VARIABILITY IN VARIETY TRIALS

by

G. K. SHUKLA

Submitted for the Degree of
Doctor of Philosophy
University of Edinburgh
March 1974

SUMMARY

This thesis examines some statistical methods that may be helpful in the planning and analysis of series of variety trials. The following aspects of variability among variety yields and their use are considered:

1) Some varieties may be more variable than others over sites and years. A parameter, the 'stability variance', is defined which gives a measure of variability after eliminating additive site effect common to all the varieties. Efficient methods of estimation and test of significance are given.

2) Some methods of investigating the causes of heterogeneity are considered. This heterogeneity can be due to differential effects of environmental factors which give rise to interactions. The procedures for studying the relationship between interaction and environmental factors are studied by extending the usual two-way model of the analysis of variance. The use of such relationships in the recommendation of varieties for specific type of environment is discussed with examples. A method is given for estimating and predicting the yield of a particular variety at a site, when the yield of other varieties at that site are known.

3) Data analytic methods for recognising patterns in the data are considered. Their use in graphical representation of variety differences and variety-site interactions is explained with examples. The following applications of these methods are discussed; i) the investigation of causes of variability among varieties, ii) choice of varieties for recommendation, and selection of sites for future trials, iii) examination of the nature

of interaction by re-arranging rows and columns of residuals. Application (iii) is of great help in detecting the abnormal behaviour of varieties in some particular types of environment.

4) A method for optimum choice of number of replications, sites and years is considered. The method is a sequential procedure which maximises the expected gain from correct choice between two varieties. Knowledge is required of cost parameters of experimentation and the value of additional produce.

In some cases the cost parameters may not be known. Another sequential approach, controlling the probability of a specified amount of error is also considered. Details are given for two separate cases in which (i) one variety is chosen from several varieties, (ii) choice is made between new varieties and known standard variety.

VARIABILITY IN VARIETY TRIALS

by

G. K. SHUKLA

Submitted for the Degree of
Doctor of Philosophy
University of Edinburgh
March 1974

DECLARATION

The following record of research work is submitted as a Thesis for the Degree of Doctor of Philosophy of the University of Edinburgh, having been submitted for no other Degree. Except where the acknowledgement is made, the work is original.

Department of Statistics,
University of Edinburgh
March 1974

ACKNOWLEDGEMENTS

The Agricultural Research Council Unit of Statistics, Edinburgh, is responsible for analysing the data of trials conducted on agricultural crops. This work was done to help in the planning of trials, and the analysis and interpretation of the data.

I am very grateful to my supervisors, Professor D. J. Finney, F.R.S. and Mr H. D. Patterson, for suggesting the problem, and for their guidance and helpful criticism.

My thanks are also due to Mr M. Talbot for help in using some computer programs. I am very grateful to Mrs V. E. Simpson for her skilful mathematical typing.

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

1.0 Introduction

New varieties of crops are produced each year by plant breeders. These must be thoroughly tested before seed can be sold for commercial use. Three stages are required. In the first stage only a small number of plants are tested, in part because of scarcity of seed and in part because the number of potential new varieties may be very large indeed. Varieties will be passed on to the next stage mainly on the basis of immunity to disease, resistance to pests, or other factors. After preliminary trials comes the yield selection phase, where many varieties are produced and reduced to a smaller number on the basis of highest yield. In the following, the word yield has been used to describe a measure of a characteristic in which the experimenter is interested. In one case it can be weight of grain and in the other it can be some meaningful combination of different characteristics of plant. Results given in this thesis will be applicable to different defintiion of yields, under certain assumptions, as will be described later. Sometimes standard varieties are also included at this stage. Finally the selected varieties are also compared with the standard varieties on a large scale and those found markedly superior are passed into commercial use. Three stages are as follows:

Preliminary trials⟶ Yield selection⟶ Comparison with the standards.

In some cases this three-stage program is not strictly followed. The preliminary trial stage is conducted by the breeder and if he is satisfied with the performance of the variety he requests for its being included in the large scale trial. All the varieties have to be tested for performance in the field before the seed can be sold for commercial purposes. In this thesis we are only concerned with the third (final) stage [illegible] and the results given here are not affected in any way by the procedure adopted in the previous stages.

The first two stages are usually conducted on a small scale in research station. The final stage is usually conducted on a large scale, over many sites, and possibly repeated over several years. This is important because when varieties differ considerably in their genetic constitution, it is likely that they are affected to various degrees by variation in environmental factors such as soil type, agronomic practices, fertility, season and climate. Differential effects of environmental factors usually described as interaction between varieties and environments will be called "variety-environmental interaction" in the present work. If these interactions are substantial then the problem of recommending a single variety for the whole region (which may be heterogeneous for the environmental factors) becomes rather difficult. The two main reasons for conducting these trials on a large scale are as follows.

Firstly the estimate of the average relative performance of one variety over another is of considerable importance in recommending one variety in preference to another, even if their relative performance is considerably affected by the environmental factors. If the environments in which trials are to be conducted are selected randomly, this should give an estimate of the overall gain (or loss) in preferring one variety to the other. Trials conducted at research stations may not give a good estimate of the average over a given region, as the environmental factors at the research station may differ considerably from the average environmental factor in that region.

Secondly, the investigation of effects of factors, such as soil type, fertility and climatic factors, on the relative performance of the varieties, is not possible unless the trials are conducted in areas which differ considerably in such environmental factors. If the interactions of such environmental factors are found substantial, it may be possible to make specific recommendations for each environmental type. In ideal conditions (i.e. if the nature of interactions is precisely known) it may be possible to predict such differences in quantitative terms. The actual nature of such interactions is generally very complicated and it may not be easy to take into account all possible important factors; this reduces the reliability of predictions. However, in some cases, a few major factors (such as soil type, rainfall, or altitude) are found responsible for most of the differences in the performance of one variety relative to another, and then it may be possible to make specific recommendation for sub-regions, where one variety is likely to perform better than others.

At the preliminary trial stage no statistical problems arise, and almost all the varieties which conform to certain criteria such as mentioned above are passed on to the next stage (yield selection). The second stage of selection of the highest yielding varieties gives rise to interesting statistical methods. Finney (1958a, b; 1960), Curnow (1961) and

Bechhofer et al (1968) gives a good account of recent work on the problem of selection.

1.1 Scope of the present work

The present work is concerned with statistical aspects of the problems arising in the third or final stage where large scale trials are conducted to compare the performance of the new varieties, both among themselves and with standards. Two aspects which will be discussed in detail are as follows.

Firstly, attempts will be made to analyse and interpret the variety-environment interaction. Some methods will be discussed which may help in the interpretation of this interaction. These methods may also help us to make specific recommendations for particular types of environment.

The second problem of taking decision, which will be discussed in detail, arises as follows:

Suppose that variety trials are conducted continuously over years, and repeated over many sites. It is desirable to select the sites randomly from the region of interest but in practice it is not always possible to do it. For obtaining some results (given later) it has been necessary to assume that the sites have been selected randomly. It is hoped that a small departure from randomness does not bias the results seriously.

For simplicity design chosen is usually a randomised complete block with two to four blocks. To study the interaction between varieties and years trials are conducted for several years. The usual practice is to take each trial for a fixed number of years, perhaps three or four. In estimating the component of variety-year interaction it is assumed that year effects are random. The mean yields of varieties over three or four years are compared with the standard and others. If a variety is found superior it is recommended in preference to the standard. If a variety is promising but less clearly superior more trials are taken before deciding whether to recommend it. On the other hand if any variety is found inferior at the end of three to four years it is replaced by another new variety in the trial (passed on from the second stage). The variety which has been rejected can be excluded from the trials unless it is performing better than others in a sub-region. In the latter case this variety should continue only in trials taken in that sub-region.

