MODEL EQUATIONS, TREATMENT POPULATION, AND MIXTURES

by

D. B. Hall and W. T. Federer
Biometrics Unit, Cornell University, Ithaca, New York 14853

BU-605-M March, 1977

Abstract

A discussion of a model or response equation, the treatment design, and the population of treatments is presented. Basically, three types of treatment populations are considered. The first is the intersection points of an n-dimensional lattice peculiar to a factorial treatment design and a discrete set of points. The second design corresponds to the continuous spaces in an n-dimensional space as is considered in regression and response surface treatment designs and populations. The third population and model equation considered is exemplified in the literature by the diallel crossing designs, which have been generalized to mixtures of k items rather than two as in the diallel cross. The population structure for this third type of treatment design is discussed and an example is presented to illustrate the three types of designs.

Introduction

An important interaction exists between the model equation used to describe a system and the information about the system gained through an experiment. When a linear or a nonlinear model is used to analyze data, it allows us to gain information about how different components of a treatment have contributed to the overall effect of the treatment. If the experiment is poorly designed components may be confounded. That is, it may be impossible to separate the contributions of the two components. If a model equation is poorly chosen, it may be impossible, or at least difficult, to make a meaningful interpretation of the experimental results.

The purpose of this paper is to discuss different approaches to finding a model equation and to consider the problem of treatment designs for mixtures. Since a model equation is an attempt to describe a system and an experiment is an attempt to gain information about a system, both must be constructed with the system in mind.

The Model Equation

A model equation is an approximate algebraic description of the effects of a treatment, where a treatment has been defined to be "a single entity, combination, or phenomenon under study in an experiment". The treatment design has been defined to be "the selection of treatments to be used in the experiment" (see Federer [1955]). The treatments are selected from some specified and precisely defined population of potential treatments, sometimes called the X-space, the factor-space or the space of interest to the experimenter. Such statistical properties as orthogonality, balance, and variance-optimality usually play some role in the selection of a treatment design. All of these statistical properties of designs are related to the model equation and the population of potential treatments.

Except in pathological cases and in some textbook examples the model equation embodies many assumptions. That is, the model equation is chosen on the basis of imperfect knowledge and is therefore an approximation to the true state of nature. The statistician tries to use this imperfect model to gain information about a population's characteristics. The model is being considered as an aid to understanding, not as a predictive tool.

The common approaches to choosing a model equation can be divided into four categories. The first category would contain approaches which begin with the simplest possible model equation and allow the data to dictate useful manipulations. The problem with such an approach is that it allows one to "explain" everything. A finite sample is taken, usually to estimate and ask questions about relationships in an infinite population. It is possible to construct several functions which will explain everything in a finite sample in terms of parameters and independent variables. Some inferences will be made concerning the population from the sample and from observed relationships which are unique characteristics of the sample chosen, which are not generalizable. The inference must remain suspect until another experiment, planned with the previously observed relationships in mind, contributes evidence of their existence. One example of an observed relationship which one would not expect to generalize is attributed to Professor G. Udny Yule. He observed a period of years and found that he could explain the divorce rate in England by considering the number of apples imported into Great Britain. "... in the years in which a large number of apples were imported into Great Britain, there were also a large number of divorces" (Fisher [1958]).

The second category would contain approaches which are primarily literature searches. The literature on related topics is scrutinized with an eye toward a paper dealing with problems similar to one's own and providing a model equation which might be borrowed. This approach has the pleasant operating characteristic that when one finds a model equation one also finds some precedent for that model equation. The main dangers with this approach category are that errors already appearing in the literature are perpetuated rather than corrected and subtle but important differences between experimental situations are ignored so that similar, but different, problems might be treated as if they were identical.

The third category would contain approaches which involve constructing the model equation. The experimenter begins with a list of factors whose levels are to be controlled, and a list of goals, questions, or objectives of the investigation. The model equation is constructed by designating a set of parameters to explain and describe the form of response. One of the main faults of such an approach is the propensity towards errors of omission; the list of factors and goals is not likely to be complete. Experimenters using approaches from this category are also susceptible to one of the errors of those who use the second category of approaches; they often consider only those factors and goals previous experimenters have considered.

The fourth category contains approaches which are primarily reductive. A very general model applicable to a wide range of subjects and problems is considered. The experimenter reduces the model by eliminating factors and parameters known to be nonexistent or negligible until the model is considered workable and reasonable. This procedure has the advantage that it requires a reason for every omission from the model. One needs to rely more on reasoning and less on memory and exhaustive library research. It is necessary to make the important assumption that the general model is "sufficiently" general.

