Arbeidsnotater

STATISTISK SENTRALBYRÅ
Dronningensgt. 16, Oslo-Dep., Oslo 1. Tlf. 41 38 20

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 74/47 15. november 1974

ON THE OPTIMAL ALLOCATION OF OBSERVATIONS IN EXPERIMENTS WITH MIXTURES

By Petter Laake

CONTENTS

ABSTRACT

Methods for designing experiments with mixtures have been devised by Scheffé. In this paper a technique for the optimal allocation of observations is presented. The optimality criterion is to choose the number of observations in each experimental point such that the integrated variance is minimized. The technique is applied to Scheffé's simplex-lattice design and Scheffé's simplex-centroid design.

Key words: Experiments with mixtures, optimal allocation of observations, simplex-lattice, simplex-centroid.

1. Introduction

Consider an experiment with mixtures, that is an experiment where the response does not depend on the total amount in the mixture, but only on the proportions of the components. For instance this is the case if we study the octane rating of a blend of gasolines.

A formal theory for such experiments was developed by Scheffé (1958, 1963).

Let x_i denote the proportion of component i in the mixture and q the number of components, so that

$$x_i \geqq 0 \quad \text{for} \quad i = i,2,\dots,q, \tag{1.1}$$

and

$$x_1+x_2+\dots+x_q = 1.$$

Hence the experimental design is restricted to the (q-1)-dimensional simplex

$$S_{q-1} = \{(x_1,\dots,x_{q-1}) | 0 \leq \sum_{i=1}^{q-1} x_1 \leq 1, \quad x_i \geqq 0, \quad i=1,2,\dots,q-1\}. \tag{1.2}$$

Scheffé (1958) introduced the $\{q,m\}$ simplex-lattice design where the proportions of component i are

$$x_i = 0,\frac{1}{m},\frac{2}{m},\dots,1 \quad \text{for} \quad i = 1,2,\dots,q. \tag{1.3}$$

The design consists of all possible mixtures with these proportions of the components. Scheffé (1958) studied the problem of fitting the response by the m-th degree polynomial

$$\eta = \alpha_0 + \sum_{i=1}^{q} \alpha_i x_i + \sum_{1\leq i\leq j\leq q} \alpha_{ij} x_i x_j + \sum_{1\leq i\leq j\leq k\leq q} \alpha_{ijk} x_i x_j x_k + \dots$$

$$\dots + \sum_{1\leq i_1\leq i_2\leq\dots\leq i_m\leq q} \alpha_{i_1 i_2 \dots i_m} x_{i_1} \dots x_{i_m}. \tag{1.4}$$

The coefficients α are not unique. We may for instance introduce the restriction (1.1), which gives a reduced polynomial in the q-1 variables $x_1,x_2,\dots,x_{q-1}$. The reduced polynomial is of degree m and has just as many coefficients as these are experimental points on the $\{q,m\}$ simplex-lattice, which makes the polynomial well adapted to the simplex-lattice design.

Becker (1970) considered the problem of choosing an optimal design on the simplex for a general q, assuming that a first degree polynomial is fitted to the observed responses and the true model is a polynomial of second degree. Becker used the optimality criterion given by Box and Draper (1959), consisting of minimizing the expected square deviation averaged over the "region of interest". This problem was earlier considered by Draper and Lawrence (1965a, b) for $q = 3$ and $q = 4$. In this paper an attempt is made to find an optimal allocation of the observations taken on the simplex (1.2) for a given number of observations and a given design. Our optimality criterion is to minimize the integrated variance of the estimated response over S_{q-1}. In section 3 this criterion is applied and the calculations carried out in detail for the second degree polynomial. Section 4 gives without proof the results for the third degree polynomial. In section 5 the alternative simplex-centroid design is introduced, and the optimal allocation of observations is given for $q = 3$ and $q = 4$. Details of the omitted proofs can be found in Laake (1973).

2. Notations and definitions

Let the response to pure component i be denoted by η_i. The response to a 1:1 mixture of the components i and j we shall denote η_{ij}. The responses to 2:1 and 1:2 mixtures of components i and j are denoted by η_{iij} and η_{ijj} respectively. Similarly the observed responses are denoted y_i, y_{ij}, y_{iij} and y_{ijj}. The notation is analogous for mixtures with more than two components. Let

$$\underset{\sim}{\eta}' = (\eta_1,\ldots,\eta_q,\eta_{12},\eta_{13},\ldots,\eta_{q-1q},\ldots) \quad \text{and}$$

$$\underset{\sim}{y}' = (y_1,\ldots,y_q,y_{12},\ldots,y_{q-1q},\ldots)$$

be the vectors of responses and observed responses, respectively, We assume that

$$E\underset{\sim}{y} = \underset{\sim}{\eta},$$

and

$$E(\underset{\sim}{y}-\underset{\sim}{\eta})(\underset{\sim}{y}-\underset{\sim}{\eta})' = \sigma^2 \underset{\sim}{I}. \tag{2.1}$$