At the end of every year it is required to take a decision, based on the data obtained up to that year. Possible decisions are:-

a) to reject a variety

b) to continue it for a few more years before taking a decision

c) to accept it (ready for recommendation for commercial purposes).

Two different approaches to taking a decision will be considered. Firstly the aim will be to maximise the expected gain to be incurred in accepting a good variety as early as possible. In this approach the cost of experimentation and the price of produce should be known, but in many cases this is not known. An alternative sequential approach of taking decision, based on probabilities, is also considered where the knowledge of cost parameters is not required.

These sequential methods may result in some saving of resources, in that the potentially bad varieties may be rejected, or good varieties accepted, before the completion of a fixed period. In the case of early rejection, other new varieties can be given a chance in such trials while in the case of acceptance, a good variety comes into commercial use earlier.

In this thesis, those sections, where the work is not referred to other authors, should be taken as original contribution. The thesis also refers to three published papers of the author; these are attached at the end of the thesis. The results obtained in these papers are used and quoted frequently in this thesis.

1.2 Review of previous work

Neyman et al (1935), while analysing the data of manurial trials conducted on a large number of sites, suggested the use of regression of yield differences among treatments on the environmental variables (corresponding to each site), a method used by Fisher (1924). In the absence of information about environmental variables for each site, Neyman et al (1935) considered the regression of each treatment yield on the yield of the other treatment from the same site. They discussed the use of such regression methods in predicting the expected gain associated with using a treatment for the given level of yield of the other. Yates (discussion of Neyman et al 1935), suggested the use of regression of the treatment yield on the average of yields of all the treatments at that site. Later, Yates and Cochran (1938) discussed in more detail the application of such regression methods to the analysis of groups of variety trials. Since then these methods have been used frequently in the analysis of genotype-environmental interaction and comprehensive references have been cited by Freeman and Perkins (1971). Perkins and Jinks (1968) and Jinks and Perkins (1970) have reformulated the techniques of Yates and Cochran (1938) for use in breeding work. Recently Shukla (1972a) and Hardwick and Wood (1972) have shown the analogy between the technique of Yates and Cochran (1938) and the non-additive model in the analysis of variance introduced by Tukey (1949), and generalised by Mandel (1961).

Another important aspect of the analysis of series of experiments is the estimation of the average yield of varieties, and the test of differences between them when variances at each site are different. Cochran (1937, 1954) discussed the efficiencies of different methods of estimation, when the precisions of treatment means at sites are different. Cochran (1937) and Cochran and Cox (1957) have also discussed the test of significance of interaction when the variances at different sites differ.

When the site effects are assumed random, and if variances and covariances cannot be assumed homogeneous, then the usual analysis of variance test (F-test) cannot be applied for testing the equality of variety means. Graybill (1954) and Scheffé (1956) have suggested the application of the T^2 - statistic for testing equality of means when the number of sites is greater than the number of varieties. When the number of sites is not much larger than the number of varieties then the test based on the T^2 - statistic is not likely to be very powerful. Caliński (1966) has studied the effects of inequality of variances and covariances, and suggested an approximate F-test using the results of Box (1954a, 1954b). He has also discussed the effects of inequality of within site variances on the F-test.

Gomes and Guimarães (1958) have considered the combined analysis of a group of experiments when individual experiments are laid out in randomised blocks and all of them have a number of treatments in common, the rest differing from experiment to experiment. Pavate (1961) has considered the combined analysis of balanced incomplete block designs with a number of treatments in common. Tyagi et al (1970) have considered the analysis of groups of experiments when treatments have a factorial structure.

As far as efficient allocation of resources at the yield selection stage (second stage) is concerned, Finney (1958a, 1960) gives a good description of statistical methods. The problem is to select a small number of highest yielding varieties (on the average) from a large number of varieties, in a given number of stages. Finney's (1958a) results give the approximate proportion of varieties to be continued from one stage to the next. Under the assumption of normality of yields, he finds the symmetric scheme (i.e. taking same proportions and allocating same resources at each stage) is nearly optimum, in the sense that this provides the maximum advance in average yield. His results do not utilise the information from previous stages of the selection trial, and only the information from the current stage is used in taking the decision to continue (or discontinue) that variety in the next stage. Curnow (1961) has extended Finney's (1958a, 1960) results to further stages and discussed the effect of non-normality, and showed that the results for the normal case hold good for any moderate departure from normality. Curnow (1961) has also discussed very briefly the gain from using information from the previous stages and the possible effect of variety by years interaction.

Finney (1958a, 1960) has introduced the term external economy, which takes into account the possible gain (or loss) by recommending varieties for commercial production, and this depends on many aspects, such as the area to be devoted to the new selected varieties, the average gain over the varieties which are already in use, and the cost of making such changes. He has tabulated the optimum values (i.e. which maximise the gain) of total area to be devoted for the trial of yield selection, and the number of varieties to be tested for one stage of selection. Curnow (1961) has tabulated similar results for a two-stage selection procedure. He has also discussed the optimum number of stages in the absence of variety-year interaction.

The problem of optimum allocation of resources with reference to external economy has also been discussed by Grundy et al (1956) for a two-stage selection procedure. Elandt (1963) has also discussed the allocation problem with reference to the distribution of allocation within trials (number of replications), and between trials (number of sites). Elandt (1963) has given expressions for the allocation of resources so that variance is minimised for a given cost; cost is minimised for a given variance; and the variance and total cost specified. Sprague and Federer (1951), by analysing variety trial data on corn, conducted over many locations and years, have estimated the variance components for year × variety and location × variety. Using these estimates, they have estimated the average genetic advance for varying number of locations, years and replication within location. They found that one replication per location with an increase in number of locations and years, is optimum for given number of total plots.

Bechhofer (1954) considered the problem of selecting the highest yielding variety out of a given number of varieties for a fixed number of stages. Bechhofer (1958) and Bechhofer and Blumenthal (1962) extended the results to the case when the number of stages, themselves, are based on the results of previous stages, thus giving a sequential procedure. Paulson (1964) also considered a simple sequential procedure for the above problem.

1.3 Layout of thesis and general notation

We shall denote by y_{ij} the mean yield of the variety j $(j = 1, 2, \dots, m)$ in the environment i $(i = 1, 2, \dots, n)$. These means are based on r replications. We assume that the parameters for environments are selected randomly from a population of environments with mean zero and variance σ_e^2. Thus we have a two-way table Y of means

$$\underset{n\times m}{Y} = (y_{ij})$$

of order $n \times m$ available. We shall represent this by the model

$$y_{ij} = \mu + e_i + v_j + \eta_{ij} + \xi_{ij} \tag{1.3.1}$$

where η_{ij} is the interaction component of the variety j with the environment i and ξ_{ij} is the within environment error component. It will be assumed that ξ_{ij} is independently and normally distributed with mean zero and variance σ_0^2 for all varieties and environments. It is also assumed that a good estimate, $\hat{\sigma}_0^2$, of σ_0^2 is available on n_0 d.f. which is independent of y_{ij}'s. μ is the overall mean; v_j is the effect of the variety j; e_i is the random effect of the environment i.

In Chapter 2 we consider a possible relation between η_{ij} and e_i of the type

$$\eta_{ij} = \beta_j e_i + \epsilon'_{ij} \quad . \tag{1.3.2}$$

The problem of estimating the β_j and testing their equality has been discussed in detail.

