STATISTICAL RESEARCH REPORT No 7
Institute of Mathematics 1973
University of Oslo

ON THE OPTIMAL ALLOCATION OF OBSERVATIONS IN EXPERIMENTS WITH MIXTURE

by

Petter Laake

Contents

1. Introduction.

Consider an experiment with mixture, that is an experiment where the property studied does not depend on the total amount in the mixture, but only on the proportions of the factors. The property studied is called the response.

Denote the i-th factor by x_i and suppose that we are studying a q-component mixture with

$$\begin{aligned} & x_i \geqq 0 \qquad i = 1,2,\dots,q \\ & x_1+x_2+\dots+x_q=1 \end{aligned} \tag{1.1}$$

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

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

Scheffé (1958) introduced the $\{q,m\}$-simplex-lattice design where the values of factor x_i are

$$x_i = 0,\tfrac{1}{m},\tfrac{2}{m},\dots,1 \qquad i = 1,2,\dots,q \tag{1.3}$$

All possible mixtures with these proportions of the factors are used. The polynomial associated with the simplex-lattice is

$$\eta = \beta_o + \sum_{i=1}^{q} \beta_i x_i + \sum_{1\leq i\leq j\leq q} \beta_{ij} x_i x_j + \sum_{1\leq i\leq j\leq k\leq q} \beta_{ijk} x_i x_j x_k + \dots$$
$$\dots + \sum_{1\leq i_1\leq i_2\leq \dots \leq i_m\leq q} \beta_{i_1 i_2 \dots i_m} x_{i_1} \dots x_{i_m} \tag{1.4}$$

This polynomial has as many coefficients as there are design-points in the $\{q,m\}$-simplex-lattice design.

Let the estimated polynomial be

$$\tilde{\eta} = \hat{\beta}_o + \sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{i\leq j} \hat{\beta}_{ij} x_i x_j + \sum_{i\leq j\leq k} \hat{\beta}_{ijk} x_i x_j x_k + \ldots$$

$$\ldots + \sum_{i_1\leq\ldots\leq i_m} \hat{\beta}_{i_1 i_2 \ldots i_m} x_{i_1} \ldots x_{i_m}$$

where the $\hat{\beta}$-s are the least-squares estimates. The results for some given simplex-lattice designs and the associated polynomials can be found in Scheffé (1958), Gorman and Hinman (1962).

Box and Draper (1959) considered the choice of design on S_{q-1} for fitting a first order polynomial model. They used the optimality criterion based on minimizing the mean square deviation averaged over the experimental region when the true model is a polynomial of second order. Draper and Lawrence (1965a,b) considered the problem for $m=3$ and $m=4$. Becker (1970) considered the choice of design for a general m and proved the generalization of the suggestions made by Box, Draper and Lawrence.

We are searching for an optimal allocation of the observations taken on the simplex-lattice. Let

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

be integrated variance over S_{q-1}. Suppose that total number of observations equals N. Our optimality criterion is to choose the number of observations in each designpoint so that W is minimized.

The fundamental results concerning $\operatorname{var} \tilde{\eta}$ can be found in section 7 in Scheffé (1958).

2.1. Optimal allocation of observations for the linear polynomial.

Consider the linear polynomial

$$\eta = \sum_{i=1}^{q} \beta_i x_i$$

and a $\{q,1\}$-simplex-lattice. We are thus studying the response of "pure components". Suppose that $\hat{\eta}_i$ is the observed response on the $\{q,1\}$-simplex-lattice. According to Scheffé (1958)

$$\tilde{\eta} = \sum_{i=1}^{q} \hat{\eta}_i x_i$$

and

$$\operatorname{var} \tilde{\eta} = \sum_{i=1}^{q} x_i^2 \frac{\sigma^2}{r_i}$$

since we assume that the observations are independent with equal variance σ^2. Let r_i be the number of observations on each lattice-point.