The parameters in the regression function are estimated by the method of least squares, and the estimated response is denoted by $\tilde{\eta}$. Let

$$W = \int_{S_{q-1}} \operatorname{var} \tilde{\eta} \, dx_1 \ldots dx_{q-1}$$

denote the integrated variance over S_{q-1}. The total number of observations equals N.

Our optimality criterion is to choose the number of observations in each experimental point so that W is minimized. The criterion is chosen in order to obtain the best possible representation of the observations over the whole region S_{q-1}.

3. Optimal allocation of observations for the second degree polynomial

In accordance with (1.4) the general second degree polynomial has the form

$$\eta = \alpha_0 + \sum_{i=1}^{q} \alpha_i x_i + \sum_{1 \leq i \leq j \leq q} \alpha_{ij} x_i x_j .$$

Subject to (1.1) the reduced polynomial can be written as

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1 \leq i < j \leq q} \beta_{ij} x_i x_j \qquad (3.1)$$

where

$$\beta_i = \alpha_0 + \alpha_i + \alpha_{ii} \quad \text{for} \quad i=1,2,\ldots,q,$$

and

$$\beta_{ij} = \alpha_{ij} - \alpha_{ii} - \alpha_{jj} \quad \text{for} \quad i < j.$$

Let the design be a {q,2} simplex-lattice. Then the experimental points consist of all the points satisfying.

$$\sum_{i=1}^{q} x_i = 1,$$

and

$$x_i = 0,\ 1/2,\ 1 \quad \text{for} \quad i = 1,2,\ldots,q.$$

The coefficients in the polynomial (3.1) are uniquely determined by the responses at the points of the $\{q,2\}$ simplex-lattice, and the estimation of the coefficients is carried out in Scheffé (1958, page 348). In accordance with earlier notation we denote the mean of the observed responses to the pure component i by $\bar{y}_i$ and the mean of the responses to a 1:1 mixture of the components i and j by $\bar{y}_{ij}$. According to Scheffé (1958, page 353) the estimated polynomial is

$$\tilde{\eta} = \sum_{i=1}^{q} a_i \bar{y}_i + \sum_{1 \leq i < j \leq q} a_{ij} \bar{y}_{ij},$$

where

$$a_i = x_i(2x_i - 1),$$

and

$$a_{ij} = 4x_i x_j. \tag{3.2}$$

By the assumption (2.1) the observations have equal variance σ^2. Let the number of observations of the responses to pure components and 1:1 mixtures be r_i and r_{ij}, respectively. Then

$$\operatorname{var} \tilde{\eta} = \sigma^2 \left(\sum_{i=1}^{q} \frac{a_i^2}{r_i} + \sum_{i<j} \frac{a_{ij}^2}{r_{ij}} \right). \tag{3.3}$$

The optimality criterion is to minimize W subject to the side condition

$$\sum_{i=1}^{q} r_i + \sum_{i<j} r_{ij} = N. \tag{3.4}$$

In order to calculate W we need the following lemma.

<u>Lemma</u>

<u>Suppose that</u>

$$x_q = 1 - x_1 - \ldots - x_{q-1},$$

<u>and</u> S_{q-1} <u>is defined by</u>. <u>Then</u>

$$\int_{S_{q-1}} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_q^{\alpha_q - 1} \, dx_1 \ldots dx_{q-1} = \frac{\prod_{i=1}^{q} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{q} \alpha_i\right)}. \tag{3.5}$$

The proof of the lemma is outlined in DeGroot (1970, page 63).

In order to find W we need the quantities

$$a_1(q) = \int_{S_{q-1}} a_i^2 dx_1 \cdots dx_{q-1} = \frac{2(q^2-7q+18)}{(3+q)!} \quad \text{for } i=1,2,\ldots,q,$$

and

$$b_1(q) = \int_{S_{q-1}} a_{ij}^2 dx_1 \cdots dx_{q-1} = \frac{64}{(3+q)!} \quad \text{for } i < j,$$

which are evaluated by means of (3.5). Hence

$$W = \int_{S_{q-1}} \sigma^2 \sum_{i=1}^{q} a_i^2 \frac{1}{r_i} dx_1 \cdots dx_{q-1} + \int_{S_{q-1}} \sigma^2 \sum_{i<j} a_{ij}^2 \frac{1}{r_{ij}} dx_1 \cdots dx_{q-1}$$

$$= \sigma^2 \Big[a_1(q) \sum_{i=1}^{q} \frac{1}{r_i} + b_1(q) \sum_{i<j} \frac{1}{r_{ij}} \Big].$$