In some cases additional information about some external variables about environments such as rainfall, altitude etc. is also available. We shall denote the value of the k^{th} external variable at the site i by z_{ik} ($i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, p$). This information may be available in the form of a $n \times p$ matrix Z.

$$\underset{n\times p}{Z} = (z_{ik}) \quad .$$

A relationship between Y and Z of the type

$$y_{ij} = \mu + e_i + v_j + b_{1j} z_{i1} + b_{2j} z_{i2} + \cdots + b_{pj} z_{ip} + \epsilon'_{ij} + \xi_{ij} \qquad (1.3.3)$$

after substituting

$$\eta_{ij} = b_{1j} z_{i1} + b_{2j} z_{i2} + \cdots + b_{pj} z_{ip} + \epsilon'_{ij}$$

will be investigated in Chapter 3. The problem of estimating the b_{kj} and testing their equality will be considered in detail.

It is also of interest to estimate and compare the variability of individual varieties in addition to the additive common variability present due to environment effects. We define

$$\underset{i}{E}\,(\eta_{ij}^2) = \sigma'^2_j \qquad (1.3.4)$$

and assume that

$$\underset{i}{E}(\eta_{ij}) = 0 \text{ and } \underset{i}{E}(\eta_{ij}\eta_{ij'}) = 0 \text{ for } j \neq j' .$$

Put

$$\epsilon_{ij} = \eta_{ij} + \xi_{ij}$$

$$V(\epsilon_{ij}) = \sigma_j'^2 + \sigma_o^2 = \sigma_j^2 \text{ , say .} \qquad (1.3.5)$$

The problem of estimating the σ_j^2 and testing their equality will be considered in Chapter 4.

An empirical method of describing η_{ij} as the sum of the products of row and column terms will be considered in Chapter 5. A graphical method will be proposed. This technique may be of great help in summarising data from a large number of environments.

A sequential decision procedure for stopping further trials will be considered in Chapter 6.

The approach considered in Chapter 6 requires the knowledge of cost parameters. In some cases it may be difficult to obtain. An alternative approach of taking decision sequentially, based on probabilities, has been described in Chapter 7.

List of symbols which remain consistent throughout the subsequent chapters:

n = Number of environments.

m = Number of varieties.

y_{ij} = Mean yield of the variety j at the environment i .

ε_{ij} = Component of experimental error (within environment) associated with y_{ij} .

η_{ij} = Component of interaction for the variety j at the site i .

$\sigma_o^2 = V(\varepsilon_{ij})$

$\sigma_j'^2 = V(\eta_{ij})$

$\sigma_j^2 = \sigma_j'^2 + \sigma_o^2$

e_i = Effect of the environment i .

v_j = Effect of the variety j .

μ = General mean.

z_{ik} = Value of the external variable k at the site i .

$\underset{n\mathrm{x}m}{Y} = (y_{ij})$

$\underset{n\mathrm{x}p}{Z} = (z_{ik})$

This does not give a complete list of all symbols as the meaning of other symbols differ from chapter to chapter. Such symbols are defined in each chapter separately.

CHAPTER 2

RELATIONSHIP BETWEEN INTERACTION AND ENVIRONMENTAL EFFECTS

2.0 Introduction and development of previous work

Neyman et al (1935) analysed data from manurial trials conducted in many locations. In a row × column arrangement (y_{ij}) rows will be taken as locations and columns as manurial treatments. The trials which they analysed had three manurial treatments and they considered all three pairs of treatments, separately, as follows. They regressed the yield of one treatment on another e.g. y_{i1} on y_{i2}. This enabled them to calculate the expected gain (or loss), to be incurred by using treatment 1 instead of treatment 2 for a given level of the yield of second treatment. They also considered a method of calculating confidence limits for such gains.

In the discussion of Neyman et al (1935), Yates suggested the use of regression of individual variety yields (y_{ij}) on the corresponding environmental means $(\bar{y}_{i.})$. Yates and Cochran (1938) followed up this proposal in some detail. They did not explicitly give any mathematical model but appear to consider the model (1.3.2) of Section 1.3. By substituting (1.3.2) in (1.3.1) one obtains

$$y_{ij} = \mu + v_j + (1 + \beta_j)e_i + \epsilon'_{ij} + \xi_{ij} \, . \qquad (2.0.1)$$

Yates and Cochran (1938) estimated $(1 + \beta_j)$ by the regression of y_{ij} on $\bar{y}_{i.}$. They suggested that the mean of all the varieties at a site $(\bar{y}_{i.})$ gives an indication of fertility (productivity) of that site in the absence of any other information about that site. Consider the difference of yields of variety j and j' at the site i using the model (2.0.1)

$$y_{ij} - y_{ij'} = v_j - v_{j'} + (\beta_j - \beta_{j'})e_i + \delta_{ij} - \delta_{ij'} \qquad (2.0.2)$$

where

$$\delta_{ij} = e'_{ij} + \xi_{ij} .$$

Assuming that the departure from linear regression considered in (2.0.1) is small ($\delta_{ij} - \delta_{ij'}$ is small), then the difference between the yields of two varieties at the i^{th} site ($y_{ij} - y_{ij'}$) deviates from the mean difference ($v_j - v_{j'}$) by an amount depending upon ($\beta_j - \beta_{j'}$) and e_i. If $\beta_j - \beta_{j'}$ is positive then as e_i (productivity) increases, the difference $y_{ij} - y_{ij'}$ increases or decreases according to the sign of $v_j - v_{j'}$. Therefore the relative magnitude of β_j and $\beta_{j'}$ gives an idea about the relative performance of the varieties at high and low productive environments. If β_j and $\beta_{j'}$ are equal then the differences tend to remain constant. These remarks hold only when the difference between δ_{ij} are relatively small.

In the same way consider the deviation of the yield of variety j at the site i from the mean over all sites ($y_{ij} - \bar{y}_{.j}$)

$$y_{ij} - \bar{y}_{.j} = (1 + \beta_j)(e_i - \bar{e}) + \delta_{ij} - \bar{\delta}_{.j} .$$

Now from the above equation it is evident that if $(1 + \beta_j)$ is positive and large then $y_{ij} - \bar{y}_{.j}$ increases as $(e_i - \bar{e})$ increases and decreases as $(e_i - \bar{e})$ decreases. If $(1 + \beta_j)$ is very near to zero then $y_{ij} - \bar{y}_{.j}$ is not affected at all by the variation in the site effects. In the same way if $(1 + \beta_j)$ is negative then $y_{ij} - \bar{y}_{.j}$ decreases as $(e_i - \bar{e})$ increases and increases when $(e_i - \bar{e})$ decreases.

This method has been used very often in the past decade and a comprehensive list of references has been cited by Freeman and Perkins (1971). They have pointed out that the above mentioned regression method for estimating $(1 + \beta_j)$ is not strictly statistically valid because of inherent dependence of $\bar{y}_{i.}$ on y_{ij}. We shall see later this gives an inconsistent estimator of the parameters. They have suggested that this difficulty can be overcome by raising some additional varieties in each environment thus giving an independent measure of environment. Regressions on this independent measure are statistically valid.

Shukla (1972a) has shown, however, that the hypothesis of equality of the β_j can be tested with strict validity by considering the model in (2.0.1) as a non-additive model and using the arguments of Tukey (1949) and Mandel (1961).

Essentially the model given in (2.0.1) is a non-linear model and estimates of the parameters can be obtained by maximum likelihood. The estimation problem will be considered in Section 2.1. Section 2.2 deals with the test of hypothesis of equality of β_j. Section 2.3 deals with the relationship of the above work with that of the other authors. Section 2.4 considers the problem of prediction and Section 2.5 deals with a worked out example.