Then

$$\int_{S_{q-1}} \operatorname{var} \tilde{\eta}\, dx_1 \ldots dx_{q-1} = \sum_{i=1}^{q} \frac{\sigma^2}{r_i} \int_{S_{q-1}} x_i^2 dx_1 \ldots dx_{q-1} \qquad (2.1.1)$$

We want to minimize (2.1.1) under the side condition

$$\sum_{i=1}^{q} r_i = N$$

According to (A.1) in Appendix

$$W = \int_{S_{q-1}} \operatorname{var} \tilde{\eta}\, dx_1 \ldots dx_{q-1} = \sigma^2 \frac{\Gamma(3)}{\Gamma(2+q)} \sum_{i=1}^{q} \frac{1}{r_i} = \sigma^2 a_1(q) \sum_{i=1}^{q} \frac{1}{r_i}$$

where

$$a_1(q) = \frac{\Gamma(3)}{\Gamma(2+q)}.$$

Here W is to be minimized under the side condition

$$\sum_{i=1}^{q} r_i = N$$

Introduce

$$W_1 = \frac{W}{\sigma^2} = a_1(q) \sum_{i=1}^{q} \frac{1}{r_i}$$

W_1 is then to be minimized under the given side condition. This extremum problem can be solved by studying

$$\Phi = a_1(q) \sum_{i=1}^{q} \frac{1}{r_i} + \lambda(\sum_{i=1}^{q} r_i - N)$$

Thus

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

$$\frac{\partial \Phi}{\partial \lambda} = \sum_{i=1}^{q} r_i - N$$

The extremum value is thus the solution of

$$\frac{\partial \Phi}{\partial r_i} = \frac{\partial \Phi}{\partial \lambda} = 0$$

Hence

$$r_i = \frac{N}{q} \qquad i = 1,2,\ldots,q$$

This indicates that, using a linear polynomial, we take equal number of observations of the response to "pure components". The result seems intuitively obvious.

If N is a multiple of q, r_i is an integer. If N is not a multiple of q, that is

$$kq < N \leq (k+1)q \ , \quad k \text{ an integer,}$$

we choose k observations of the response to each "pure component". The remainding $N-kq$ observations can either be distributed randomly on the lattice-points or according to special interest in the coefficients β_i .

Obviously the solution of the extremum problem gives a minimum value of W . Suppose that $r_1, r_2, \ldots, r_{q-1}$ are chosen sufficiently close to 0 , and

$$r_q = N - \sum_{i=1}^{q-1} r_i$$

Thus

$$\sum_{i=1}^{q} \frac{1}{r_i}$$

can be made as large as we want. Consequently we can make W as large as we want at the same time as the side condition

$$\sum_{i=1}^{q} r_i = N$$

is fullfilled. The extremum point is thus a minimum point.

2.2. Optimal allocation of observations for the quadratic polynomial.

Consider the polynomial

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1 \le i < j \le q} \beta_{ij} x_i x_j \tag{2.2.1}$$

and a $\{q,2\}$-simplex-lattice, which means that the q factors are given by

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

$$x_i = 0, \tfrac{1}{2}, 1 \qquad i = 1,2,\ldots,q$$

From this design the coefficients in the polynomial (2.2.1) are estimated. This is carried out in Scheffé (1958). Suppose that $\hat{\eta}_i$ and $\hat{\eta}_{ij}$ are the means of the observed responses on the simplex-lattice. According to Scheffé (1958) the estimated polynomial is

$$\tilde{\eta} = \sum_{i=1}^{q} a_i \hat{\eta}_i + \sum_{1 \le i < j \le q} a_{ij} \hat{\eta}_{ij}$$

where

$$a_i = x_i(2x_i - 1)$$
$$a_{ij} = 4x_i x_j \tag{2.2.2}$$

Suppose that the observations are independent with equal variance, σ^2 and the numbers of observations of the response to "pure components" and mixtures with $x_i = x_j = \frac{1}{2}$ are r_i and r_{ij}. We then get