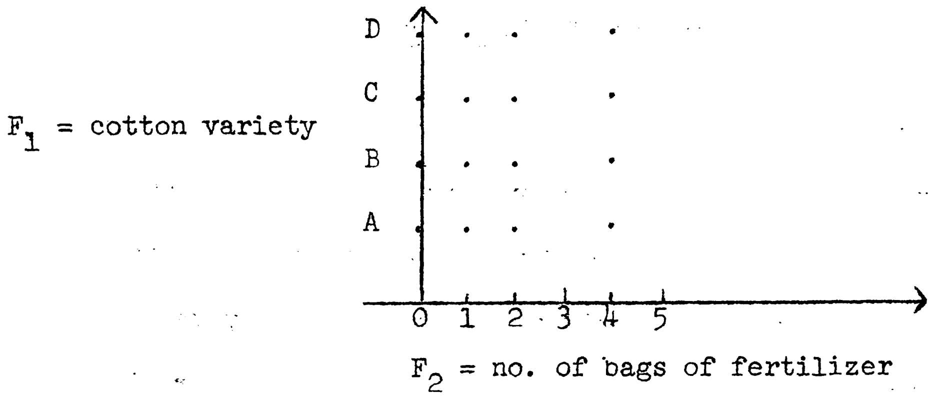
To operate within the second category someone must have preceded the experimenter and worked through a similar problem using an approach from one of the other categories. The differences between the other three categories are mostly questions of timing of construction. In the fourth category a model for a general problem is constructed before the specific problem has been formulated. The equation is for a general problem. In the third category a model is constructed after the specific problem has been formulated and before the data has been collected. The equation is for the specific problem. In the first category a model is constructed after the data has been collected. The equation is for a specific set of observations.

Treatment Designs for Mixtures

A large share of treatment designs utilized in experimentation involve combinations of levels of two or more factors or agents. A treatment is a mixture of two or more agents which are capable of assuming more than one value. For example, one agent might be rainfall in a week, which could have any value from zero to ten inches. Another agent might be sunlight which could be either direct or indirect. One response is obtained for each treatment unit, for each mixture. Factorial treatment designs, regression designs, genetic crossing designs, response surface designs, etc. are all designs involving mixtures of levels of two or more agents. Three general types of mixtures are distinguishable when one considers the population of possible mixtures, the factor space.

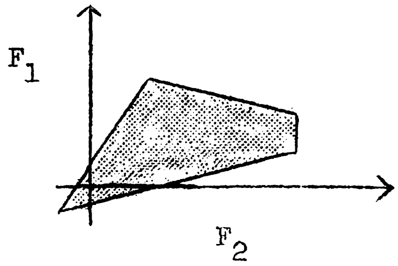
Factorial treatment designs involve all combinations of specified levels of two or more factors (see, e.g. Yates [1937]). The n-factor, F_1, F_2, $\cdots$, F_n, design has a finite number of levels for each factor. If k_i is the number of levels for the ith factor, the design consists of the $\prod_{i=1}^{n} k_i$ combinations or treatments. When this design is represented in n-space, the set of potential mixtures is an n-dimensional grid of lattice points or intersection points of the levels for each factor. In a complete factorial design all the points in the set of potential mixtures are used in the experiment. In a fractional factorial design the set of potential mixtures is the same but due to considerations of time, space, money, or special interests only the points in some subset are used in the experiment.

Let F_1 be cotton variety and let A, B, C, D be the particular four cotton varieties of interest. Let F_2 be fertilizer of brand X with the levels being number of bags (50 kilogram) of fertilizer per hectare with 0, 1, 2, and 4 bags as the levels of interest. The complete factorial arrangement of the 4 x 4 = 16 treatments could be represented graphically as follows:

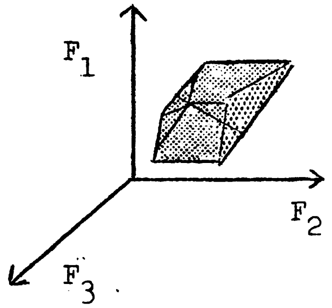


The points (dots) in the above diagram represent the population of potential treatments. Inferences are to and about this set of points. In designing for factorial experiments consideration is given to the existence of interrelationships, interactions, among the factors. If all the factors interact, making inferences requires at least one observation for each possible mixture. If none of the factors interact, inference can be based on as few as seven points, for example, (A,0), (A,1), (B,1), (B,2), (C,2), (C,4), and (D,4) . The model equation used in the last situation would be a general factorial design model equation reduced by the removal of interaction terms. Inferences would still apply to all the points in the factor space.