2.1 Estimation of parameters

We start with the model in (2.0.1)

$$y_{ij} = \mu + v_j + (1 + \beta_j)e_i + \delta_{ij}. \tag{2.1.0}$$

At present we wish to estimate the parameters $(1 + \beta_j)$ and σ_e^2 . We shall work with the sample variance-covariance matrix S given by

$$S = (s_{jj'})$$

where

$$s_{jj'} = \sum_i (y_{ij} - \bar{y}_{.j})(y_{ij'} - \bar{y}_{.j'})/(n - 1) .$$

S is independent of parameters μ and v_j as the corresponding population variance-covariance matrix is given by

$$\Sigma = LL' + \sigma_\delta^2 I \tag{2.1.1}$$

where L is a vector of order m such that

$$L' = \left[(1 + \beta_1)\sigma_e, (1 + \beta_2)\sigma_e, \ldots, (1 + \beta_m)\sigma_e\right] .$$

There is an indeterminacy between $(1 + \beta_j)$ and σ_e as by suitably multiplying and dividing $(1 + \beta_j)$ and σ_e by a constant Σ remains unaffected. Under these circumstances we can estimate the parameters uniquely only under a constraint on $(1 + \beta_j)$ which we shall discuss later. Note that the simple form of Σ gives the following expressions for $|\Sigma|$ and Σ^{-1},

$$|\Sigma| = (1 + \frac{L'L}{\sigma_\delta^2})\ \sigma_\delta^{2m}$$

$$\Sigma^{-1} = \frac{1}{\sigma_\delta^2}\left(I - \frac{LL'}{\lambda^2}\right)$$

where

$$\lambda^2 = L'L + \sigma_\delta^2 \text{ .} \tag{2.1.2}$$

The log likelihood $L(\Omega)$ of S (using Wishart's distribution) can now be written as

$$2L(\Omega) = -(n-1)\left[\log|\Sigma| + \operatorname{tr}(S\Sigma^{-1})\right] + \text{constant} \tag{2.1.3}$$

and using the expressions given above it reduces to

$$2L(\Omega) = -(n-1)\left[\log\left(1 + \frac{L'L}{\sigma_\delta^2}\right) - \frac{L'SL}{\lambda^2\sigma_\delta^2} + m\log(\sigma_\delta^2) + \frac{\operatorname{tr}(S)}{\sigma_\delta^2}\right] + \text{constant} \text{ .} \tag{2.1.4}$$

Now maximising (2.1.4) with respect to L subject to condition (2.1.2) we obtain

$$SL = \hat{\lambda}^2 L$$

where

$$\hat{\lambda}^2 = \hat{L}'\hat{L} + \sigma_\delta^2 \text{ .}$$

Thus $\hat{\lambda}^2$ is the largest latent root of S and $\hat{L}$ can be obtained as

$$\hat{L} = (\hat{\lambda}^2 - \sigma_\delta^2)^{\frac{1}{2}}\hat{Q} \tag{2.1.5}$$

where $\hat{Q}$ is the corresponding latent vector satisfying the condition

$$\hat{Q}'\hat{Q} = 1 \; .$$

σ_δ^2 can also be estimated by maximising (2.1.4) w.r. to σ_δ^2. This gives an additional equation

$$\sigma_\delta^2 = \left[\mathrm{tr}(S) - \hat{\lambda}^2 \right] /(m-1) \; . \qquad (2.1.6)$$

The estimate of σ_δ^2 obtained here should be substituted in (2.1.5).

To obtain the estimate of $(1 + \beta_j)$ we impose an additional condition on $(1 + \beta_j)$ i.e.

$$\sum_j (1 + \beta_j) = m \; . \qquad (2.1.7)$$

The values of $(1 + \hat{\beta}_j)$ satisfying this condition can now be obtained. Let $\hat{l}_j$ and $\hat{q}_j$ be the jth element of vector $\hat{L}$ and $\hat{Q}$ then

$$(1 + \hat{\beta}_j) = \frac{\hat{l}_j}{\hat{\bar{l}}} = \frac{\hat{q}_j}{\hat{\bar{q}}} = \hat{\theta}_j, \text{ say,} \qquad (2.1.8)$$

i.e. the ratio of the element of the latent vector corresponding to the largest latent root to the mean of all the elements of that latent vector. Because of indeterminacy absolute value of $(1 + \beta_j)$ cannot be estimated and only the ratio of $(1 + \beta_j)$ remains constant. In this case we have overcome this problem by putting a constraint on β_j by making them add to zero. Under different constraints different values will be obtained but ratio will remain the same. The estimate of σ_e^2 is obtained as

$$\hat{\sigma}_e^2 = \left(\hat{\lambda}^2 - \hat{\sigma}_\delta^2\right) / \sum_j (1 + \hat{\beta}_j)^2 .$$

2.2 Test of hypothesis

It is of great interest to test the equality of the $1 + \beta_j$ in the model (2.0.1) which is equivalent to testing whether all $\beta_j = 0$ or not. If they are found equal then there is hardly any justification in characterising varieties on the basis of their β_j. However, if significant differences are found among β_j then it may be of further interest to compare a pair of β_j and give confidence limits for each β_j. In the following we shall give an exact test and a likelihood ratio test.

Here we use the results of Kshirsagar (1961) for testing the proportionality of the coefficients of the first principal component to an arbitrary chosen vector. In our notation the result can be stated as follows. Suppose we want to test the proportionality of the first principal component of S to an arbitrary given vector θ_0 where

$$\theta_0' = (\theta_{01}, \theta_{02}, \ldots, \theta_{0m}) .$$

This may be tested by calculating F^*, where

$$F^* = \frac{\theta_0' S^2 \theta_0}{\theta_0' S \theta_0} - \frac{\theta_0' S \theta_0}{\theta_0' \theta_0} \tag{2.2.1}$$

and under the null hypothesis

$$F^*/\sigma_\delta^2$$

will have a χ^2 distribution with $(m-1)$ D.F. .

We are particularly interested in testing whether or not all θ_j are equal. The value of F^* in (2.2.1) after putting all θ_{0j} equal to 1 is

$$F^* = \frac{\sum_j \left[\sum_i (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})\right]^2}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2} \quad . \tag{2.2.2}$$

Moreover, it is not difficult to see that

$$Q = \left[\sum_{i,j} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - F^*\right] / \sigma_\delta^2 \tag{2.2.3}$$

is distributed as χ^2 on $(m-1)(n-2)$ D.F. when all β_j are zero. By using arguments similar to Mandel (1961) (or Tukey, 1949) it can be shown that expressions in (2.2.2) and (2.2.3) are independently distributed under the null hypothesis thus providing a test criterion R

$$R = \frac{F^*/(m-1)}{\left[\sum_{i,j} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - F^*\right] / (m-1)(n-2)} \quad . \tag{2.2.4}$$

When the null hypothesis is true, R follows the F-distribution with $(m-1)$ and $(m-1)(n-2)$ D.F. . This provides an exact test of the hypothesis of the equality of all β_j to 0 . It is not easy to find the distribution of R under the alternative hypothesis.