The problem of minimizing W subject to the side condition (3.4) is solved by the method of Lagrange multipliers. This is done by differentiation of the function

$$\Phi = a_1(q) \sum_{i=1}^{q} \frac{1}{r_i} + b_1(q) \sum_{i<j} \frac{1}{r_{ij}} + \lambda\Big(\sum_{i=1}^{q} r_i + \sum_{i<j} r_{ij} - N\Big),$$

which yields

$$\frac{\partial \Phi}{\partial r_i} = -a_1(q) \frac{1}{r_i^2} + \lambda,$$

$$\frac{\partial \Phi}{\partial r_{ij}} = -b_1(q) \frac{1}{r_{ij}^2} + \lambda,$$

and

$$\frac{\partial \Phi}{\partial \lambda} = \sum_{i=1}^{q} r_i + \sum_{i<j} r_{ij} - N.$$

These partial derivatives equal zero for

$$r_i = \sqrt{a_1(q)\lambda^{-1}} \quad \text{for } i=1,2,\ldots,q, \tag{3.6}$$

and

$$r_{ij} = \sqrt{b_1(q)\lambda^{-1}} \quad \text{for } i < j.$$

Substituting (3.6) into the side condition gives

$$r_i = N\sqrt{a_1(q)}/\{q\sqrt{a_1(q)} + \tbinom{q}{2}\sqrt{b_1(q)}\} \quad \text{for } i=1,2,\ldots,q,$$

and

$$r_{ij} = N\sqrt{b_1(q)}/\{q\sqrt{a_1(q)} + \tbinom{q}{2}\sqrt{b_1(q)}\} \quad \text{for } i < j. \tag{3.7}$$

That this point really gives the minimum is seen as follows: The function W is convex. Since (3.4) is a linear combination of r_i and r_{ij}, the function ϕ is a convex function and clearly takes its minimum in the interior of the region $\{r_i, r_{ij} | r_i>0, r_{ij}>0\}$. Hence the only extremal point must be the minimum point.

Thus the optimal allocation consists of choosing the same number of observations of the responses to each "pure component", and the same number of observations of the responses to 1:1 mixtures. The relative proportion of the number of observations is given by

$$r_i/r_{ij} = \sqrt{a_1(q)}/\sqrt{b_1(q)} \quad \text{for } i < j. \tag{3.8}$$

Let

$$s_{ij} = qr_i/\binom{q}{2}r_{ij} \quad \text{for } i < j. \tag{3.9}$$

denote the relative proportions of the observations used to estimate the "main effects" and the "interactions".

In table 1 is given r_i/r_{ij} and s_{ij} for different values of q.

Table 1

q	r_i/r_{ij}	s_{ij}
3	0,433	0,433
4	0,433	0,289
5	0,500	0,250
6	0,612	0,245
7	0,750	0,250
8	0,901	0,257
9	1,060	0,265
10	1,225	0,272
20	2.948	0.310

Table 1 indicates that if $q \leq 8$, r_i is chosen smaller than r_{ij} for $i<j$. If $q>8$, the inequality is reversed. However, s_{ij} does not vary much for $4 \leq q \leq 20$.

4. Optimal allocation of observations for the special cubic and general cubic polynomial

This section gives an outline of the results. The proofs can be found in Laake (1973).

Consider the special cubic polynomial

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1\le i<j\le q} \beta_{ij} x_i x_j + \sum_{1\le i<j<k\le q} \beta_{ijk} x_i x_j x_k . \tag{4.1}$$

In designing the experiment we adopt the {q,2} simplex-lattice augmented by an experimental point corresponding to a 1:1:1 mixture. Applying the optimality criterion we find: Choose r_i, r_{ij} and r_{ijk} so that

$$r_i : r_{ij} : r_{ijk} = \sqrt{a_2(q)} : \sqrt{b_2(q)} : \sqrt{c_2(q)},$$

where

$$a_2(q) = \frac{q^4-10q^3+59q^2-218q+1608}{2(5+q)!},$$

$$b_2(q) = \frac{16}{(5+q)!}(16q^2-144q+392), \tag{4.2}$$

and

$$c_2(q) = (27)^2 \frac{8}{(5+q)!} .$$

Scheffé (1958, page 347) shows that the most general polynomial of third degree can be written in the form