Response surface, multiple regression designs involve two or more agents, each with an uncountably infinite number of possible levels. If there are n agents in each mixture the set of potential mixtures can be represented graphically as an n-dimensional solid. An example of a factor space for n = 2 would be:



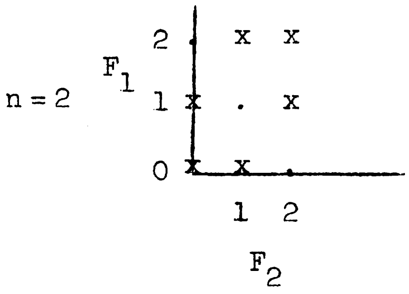
Any point within the shaded figure represents a potential treatment and experimentation aims at inference about the entire shaded figure. An example of a factor space for n = 3 should be:



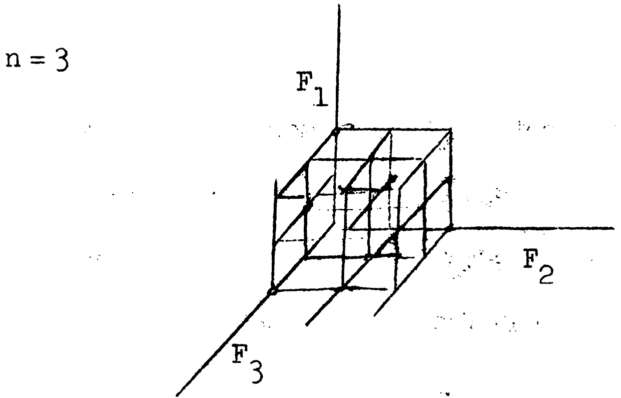
Any point within the shaded figure represents a potential treatment. A general model equation with infinitely many terms is reduced on the basis of assumptions, knowledge and pragmatic considerations to a model that allows for valid inferences about this infinite factor space on the basis of a finite set of observations.

When one designs for a response surface or regression problem a finite set of points must be chosen to be sampled from and to be observed. It is assumed that these points are known or are measured without error. These points may be considered as a subset of a suitably chosen factorial design. Inference is made to the populations from which the observed values are random samples, to the realized mixtures or the realized factorial. The inference is expressed in the context of a model which is designed to describe what can happen for any potential mixture.

Diallel crossing plans are the basis for designs which can be viewed as the mixture of two parents or parent strains as a treatment with potential progeny as the objects to which these genetic mixtures are applied. All possible crosses between n lines produce n(n-1)/2 crosses, all possible crosses plus reciprocal crosses produce n(n-1) crosses, all possible crosses plus n selfs produce n(n+1)/2 crosses, and all crosses, reciprocals, and self produce n^2 crosses. All of these crossing plans without reciprocals are subsets of a 3^n factorial arrangement as can be demonstrated graphically for n = 2 or 3:



The dot represents points in the diallel cross and the x's are the non-observable points of the 3^2 factorial.



The dots represent potential points in the diallel cross and the other grid points for the rest of the 3^3 factorial.

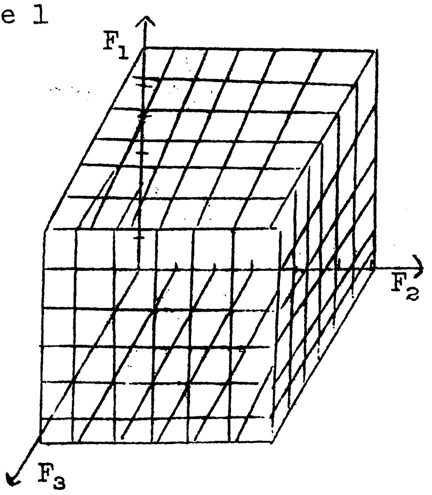
If L_{F_i} is the level of factor i, the subset of the factorial design is the intersection of the surface in n-space where $\sum_{i=1}^{n} L_{F_i} = 2$ and the points of the complete 3^n factorial. Inferences are to be made only about the points in the subset of the factorial and the model equation has meaning only at those points.