For comparison purposes we shall also consider the likelihood ratio test for the null hypothesis (H_0) that all β_j are zero. Under H_0 the log-likelihood $L(w)$ is given by

$$2L(w) = -(n-1)\left[\log|\Sigma_0| + \mathrm{tr}(S\,\Sigma_0^{-1})\right] + \text{constant}$$

where

$$\Sigma_0 = \sigma_{0e}^2\,11' + \sigma_{0\delta}^2\,I$$

$$|\Sigma_0| = (1 + m\frac{\sigma_{0e}^2}{\sigma_{0\delta}^2})\sigma_{0\delta}^{2m}$$

$$\Sigma_0^{-1} = \frac{1}{\sigma_{0\delta}^2}\,(I - \frac{\sigma_{0e}^2\,11'}{m\sigma_{0e}^2 + \sigma_{0\delta}^2})\ .$$

Here σ_{0e}^2 and $\sigma_{0\delta}^2$ are the values of the parameters σ_e^2 and σ_δ^2 under the null hypothesis H_0. It is not difficult to see that M.L. estimates are given by

$$\hat{\sigma}_{0e}^2 = \sum_{j\neq j'} s_{jj'}/m\,(m-1)$$

$$\hat{\sigma}_{0\delta}^2 = \left[\sum_j s_{jj} - \sum_{j\neq j'} s_{jj'}/(m-1)\right]/m\ .$$

Using the above estimates likelihood ratio statistic

$$-2\log\lambda^* = -2[L(w) - L(\Omega)]$$

can be calculated. Under the null hypothesis $-2\log\lambda^*$ should have a χ^2 distribution (asymptotically) on $(m-1)$ D.F.

The statistic R in (2.2.4) and $-2\log\lambda^*$ of LR test give different test statistic. R gives an exact test while LR is only an approximate but asymptotically more powerful. In the present case of variety trials it is unlikely that these two statistics will differ much at least when n is large.

The problem of finding the standard errors of $(1 + \hat{\theta}_j)$ is not easy but confidence interval for each can be obtained with the help of (2.2.1). This involves finding the roots of a fourth degree polynomial for each θ_j ;

$$(\theta_j^{*\prime} S^2 \theta_j^*)(\theta_j^{*\prime} \theta_j^*) - (\theta_j^{*\prime} S \theta_j^*)^2 - F_{(1,N)} \hat{\sigma}_\delta^2 (\theta_j^{*\prime} S \theta_j^*)(\theta_j^{*\prime} \theta_j^*) = 0 \qquad (2.2.5)$$

where

$$\theta_j^{*\prime} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \theta_j, \ldots, \hat{\theta}_m) \text{ and } N = (m-1)(n-2) .$$

Suppose that the equation (2.2.5) corresponding to θ_j is given by

$$a_0 \theta_j^4 + a_1 \theta_j^3 + a_2 \theta_j^2 + a_3 \theta_j + a_4 = 0 \qquad (2.2.6)$$

where a_0, a_1, a_2, a_3 and a_4 are the coefficients of the polynomial in θ_j. Assume the values of the four roots of this equation are real and given by

$$\theta'_{j1}, \theta'_{j2}, \theta'_{j3}, \theta'_{j4} .$$

The confidence intervals can be obtained by pairing the roots in such a way that for all the values of θ_j, the L.H.S. of (2.2.6) is less than zero, where

$$\theta'_{jl} \leq \theta_j \leq \theta'_{jl'}, \qquad (l \neq l' = 1,2,\ldots,4).$$

The pair which contains the estimated value of θ_j obtained from (2.1.8) gives the appropriate confidence interval. This method may fail when all the roots are imaginary. In some cases it may give disjoint confidence intervals. However, in the examples considered at the end of this chapter, regular intervals containing the estimated values occur.

There is also considerable interest in testing the equality of $(1 + \beta_j)$ for a pair of varieties, say, j and j'. This can be done by calculating confidence limits of $\frac{1 + \beta_j'}{1 + \beta_j}$ by putting $\hat{\theta}_j = 1$ in (2.2.5) and calculating limits of θ_j' from (2.2.5). If the confidence interval for $\frac{1 + \beta_j'}{1 + \beta_j}$ contains 1 then both can be taken as equal. This shall be discussed in the example.

2.3 Relationship with other work and discussion

The essential difference between the approach for estimation of $1 + \beta_j$ given in Section 2.1 and the method of Yates and Cochran (1938) is as follows. Yates and Cochran (1938) take the mean of all varieties as a measure of productivity of a particular environment, thus giving equal weights to all varieties. In cases where varieties are of different sensitivities (in relation to environments) it may be more appropriate to give the weights proportional to their sensitivities in calculating the measure of environments. In the present approach this has been done by calculating sensitivities $(1 + \beta_j)$ and then calculating measure of environments as a weighted mean, and thus proceeding iteratively.

It is not appropriate to estimate the random effect e_i. In cases where a similar model (2.0.1) is considered, as in Factor Analysis $(1 + \beta_j)$

are termed 'loadings' and e_i 'scores'. The e_i are estimated after substituting for $(1 + \beta_j)$. By following analogous arguments to Factor Analysis and using equation (2.1.0) and substituting the estimates $(1 + \beta_j)$ as obtained in (2.1.8) we obtain

$$y_{ij} = \bar{y}_{.j} + \hat{\theta}_j e_i + \delta_{ij} .$$

Now multiplying both sides by θ_j $(j = 1,2,\ldots,m)$ we obtain an estimate E_i of e_i by ignoring δ_{ij}

$$E_i = \frac{\Sigma(y_{ij} - \bar{y}_{.j})\hat{\theta}_j}{\Sigma\, \hat{\theta}_j^2} .$$

The weights $(1 + \beta_j)$ are such that the varieties which are most variable over environments get most weight. This may not always be very desirable. For example, if one particular variety is very prone to lodging, and the sites selected differ appreciably for lodging (which in turn affects yield), then this variety will get a relatively large weighting in the measure of productivity of sites. In this case the indices E_i will not be a good measure of productivity as far as average varieties are concerned, but may be a good measure of lodging as measured by yield. In Yates and Cochran's method the productivity index for each environment is calculated by giving equal weights (by $\bar{y}_{i.}$) to all the varieties. Unless the varieties are not very different in their interactions the estimates of $(1 + \beta_j)$ obtained by both methods will not differ very much.

The coefficients $(1 + \beta_j)$ have been calculated under the assumption that all the varieties are equally variable within environment. However, if within environment the error variance is larger for some varieties than others then more general model may be necessary for estimating variances as well as coefficients. The methods will then be similar to that used in Factor Analysis. The model (2.0.1) can be taken as a particular case of Factor Analysis model (Lawley and Maxwell, 1971) with only one factor and variances of δ_{ij} equal for all j.

The use of estimates obtained by Yates and Cochran has come under criticism because of dependence of y_{ij} on $\bar{y}_{i.}$ and their inconsistency. As far as inconsistency is concerned this can be removed as suggested by Yates (discussion of Neyman et al 1935). Usually the bias is of the order of $\frac{\sigma_\delta^2}{\sigma_e^2}$ and when environments are relatively much more variable, the bias is negligible.

Williams (1952) used the multiplicative model for interpretation of interactions. He used a least squares method for his fixed effect model and obtained similar estimates of $(1 + \beta_j)$ as that obtained here in (2.1.8). In our case of mixed model we had to consider a maximum likelihood method but the final results obtained are similar.

As far as testing the equality of the $(1 + \beta_j)$ is concerned the statistic R given in (2.2.4) is the same as that proposed by Mandel (1961) for testing the presence of non-additivity in the model (2.0.1) which is a generalisation of Tukey's (1949) statistic. Shukla (1972a) and Hardwick and Wood (1972) have pointed out that the statistic R is the same as obtained through the method of Yates and Cochran.

There is also a very close relationship between the method described in this chapter for estimating and interpreting $(1 + \beta_j)$ and the method for estimating interaction between qualitative and quantitative factors. Fisher (1949) considered an experiment comparing two qualities of nitrogen fertilisers, with single and double applications. He argued that a reasonable hypothesis might suppose that the difference between yields is twice as much at the double as that at the single dressing. And thus he calculated the effect of quality by giving double weight to the yield at double dressing than at single dressing instead of simple mean which gives equal weights. He calculated interaction by taking an orthogonal contrast to that of the effect of quality.