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

The optimality criterion is now to minimize

$$W = \int_{S_{q-1}} \text{var}\ \tilde{\eta}\ dx_1 \dots dx_{q-1}$$

subject to the side condition

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

We consider

$$\int_{S_{q-1}} \text{var}\ \tilde{\eta}\ dx_1 \dots dx_{q-1} = \int_{S_{q-1}} \sigma^2 \sum_{i=1}^{q} a_i^{\,2} \frac{1}{r_i}\ dx_1 \dots dx_{q-1}$$

$$+ \int_{S_{q-1}} \sigma^2 \sum_{i<j} a_{ij}^{\,2} \frac{1}{r_{ij}}\ dx_1 \dots dx_{q-1} \qquad (2.2.3)$$

and calculate

$$a_2(q) = \int_{S_{q-1}} a_i^{\,2} dx_1 \dots dx_{q-1}$$

$$b_2(q) = \int_{S_{q-1}} a_{ij}^{\,2} dx_1 \dots dx_{q-1}$$

According to (A.1) in Appendix we get

$$a_2(q) = \int_{S_{q-1}} x_i^{\,2}(2x_i-1)^2 dx_1 \dots dx_{q-1} = \frac{2(q^2-7q+18)}{(3+q)!} \qquad (2.2.4)$$

$$i = 1,2,\dots,q$$

$$b_2(q) = \int_{S_{q-1}} 16x_i^2 x_j^2 dx_1 \ldots dx_{q-1}$$

$$= \frac{64}{(3+q)!} \qquad \begin{matrix} i = 1,2,\ldots,q \\ j = 1,1,\ldots,q \\ i<j \end{matrix} \tag{2.2.5}$$

Substituting (2.2.4) and (2.2.5) into (2.2.3), we get

$$\int_{S_{q-1}} \text{var}\, \tilde{\eta}\, dx_1 \ldots dx_{q-1} = \sigma^2[a_2(q) \sum_{i=1}^{q} \frac{1}{r_i} + b_2(q) \sum_{i<j} \frac{1}{r_{ij}}]$$

We introduce

$$W_1 = a_2(q) \sum_{i=1}^{q} \frac{1}{r_i} + b_2(q) \sum_{i<j} \frac{1}{r_{ij}} \tag{2.2.6}$$

and we are interested in minimizing (2.2.6) subject to the side condition

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

The problem is solved by differentiating

$$\Phi = a_2(q) \sum_{i=1}^{q} \frac{1}{r_i} + b_2(q) \sum_{i<j} \frac{1}{r_{ij}} + \lambda(\sum_{i=1}^{q} r_i + \sum_{i<j} r_{ij} - N) \tag{2.2.7}$$

which yields

$$\frac{\partial \Phi}{\partial r_i} = -a_2(q) \frac{1}{r_i} + \lambda$$

$$\frac{\partial \Phi}{\partial r_{ij}} = -b_2(q) \frac{1}{r_{ij}} + \lambda$$

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

We then solve the equations

$$\frac{\partial \Phi}{\partial r_i} = \frac{\partial \Phi}{\partial r_{ij}} = \frac{\partial \Phi}{\partial \lambda} = 0$$

and get

$$r_i = \frac{\sqrt{a_2(q)}}{\sqrt{\lambda}}$$

(2.2.8)

$$r_{ij} = \frac{\sqrt{b_2(q)}}{\sqrt{\lambda}}$$

Substituting (2.2.8) into the side condition we get

$$r_i = N \frac{\sqrt{a_2(q)}}{q\sqrt{a_2(q)}+\binom{q}{2}\sqrt{b_2(q)}} \qquad i = 1,2,\ldots,q$$

(2.2.9)

$$r_{ij} = N \frac{\sqrt{b_2(q)}}{q\sqrt{a_2(q)}+\binom{q}{2}\sqrt{b_2(q)}} \qquad \begin{array}{l} i = 1,2,\ldots,q \\ j = 1,2,\ldots,q \\ \quad i<j \end{array}$$

We are thus led to the conclusion of taking the same number of observations of the responses to each "pure component" and the same number of observations of the responses to mixtures where $x_i = x_j = \frac{1}{2}$. The relative proportion of the number of observations is given by

$$\frac{r_i}{r_{ij}} = \frac{\sqrt{a_2(q)}}{\sqrt{b_2(q)}} \qquad \begin{array}{l} i = 1,2,\ldots,q \\ j = 1,2,\ldots,q \\ \quad i<j \end{array} \qquad (2.2.10)$$

Using an argument similar to the argument used in section 2.1, we get that the solution (2.2.9) gives minimum value of W.