There is an extensive literature discussing designs, model equations and analysis for diallel crossing plans (see Randall [1975], Griffing [1956], Kempthorne [1957]). Recently Federer [1975] has proposed a generalization of the diallel cross concepts of general and specific combining ability for use when the set of potential treatments is a proper subset of a complete factorial arrangement. Hall [1976] developed a general model equation for this situation with emphasis on designing when there are restrictions that $\sum_{i=1}^{n} L_{F_i} = K$ $(K > 0)$ and if $L_{F_i}, L_{F_j} \neq 0$ then $L_{F_i} = L_{F_j}$. The concept of general and specific combining abilities was extended and relabeled as general and specific mixing abilities.

An example:

Consider an experiment testing three different fertilizers for effect. It is possible to apply up to six bags of each variety to each field. Two response surface type arrangements might be of interest. Any combination of fertilizers using whole bags and fractions of bags might be the treatment population of interest or any combination with the restriction that a total of exactly six bags full of fertilizer are to be applied might be the interesting treatment population. If the former arrangement was amended with a restriction that only whole bags could be used the arrangement would be a complete factorial. If the latter arrangement was amended with restrictions to whole bags and equal quantities from each variety applied, the arrangement would be one of the generalized genetic type discussed by Hall (1976). Figures 1 and 2 depict these treatment populations graphically.

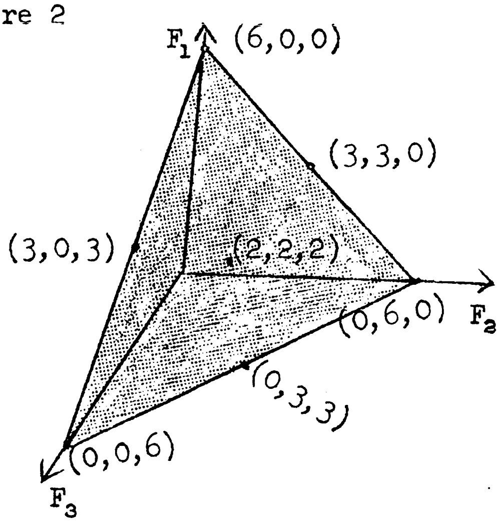
Figure 1



r_1 ≡ the first response surface
≡ the entire 6^3 cube.

f ≡ the factorial
≡ the 6^3 grid points of the cube.

Figure 2



r_2 ≡ the second response surface
≡ the shaded triangle.

g ≡ the generalized genetic
≡ the 7 dots.

The factorial design has been widely studied (see Federer and Balaam [1972] for list of references). Designs for response surface type populations have been reported and reviewed (see Mead and Pike [1975] and Cornell [1973] for review articles). The purpose of this paper has been presentation of a third class of mixture populations which has been slighted in statistical literature except in the case of diallel cross designs. The population structure and the treatment design problem are different for the three classes of designs discussed. Clear thinking and formulation is required in comprehending their differences and hence in the solution of statistical design and analysis problems for the three types.

References

1. Cornell, J. A. (1973). Experiments with mixtures: a review. Technometrics 15: 437-455.

2. Federer, W. T. (1955). Experimental Design. The Macmillan Company, New York.

3. Federer, W. T. (1975). Statistical designs for mixtures of crops and other applications. Paper No. BU-560-M in the Biometrics Unit Mimeo Series, Cornell University, Ithaca, New York.

4. Federer, W. T. and Balaam, L. N. (1972). Bibliography on Experiment and Treatment Design Pre-1968. Oliver and Boyd, Edinburgh.

5. Federer, W. T., Hedayat, A., Lowe, C. C., and Raghavarao, D. (1976). An application of statistical design theory to crop estimation with special reference to legumes and mixtures of cultivars. Agronomy Journal 68(6): 914-919.

6. Fisher, R. A. (1958). Cigarettes, cancer, and statistics. Centennial Review 2: 151-166.

7. Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. Aust. J. Biol. Sci. 9: 463-493.

8. Hall, D. B. (1976). Mixing designs: a general model, simplifications, and some minimal designs. M. S. Thesis, Cornell University, Ithaca, New York.

9. Kempthorne, O. (1957). An Introduction to Genetic Statistics. John Wiley and Sons, Inc., New York and London.

10. Mead, R. and Pike, D. J. (1975). A review of response surface methodology from a biometrics viewpoint. Biometrics 31: 803-851.

11. Randall, J. H. (1976). The diallel cross. M. S. Thesis, Cornell University, Ithaca, New York.

12. Smith, L. L., Federer, W. T., and Raghavarao, D. (1974). A comparison of three techniques for eliciting truthful answers to sensitive questions. *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 447-452.

13. Yates, F. (1937). *The Design and Analysis of Factorial Experiments*. Technical Communication No. 35 of the Commonwealth Bureau of Soils, Harpenden, England.