A similar interpretation can be made in variety trial work by replacing types of nitrogen by environments, and doses of nitrogen by varieties. As varieties are qualitative factors the weights $(1 + \beta_j)$ are calculated from the data itself so as to give best measures of environments instead of an arbitrary chosen constant as done by Fisher (1949).

2.4 Estimation and prediction

When the interaction component η_{ij} is not related to the additive environment component e_i then the yield y_{ij} differs from environment to environment by a constant additive environment component e_i and a random interaction component η_{ij} which cannot be predicted. However, when η_{ij} is related to e_i by a relationship of the type described in (2.1.0) then y_{ij} at different sites will differ systematically depending on the magnitudes of the coefficient β_j and the additive environmental component e_i. If e_i is given for any particular site then it is not difficult to estimate y_{ij} and the differences $y_{ij} - y_{ij'}$. As we have seen in the preceding sections, the estimate of e_i itself depends on the yield of all the varieties (including the yield of the variety which we want to estimate) and therefore the relationship in (2.1.0) is not very helpful in predicting y_{ij}.

Suppose we have the yield of m varieties available at the n sites. If the tests described in this chapter suggest that the relationship described in (2.1.0) is more appropriate then sometimes it may be of interest to estimate (or predict) the yield of a variety, say, the first one given the yield of the remaining varieties at a site where the first variety is not grown. This may be used in estimating and predicting the difference between the first variety and one of the remaining varieties. This can be done by regressing y_{i1} on $y_{i2},\ldots,y_{im}$. The prediction equation for y_{i1} given $y_{i2},\ldots,y_{im}$ is given by

$$\hat{y}_{i1} = \hat{a}_0 + \hat{a}'Y_i \tag{2.4.1}$$

where

$$\hat{a}_0 = \bar{y}_{.1};\ \hat{a} = S_{22}^{-1}S_{12};\ Y'_i = (y_{i2} - \bar{y}_{.2},\ldots,y_{im} - \bar{y}_{.m})$$

and

$$S = \begin{bmatrix} s_{11} & S'_{12} \\ S_{12} & S_{22} \end{bmatrix} .$$

Thus for any given value of Y_i we can estimate y_{i1} using the equation (2.4.1). A case of special interest occurs when only two varieties are involved and, given the yield of the second variety at the ith site, we wish to estimate the yield of the first variety, although this first variety ought not to have been grown on that site.

Using the equation (2.4.1) the estimate can be easily obtained, namely

$$\hat{y}_{i1} = \bar{y}_{.1} + \frac{s_{12}}{s_{22}} (y_{i2} - \bar{y}_{.2}) \quad . \tag{2.4.2}$$

Additional expected yield by using the first variety instead of the second is given by

$$\text{Expected gain} = \hat{y}_{i1} - y_{i2} = (\bar{y}_{.1} - \frac{s_{12}}{s_{22}} \bar{y}_{.2}) + y_{i2} (\frac{s_{12}}{s_{22}} - 1) \; . \tag{2.4.3}$$

The first part of the expression on the R.H.S. of the above relationship does not vary with variation in sites, whereas the second part does. If the regression coefficient s_{12}/s_{22} is greater than one, then the second part increases as y_{i2} increases; if it is less than one, it decreases as y_{i2} increases. If the regression coefficient is very nearly equal to one then this part vanishes and the expected gain remains constant. Thus, the above relationship can be used to estimate the gain and this provides some guidance towards choosing a suitable variety for a given site. The variance of this gain can be calculated without much difficulty.

Let us examine the relationship between this approach and that described by the model (2.1.0). By using the population values of S_{22} and S'_{12} in the expression $\hat{a}$ and using the form of Σ given in (2.1.1) we obtain the limiting value of $\hat{a}$

$$\hat{a} \xrightarrow[n \to \infty]{\text{in limit}} \Sigma_{22}^{-1} \Sigma_{12}$$

where

$$\Sigma = \begin{bmatrix} \sigma_{11} & \Sigma'_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} = LL' + \sigma_\delta^2 I \, .$$

Using the expression of Section 2.1 it is not difficult to see that

$$\Sigma_{22}^{-1} \Sigma'_{12} = \frac{(1 + \beta_1)}{\sum_{j=2}^{m} (1+\beta_j)^2 + \sigma_\delta^2/\sigma_e^2} \begin{bmatrix} 1 + \beta_2 \\ \vdots \\ 1 + \beta_m \end{bmatrix} . \qquad (2.4.4)$$

Now substituting this limiting value of $\hat{a}$ in (2.4.1) and taking the conditional expectation given e_i and using (2.1.0) we obtain

$$E(\hat{y}_{i1}/e_i) = \mu + v_j + (1 + \beta_1)\, e_i \; \frac{\sum_{j=2}^{m} (1 + \beta_j)^2}{\sum_{j=2}^{m} (1 + \beta_j)^2 + \sigma_\delta^2/\sigma_e^2} \, . \qquad (2.4.5)$$

Therefore $\hat{y}_{ij}$ is an inconsistent estimator of the expected yield at the ith site, namely $\mu + v_j + (1 + \beta_j)e_i$.

This inconsistency has been introduced because of the errors in the predictor variables $y_{i2}, \ldots, y_{im}$. Usually $\sigma_\delta^2/\sigma_e^2$ will be negligible in practice and $\hat{y}_{i1}$ thus provides a reasonable estimator.

When β_j are all zero then the coefficients $\hat{a}$ approach

$$\hat{a} = \frac{1}{m-1+\sigma_\delta^2/\sigma_e^2}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} \quad \xrightarrow[\text{as } \frac{\sigma_\delta^2}{\sigma_e^2} \to 0]{\text{is limit}} \quad \frac{1}{m-1}\begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix}$$

as expected.

2.5 An example

The data from co-ordinated variety trials on spring barley, conducted in Scotland in 1972, have been used to provide an example. Table 2.5.1 contains the data for eight varieties from twenty sites. The sites will be assumed to be selected randomly. At each site there were four replications but Table 2.5.1 gives only means (over replications) for each variety at each site. The data from each site were analysed separately and here we are concerned only with the combined analysis. The variances of individual trials did not differ very much and the average S.E. for means of four replications is given in the table. The data on external variates i.e. altitude, rainfall and sunshine from each site (wherever available) are also given.

We shall use the methods described in this chapter for estimation of parameters $(1 + \beta_j)$ and test for their equality. For comparison purposes we have also used the Yates and Cochran (1938) technique. If Y_C denotes the matrix Y after subtracting column means from each unit then

$$S = Y_C'Y_C/(n-1).$$

The largest latent root of the matrix S is

$$\lambda^2 = 10.193; \quad \% \text{ variation} = 92.16;$$

and about 92% of variation is accounted by the largest latent root. The corresponding latent vector is given in Table 2.5.2 under the heading $\hat{Q}$. The M.L. estimates of $1 + \beta_j$ are given under the heading $1 + \hat{\beta}$. Estimates of $1 + \beta_j$ obtained by the regression technique (Yates and Cochran) are given under the heading $1 + \beta^*$. Lower and upper 95% confidence limits of $(1 + \beta_j)$ are given under the heading $(1 + \hat{\beta})_L$ and $(1 + \hat{\beta})_U$, respectively. The estimate of σ_δ^2 used for calculating confidence limits is given

$$\hat{\sigma}^2_{\delta*} = \left[\sum_{i,j}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - F^*\right] /(m-1)(n-2) .$$

There is little difference in the estimates of σ_δ^2 given here and those obtained in (2.1.6).

Table 2.5.3 gives the analysis of variance table. The interaction sum of squares is partitioned into the sum of two components; (i) the sum of squares due to heterogeneity among regressions (F^*) and (ii) the sum of squares due to deviation from regression obtained by subtracting F^* from interaction sum of squares.