Ex. 1: We are interested in studying the relative proportions of observations, given by (2.2.10) for some values of q. The result is given in table 1

q	r_i/r_{ij}
3	0.433
4	0.433
5	0.500
6	0.612
7	0.750
8	0.901
9	1.060
10	1.225
20	2.948

Table 1

For each value of q we choose

$$r_1 = r_2 = \ldots = r_q$$

$$r_{12} = r_{13} = \ldots = r_{q-1q}$$

Table 1 indicates that if $q \leqq 8$, r_i and r_{ij} are chosen, according to the optimality criterion, so that $r_i < r_{ij}$. This signifies that when there are few components in the mixture, most of the observations are used to estimate the "interaction" between the factors. When there are many components in the mixture, most of the observations are used to estimate the "main effects".

2.3. Optimal allocation of observations for the special cubic polynomial.

Consider the special cubic polynomial

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{1\leq i<j\leq q} \beta_{ij} x_i x_j + \sum_{1\leq i<j<k\leq q} \beta_{ijk} x_i x_j x_k \qquad (2.3.1)$$

When we have chosen the polynomial, we adopt the $\{q,2\}$-simplex-lattice argumented by the designpoints corresponding to mixture with $x_i = x_j = x_k = \frac{1}{3}$, $i,j,k = 1,2,\ldots,q$, $i < j < k$. Scheffé (1958) found that estimated response is given by

$$\tilde{\eta} = \sum_{i=1}^{q} b_i \hat{\eta}_i + \sum_{i<j} b_{ij} \hat{\eta}_{ij} + \sum_{i<j<k} b_{ijk} \hat{\eta}_{ijk}$$

and

$$b_i = \tfrac{1}{2} x_i (6x_i^2 - 2x_i + 1 - 3\sum_{j=1}^{q} x_j^2)$$

$$b_{ij} = 4x_i x_j (3x_i + 3x_j - 2)$$

$$b_{ijk} = 27 x_i x_j x_k$$

The observations are assumed to be independent with equal variance σ^2 , and r_i , r_{ij} and r_{ijk} are the numbers of observations on η_i , η_{ij} and η_{ijk} . The variance of the estimated response is

$$\text{var}\ \tilde{\eta} = \sum_{i=1}^{q} b_i^2 \frac{\sigma^2}{r_i} + \sum_{i<j} b_{ij}^2 \frac{\sigma^2}{r_{ij}} + \sum_{i<j<k} b_{ijk}^2 \frac{\sigma^2}{r_{ijk}}$$

Minimizing

$$W = \int_{S_{q-1}} \text{var}\ \tilde{\eta}\ dx_1 \cdots dx_{q-1}$$

subject to the side condition

$$\sum_{i=1}^{q} r_i + \sum_{i<j} r_{ij} + \sum_{i<j<k} r_{ijk} = N$$

leads to the following conclusion: Choose r_i , r_{ij} and r_{ijk} so that

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

and

$$a_3(q) = \frac{q^4-10q^3+59q^2-218q+1608}{2(5+q)!}$$

$$b_3(q) = \frac{16}{(5+q)!}(16q^2-144q+392) \qquad (2.3.1)$$

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

For details concerning the proof, the reader is referred to Laake (1973). An application of (2.3.1) will be developed in section 3.1.

2.4. Optimal allocation of observations for the general cubic polynomial.

Consider the polynomial

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

$$+ \sum_{1\leq i<j<k\leq q} \beta_{ijk} x_i x_j x_k$$

and adopt the $\{q,3\}$-simplex-lattice. Applying the optimality criterion, we obtain the following conclusion: Choose r_i, r_{iij}, r_{ijj} and r_{ijk} so that

$$r_{iij} = r_{ijj} \qquad \begin{array}{l} i = 1,2,\dots,q \\ j = 1,2,\dots,q \\ i < j \end{array}$$

and

$$r_i : r_{iij} : r_{ijk} = \sqrt{a_4(q)} : \sqrt{b_4(q)} : \sqrt{c_4(q)}$$

where

$$a_4(q) = \frac{1}{4(5+q)!}(8q^4-104q^3+784q^2-3088q+5280)$$

$$b_4(q) = \frac{81}{(5+q)!}(q^2-9q+38)$$

$$c_4(q) = \frac{8(27)^2}{(5+q)!}$$

For details the reader is referred to Laake (1973).