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1\le i<j\le q} \beta_{ij} x_i x_j + \sum_{1\le i<j\le q} \gamma_{ij} x_i x_j (x_i - x_j)$$

$$+ \sum_{1\le i<j<k\le q} \beta_{ijk} x_i x_j x_k . \tag{4.3}$$

To fit the general cubic polynomial the observations are taken on the {q,3} simplex-lattice. The optimal allocation gives: Choose r_i, r_{ij}, r_{ijj} and r_{ijk} so that

$$r_{iij} = r_{ijj} \text{ for } i<j,$$

and

$$r_i : r_{iij} : r_{ijk} = \sqrt{a_3(q)} : \sqrt{b_3(q)} : \sqrt{c_3(q)},$$

where

$$a_3(q) = \frac{1}{4(5+q)!}(8q^4-104q^3+784q^2-3088q+5280),$$

$$b_3(q) = \frac{81}{(5+q)!}(q^2-9q+38), \tag{4.4}$$

and

$$c_3(q) = \frac{8(27)^2}{(5+q)!} .$$

5. Definition of the simplex-centroid design and the optimal allocation of observations

Scheffé (1963) introduced the alternative simplex-centroid design on the simplex. This design comprises 2^q-1 experimental points corresponding to the q permutations of $(1,0,\dots,0)$, the $\binom{q}{2}$ permutations of $(\frac{1}{2},\frac{1}{2},0,\dots,0)$, the $\binom{q}{3}$ permutations of $(\frac{1}{3},\frac{1}{3},\frac{1}{3},0,\dots,0),\dots,$ and the point $(\frac{1}{q},\frac{1}{q},\dots,\frac{1}{q})$. The $\{q,m\}$ simplex-lattice design differs from the simplex-centroid design in that for a given q there is a family of alternative $\{q,m\}$ designs for $m=1,2,\dots,$ but there is only a single simplex-centroid design.

Under the simplex-centroid design, the regression function

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1\le i<j\le q} \beta_{ij} x_i x_j + \dots + \beta_{12\dots q} x_1 x_2 \dots x_q \tag{5.1}$$

is fitted to the response by means of the method of least squares.

In order to find an optimal allocation of observations we observe that the polynomial regressions (4.1) and (5.1), and the simplex-centroid design and the augmentented $\{q,2\}$ simplex-lattice are identical for $q=3$. The optimal allocation of observations for $q=3$ is therefore given by substituting $q=3$ in (4.2). Hence the conclusion is to choose r_i, r_{ij} and r_{123} so that

$$r_i : r_{ij} : r_{123} = 1 : 1.60 : 4.00 \quad \text{for } i<j. \tag{5.2}$$

Suppose $q=4$ and consider the polynomial

$$\eta = \sum_{i=1}^{4} \beta_i x_i + \sum_{1\le i<j\le 4} \beta_{ij} x_i x_j + \sum_{1\le i<j<k\le 4} \beta_{ijk} x_i x_j x_k + \beta_{1234} x_1 x_2 x_3 x_4$$

together with the simplex-centroid design. Omitting the proof in Laake (1973), the optimal allocation leads to the following conclusion: Choose r_i, r_{ij}, r_{ijk} and r_{1234} so that

$$r_i : r_{ij} : r_{ijk} : r_{1234} = 1 : 1.30 : 2.10 : 3.84 \quad \text{for } i<j<k. \tag{5.3}$$

Acknowledgments

My thanks are due to Professor Emil Spjøtvoll for his helpful suggestions and the referees for various comments.

References

Becker, N.G. (1970): Mixture designs for a model linear in the proportions. Biometrika **57**, 329-338.

Box, G.E.P. and Draper, N.R. (1959): A basis for the selection of a response surface design. J. Am. Statist. Ass. **54**, 622-654.

DeGroot, M.H. (1970): Optimal Statistical Decisions. McGraw Hill, New York.

Draper, N.R. and Lawrence, W.L. (1965a): Mixture designs for three factors. J. Roy. Statist. Soc. Ser. B. **27**, 450-465.

Draper, N.R. and Lawrence, W.L. (1965b): Mixture designs for four factors. J. Roy. Statist. Soc. Ser. B. **27**, 473-478.

Laake, P. (1973): Some optimal properties for experiments with mixtures. (In Norwegian.) Institute of Mathematics, University of Oslo.

Scheffé, H. (1958): Experiments with mixtures. J. Roy. Statist. Soc. Ser. B. **20**, 344-360.

Scheffé, H. (1963): The simplex-centroid design for experiments with mixtures. J. Roy. Statist. Soc. Ser. B. **25**, 235-263.