The sum of squares due to F^* is further partitioned into two components. The first is due to proportionality of β_j to v_j and is as given by Tukey (1949). This has been denoted by G

$$G = \frac{\left[\sum_{i,j} y_{ij}(y_i - \bar{y}_{.})(\bar{y}_{.j} - \bar{y}_{..})\right]^2}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2} .$$

The second component is the deviation from proportionality, obtained by subtracting G from F*. It must be noted that this sum of squares (F* and G) is not the sum of squares in the usual sense of least squares analysis but they have only been obtained for testing purposes. The sum of squares due to β_j estimated by the method of M.L. can be calculated by subtracting sum of squares due to sites (obtained from AOV.) from $(n-1)\ \lambda^2$ and is given by

$$\text{S.S. due to } \beta_j = 2.09 .$$

In this particular case this is not very different from the S.S. due to heterogeneity of regression F* (1.94) obtained for testing purposes.

For calculating likelihood ratio statistic the expressions $L(\omega)$ and $L(\Omega)$ can be further simplified and are given as follows.

$$2L(\omega) = -(n-1)\left[\log(\hat{\sigma}_{0\delta}^2 + m\hat{\sigma}_{0e}^2) + \overline{m-1}\log(\hat{\sigma}_{0\delta}^2) + \frac{tr(S)}{\hat{\sigma}_{0\delta}^2} - \frac{(\sum_{i,j} s_{ij})\hat{\sigma}_{0e}^2}{\hat{\sigma}_{0\delta}^2(\hat{\sigma}_{0\delta}^2 - m\hat{\sigma}_{0e}^2)}\right]$$

$$2L(\Omega) = -(n-1)\left[\log(\hat{\lambda}^2) + \overline{m-1}\log(\hat{\sigma}_{\delta}^2) + \frac{tr(S)}{\hat{\sigma}_{\delta}^2} + (1 - \frac{\hat{\lambda}^2}{\sigma_{\delta}^2})\right] .$$

M.L. estimates of parameters are given as

$$\hat{\sigma}_{\delta}^{2} = 0.124 \; ; \quad \hat{\lambda}^{2} = 10.193 \; ; \quad \hat{\sigma}_{e}^{2} = 1.246$$

$$\hat{\sigma}_{0\delta}^{2} = 0.139 \; ; \quad \hat{\sigma}_{0e}^{2} = 1.244$$

$$\mathrm{tr}(S) = 11.064 \; ; \quad \sum_{i,j} s_{ij} = 80.728$$

$$2L(\omega) = -19 \left[2.3116 - 13.8129 + 79.5970 - 71.6000\right] = 66.5825$$

$$2L(\Omega) = -19 \left[2.3214 - 14.6123 + 89.2258 - 81.2016\right] = 81.0673$$

$$-2 \log \lambda^{*} = 14.485 \sim \chi^{2}_{7} \; .$$

Hence it is significant at 5% probability level. As far as comparison with exact test is concerned the value of F obtained in Table 2.5.3 should be nearly equal to $\chi^2_7/7$ as $n \to \infty$,

$$F = 2.15 \; ; \quad \chi^2_7/7 = 2.069 \; .$$

This shows a good agreement.

For testing the equality of a pair of $1 + \beta_j$ we have calculated the confidence limits of ratio $\dfrac{1 + \beta_j}{1 + \beta_7}$ for all $j (\neq 7)$. The 95% confidence limits are given by:-

Varieties	1	2	3	4	5	6	7	8
Upper limit	1.340	1.340	1.317	1.289	1.511	1.342	-	1.622
Lower limit	0.969	0.972	0.948	0.917	1.126	0.974	-	1.235

Varieties 1, 2, 3, 4, and 6 contain one among their limits and therefore the corresponding $1 + \beta_j$ can be inferred to be not significantly different from $1 + \beta_7$. Confidence intervals of $1 + \beta_j$ for the varieties 5 and 8 do not contain one and therefore they differ significantly from $1 + \beta_7$.

As far as practical application of these results is concerned there appears to be heterogeneity among regressions and thus the differences among varieties depend to some extent upon the general productivity of sites. Universe is the highest yielding variety and as its coefficient is larger than one this shows that it is expected to do better than the other varieties at higher productive sites. As its general mean is the highest it is likely to do better than the other varieties in average productive sites but this may not be the best variety in very poor sites. Let us examine its relationship with Mazurka which is slightly lower in yield but has a considerably smaller regression coefficient than that of Universe. We shall calculate the regression of yield of Mazurka on the yield of Universe and then calculate the prediction equation.

$$\hat{y}_{i6} = 5.72 + .743\,(y_{i8} - 5.82)$$

$$\text{Expected gain} = \hat{y}_{i6} - y_{i8} = 1.396 - 0.257\,y_{i8}$$

Therefore, at the sites where Universe is expected to yield less than 5.43 it is advantageous to grow Mazurka. However, at more productive sites where yield of Universe is higher than 5.43 it is expected that Universe will provide a higher yield than Mazurka.

Variety Pegasus has a considerably lower regression coefficient than the other varieties. Unfortunately this variety has a small general mean. If this variety had a higher mean it is likely that at lower yielding sites it would have yielded higher than other varieties. From Table 2.5.1 it is apparent that this variety does not do very badly in comparison with other varieties in lower yielding sites. Therefore these coefficients $(1 + \beta_j)$, in addition to variety means, give a good guidance in recommending varieties for individual sites if some information on the general fertility level of the site is available.

TABLE 2.5.1

YIELD OF SPRING BARLEY (t/ha) IN CO-ORDINATED VARIETY TRIAL CONDUCTED IN SCOTLAND 1972

Geno-types Sites	Ymer 1	Gerkra 2	Gold-field 3	Imber 4	Maris Mink 5	Mazurka 6	Pegasus 7	Universe 8	Mean	Altitude (m.)	Rainfall (mm.)	Sunshine (hrs.)
1	6.88	7.11	6.80	6.58	7.05	6.98	6.61	7.62	6.95	61	-	-
2	6.46	7.07	7.24	6.77	7.58	7.32	6.92	8.16	7.19	97	-	-
3	3.62	4.14	4.35	3.74	4.80	3.79	3.88	4.57	4.11	58	-	-
4	5.50	5.42	5.17	5.02	5.64	5.38	5.30	6.04	5.43	46	-	-
5	5.67	5.81	6.01	5.44	6.68	6.30	6.05	6.98	6.12	61	-	-
6	5.14	5.39	6.44	5.98	5.72	6.25	5.97	6.22	5.89	180	492	930
7	6.70	6.68	6.30	6.27	6.78	6.81	6.27	7.21	6.63	30	205	901
8	5.97	5.84	5.92	5.48	6.22	5.95	5.42	6.49	5.91	90	212	740
9	6.25	6.21	6.49	6.18	7.03	6.56	5.82	6.48	6.38	27	309	893
10	5.54	5.51	4.99	5.03	5.07	5.78	5.43	5.54	5.36	61	257	887
11	6.22	5.93	5.58	6.01	5.55	5.99	5.87	6.08	5.90	137	233	846
12	4.80	4.94	5.28	4.58	5.76	5.22	4.96	5.47	5.13	21	234	959
13	5.62	5.54	5.48	5.33	6.07	5.94	5.47	6.16	5.70	76	212	973
14	6.28	6.16	6.30	6.22	4.57	6.03	6.29	5.82	5.96	40	528	976
15	5.84	5.74	5.74	5.94	5.64	5.80	5.71	5.86	5.78	15	426	955
16	6.73	6.82	7.38	7.16	6.86	7.17	6.82	7.27	7.03	30	426	955
17	4.54	3.99	4.35	4.91	3.67	4.35	4.74	3.39	4.24	105	477	923
18	4.58	4.74	4.51	5.32	4.71	4.97	4.81	4.01	4.71	45	502	834
19	4.10	3.76	3.77	4.46	3.68	4.76	4.46	4.08	4.13	34	343	849
20	2.46	3.04	3.32	2.28	2.25	3.03	3.01	2.95	2.79	180	481	939
Mean	5.44	5.49	5.57	5.44	5.57	5.72	5.49	5.82	5.57	69.70	355.80	904.00
S.E.					0.158				S.D.	47.61	119.98	62.74