3. Definition of the simplex-centroid design.

Scheffé (1963) has proposed an alternative design on the simplex. The design is called the simplex-centroid design and is defined by

q observations of "pure components"

$\binom{q}{2}$ observations of mixtures of two components with equal proportions

$\binom{q}{3}$ observations of mixtures of three components with equal proportions

.

.

.

1 observation of the mixture with q components all equal to $\frac{1}{q}$.

Suppose that the response can be expressed by the polynomial

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

Estimated response is given by

$$\tilde{\eta} = \sum_{i=1}^{q} \hat{\beta}_i x_i + \sum_{1 \le i < j \le q} \hat{\beta}_{ij} x_i x_j + \ldots + \hat{\beta}_{12 \ldots q} x_1 x_2 \ldots x_q$$

where the $\hat{\beta}$-s are least squares estimates. The $\{q,m\}$-simplex-lattice designs differ from the simplex-centroid design in that for a given q there is a family of alternative $\{q,m\}$ designs for $m = 1,2,\ldots$, but there is a single simplex-centroid design.

3.1 Optimal allocation of observations for the simplex-centroid design with $q = 3$.

In section 2.3 we considered an optimal allocation of observations for the special cubic polynomial and for a general q . Comparing the simplex-lattice design and the associated polynomial in section 2.3 with the simplex-centroid design in section 3, we see that the models are identical for $q = 3$. The optimal allocation of observations for $q = 3$ is therefore given by substituting $q = 3$ in (2.3.1). Hence the conclusion is to choose r_i, r_{ij} and r_{123} so that

$$r_i : r_{ij} : r_{123} = 1:1.60:3.00 \qquad \begin{array}{c} i = 1,2,3 \\ j = 1,2,3 \\ i < j \end{array}$$

3.2 Optimal allocation of observations in the simplex-centroid design with q = 4.

Consider the polynomial

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

and the simplex-centroid design with $q = 4$. The optimum procedure now leads to the following conclusion: Choose r_i, r_{ij}, r_{ijk} and r_{1234} so that

$$r_i : r_{ij} : r_{ijk} : r_{1234} = 1:1.30:2.10:3.84 \qquad \begin{array}{c} i = 1,2,3,4 \\ j = 1,2,3,4 \\ k = 1,2,3,4 \\ i < j < k \end{array}$$

For details concerning the proof the reader is referred to Laake (1973).

Appendix

Suppose a random vector $x = (x_1,x_2,\ldots,x_q)$ has a Dirichlet distribution with parameter vector $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_q)$, $\alpha_i > 0$, $i = 1,2,\ldots,q$. According to DeGroot (1970) page 51 we have

Lemma A.1: Suppose that

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

and S_{q-1} is defined by (1.2). Then

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

References

N.G. Becker (1970): Mixture design for a model linear in the proportions. Biometrika, 57, 329-38.

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

M. DeGroot (1970): Optimal statistical decisions. McGraw Hill Book Company.

N.R. Draper and W.L. Lawrence (1965a): Mixture designs for three factors. J. Roy. Statist.Soc., B27, 450-65.

N.R. Draper and W.L. Lawrence (1965b): Mixture designs for four factors. J. Roy. Statist.Soc., B27, 473-8

J.W. Gorman and J.E. Hinman (1962): Simplex-lattice-designs for multicomponent systems. Technometrics, 4, 463-88.

P. Laake (1973): Noen optimale egenskaper i eksperimenter med blanding. Hovedoppgave i statistikk. Matematisk institutt, Universitetet i Oslo.

H. Scheffé (1958): Experiments with mixtures. J. Roy. Statist. Soc., B20, 344-60.

H. Scheffé (1963): The simplex-centroid design for experiments with mixtures. J. Roy. Statist. Soc., B25, 235-63.