C}'$,

$$V_P(T) \geq V_P(T^*) = \frac{2\sigma^2}{r} = V_{P^*}(T^*) \quad \text{if} \quad x_1 = \dots = x_{vr} .$$

Otherwise, $V_P(T) \geq V_P(T^*)$

$$= \frac{2\sigma^2}{r} \left[1 + \frac{1}{v-1} \cdot \sum_{s=1}^{M} P(s) \cdot \frac{T_{xx} - W_{xx}(s)}{W_{xx}(s)}\right]$$

$$\geq \frac{2\sigma^2}{r} \left[1 + \frac{1}{v-1} \cdot \frac{T_{xx} - W_{xx}(s^*)}{W_{xx}(s^*)}\right] \quad \text{by (5)}$$

$$= V_{P^*}(T^*) .$$

Theorem 3. $(P^*, T^*) \in \mathcal{C}$ and for all $(P, T) \in \mathcal{C}$,

$$V_P(T) \geq V_{P^*}(T^*) .$$

Proof. Since $\mathcal{C} \subset \mathcal{C}'$, the latter part of the theorem follows from Theorem 2. To show that $(P^*, T^*) \in \mathcal{C}$, we note that since T^* obviously satisfies condition (i) and since $V_{P^*}(T^*)$ is always a finite quantity not depending on $\beta, \Theta_1, \dots, \Theta_v$, it will be enough to verify condition (ii). For any $j \neq j' = 1, \dots, v$,

$$\sum_{s=1}^{M} P^*(s) \; T^*_{jj'}|s$$

$$= \begin{cases} \frac{1}{v!} \sum_{s \epsilon S^*} [\, \bar{y}_j(s) - \bar{y}_{j'}(s) \,] & \text{if } x_1 = \dots = x_{vr} \\ \frac{1}{v!} \sum_{s \epsilon S^*} [\, \{ \bar{y}_j(s) - \bar{y}_{j'}(s) \} - \frac{W_{yx}(s)}{W_{xx}(s)} \{ \bar{x}_j(s) - \bar{x}_{j'}(s) \} \,] & \text{otherwise} \end{cases}$$

$$= \begin{cases} (\Theta_j - \Theta_{j'}) + \frac{1}{v!} \sum_{s \epsilon S^*} [\, \bar{a}_j(s) - \bar{a}_{j'}(s) \,] & \text{if } x_1 = \dots = x_{vr} \\ (\Theta_j - \Theta_{j'}) + \frac{1}{v!} \sum_{s \epsilon S^*} [\, \bar{a}_j(s) - \bar{a}_{j'}(s) \,] \\ \quad - \frac{\sum_{i=1}^{vr} a_i x_i - r \sum_{j=1}^{v} \bar{a}_j(s^*) \; \bar{x}_j(s^*)}{W_{xx}(s^*) \,.\, v\,!} \,.\, \sum_{s \epsilon S^*} [\, \bar{x}_j(s) - \bar{x}_{j'}(s) \,] \end{cases}$$

$$= \Theta_j - \Theta_{j'} \; ,$$

and that completes the proof.

Remarks. 1) It is doubtful whether the design P* is unbiased in the sense of Yates [3] and the author is inclined to believe that it is not though he could not prove it because even though the expected value of the adjusted residual mean square under randomization becomes involved, there still remains the possibility of getting an unbiased estimator of the variance of the estimation rule T* under (1) and under randomization according to P*, in some other way. We also have to keep in mind the further possibility that there may be a different estimation rule which is unbiased under (1) and under P* and the variance of which can be estimated in the desired manner. However, if we restrict ourselves to the traditional analysis, then Cox's [1] argument shows that P* is not unbiased.

2) The result given in this note can be extended without any difficulty to the case where more than one concomitant measurements are available for each experimental unit.

The problem as formulated in section 2, arose in course of a discussion that the author had with J. Roy in the Indian Statistical Institute.

REFERENCES

[1] D. R. Cox (1956). A note on weighted randomization. Ann. Math. Stat. 27. pp. 1144-1151.

[2] R. A. Fisher (1949). Design of Experiments. Hafner Publishing Company Inc., New York. (Fifth edition) .

[3] F. Yates (1951). Bases logiques de la planification des expériences. Ann. Inst. H. Poincaré. 12. pp. 97-112.