TABLE 2.5.2

ESTIMATES OF $(1 + \hat{\beta}_j)$ AND 95% CONFIDENCE LIMITS

Variety Names	$\hat{\alpha}$	$(1 + \hat{\beta})$	$(1 + \beta^*)$	$(1 + \hat{\beta})_L$	$(1 + \hat{\beta})_U$
Ymer	0.344	0.977	0.980	0.823	1.138
Gerkra	0.344	0.978	0.978	0.825	1.138
Goldfield	0.337	0.958	0.959	0.805	1.118
Imber	0.328	0.933	0.940	0.779	1.094
Maris Mink	0.392	1.115	1.108	0.956	1.282
Mazurka	0.345	0.980	0.981	0.827	1.139
Pegasus	0.298	0.849	0.852	0.697	1.005
Universe	0.425	1.208	1.202	1.048	1.377

$\hat{\lambda}^2 = 10.193$

TABLE 2.5.3

AOV. OF SPRING BARLEY (t/ha)

Source	D.F.	S.S.	M.S.
Sites	19	191.57	10.08
Varieties	7	2.62	0.37
Sites × Varieties	133	18.47	0.14
Heterogeneity in regression (F*)	7	1.94	0.28*
Proportionality (G)	1	0.55	0.55
Deviation from proportionality	6	1.39	0.23
Deviation from regression	126	16.53	0.13
Total	159	212.66	

CHAPTER 3

ANALYSIS OF INTERACTION USING INFORMATION ON ENVIRONMENTAL VARIABLES

3.0 Introduction

In this chapter we shall use information provided by p external variables. This information is given in the form of a matrix Z of order $n \times p$. It is assumed that for each environment i, the z_{ik} $(k = 1,2,\dots,p)$ are independent of j $(j = 1,2,\dots,m)$; Z_{ik} may be taken as information on sunshine, rainfall and altitude etc. We shall use the model (1.3.3) i.e.

$$\eta_{ij} = b_{1j} z_{i1} + b_{2j} z_{i2} + \dots + b_{pj} z_{ip} + \varepsilon'_{ij} \qquad (3.0.1)$$

with η_{ij} defined as in (1.3.1). We require to make inferences about the parameters b_{kj}. A significant deviation of the value of b_{kj} from the mean (over j) shows that $(y_{ij} - \bar{y}_{i.})$ is dependent on the kth environmental variable z_{ik}. The sign and magnitude of b_{kj} give the direction and degree of dependence.

Neyman et al (1935), briefly considered the possibility of regressing yields on external variates. They suggested the use of methods analogous to those used by Fisher (1924). Freeman and Perkins (1971) considered the case when an independent measurement of environment is available and used regression techniques. Shukla (1972a) and Hardwick and Wood (1972) have considered the use of information on one or more external variables. The method used here is an extension of the method discussed in Shukla (1972a).

The problems of estimation and hypothesis testing are discussed in Section 3.1. Section 3.2 deals with the relationship of this work with the work of other authors and in Section 3.3 a worked example is discussed.

3.1 Estimation of regression parameters and test of hypothesis

By substituting (3.0.1) in (1.3.1) we obtain

$$y_{ij} = \mu + e_i + v_j + b_{1j}z_{i1} + b_{2j}z_{i2} + \dots + b_{pj}z_{ip} + e'_{ij} + \varepsilon_{ij} \; . \qquad (3.1.1)$$

For convenience each z_{ik} is assumed to be measured from the mean over all environments and thus

$$\sum_i z_{ik} = 0 \; ; \quad (k = 1, 2, \dots, p) \; .$$

To standardise, each z_{ik} has been divided by $\sqrt{\sum_i z_{ik}^2}$. Without any loss in generality, the mean of all b_{kj} can be absorbed in e_i so that

$$\sum_j b_{kj} = 0 \; ; \quad (k = 1, 2, \dots, p) \; .$$

It can be easily demonstrated that the least squares estimate of b_{kj}, here denoted by $\hat{b}_{kj}$, is given by $(k, j)^{th}$ element of $\hat{B}$, where

$$\hat{B} = (Z'Z)^{-1}Z'R, \qquad (3.1.2)$$

$$Z = (z_{ik}) \; ; \quad R = (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{\bar{y}}_{..}) \; .$$

It is not difficult to observe that

$$\sum_j \hat{b}_{kj} = 0$$

from the properties of R . The sum of squares due to fitted coefficients $\hat{b}_{kj}$ is given as usual by

$$\mathrm{Tr}(R'Z(Z'Z)^{-1}Z'R) .$$

Now write

$$\hat{B} = (\hat{B}_1, \hat{B}_2, \ldots, \hat{B}_m)$$

where each $\hat{B}_j$ is a vector of order p such that

$$\hat{B}_j' = (\hat{b}_{1j}, \hat{b}_{2j}, \ldots, \hat{b}_{pj}) .$$

It can be shown that

$$V(\hat{B}_j) = \sigma_\delta^2(1 - \frac{1}{m})(Z'Z)^{-1} \qquad (3.1.3)$$

$$\mathrm{Cov}\ (\hat{B}_j, \hat{B}_{j'}) = -\frac{\sigma_\delta^2}{m}(Z'Z)^{-1}; \quad (j \neq j' = 1, 2, \ldots, m) \qquad (3.1.4)$$

where σ_δ^2 is the variance of $\epsilon'_{ij} + \xi_{ij}$.

The estimated difference between variety j and j' at the site i is given by

$$\bar{y}_{.j} - \bar{y}_{.j'} + (\hat{B}'_j - \hat{B}'_{j'})\ Z_i \ ;$$

where

$$Z_i' = (z_{i1}, z_{i2}, \cdots z_{ip}) .$$

The variance of the estimated difference can be shown to be

$$2\sigma_\delta^2 \left[\frac{1}{n} + Z_i' (Z'Z)^{-1} Z_i\right] .$$

When $p = 1$, estimators for a single set of regression coefficients (dropping the subscripts of b's and z's) b_j are given by

$$\hat{b}_j = \sum_i (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})z_i = \sum_i (y_{ij} - \bar{y}_{i.})z_i \quad . \qquad (3.1.5)$$

As $\sum_i z_i^2 = 1$, it follows that

$$V(\hat{b}_j) = \sigma_\delta^2 \left(1 - \frac{1}{m}\right); \quad \text{cov}\,(\hat{b}_j, \hat{b}_{j'}) = -\frac{\sigma_\delta^2}{m} \text{ for } j \neq j' .$$

The estimated difference between variety j and j' at the site i is given by

$$\bar{y}_{.j} - \bar{y}_{.j'} + (\hat{b}_j - \hat{b}_{j'})\, z_i \quad ;$$

with variance

$$2\sigma_\delta^2 \left(\frac{1}{n} + z_i^2\right) .$$

The analysis of variance showing the break up of the interaction sum of squares is given in Table 3.1.

TABLE 3.1

AOV OF INTERACTION FOR EXTERNAL VARIABLES

Source	D.F.	S.S.
Heterogeneity in regression	$p(m-1)$	$Tr(R'Z(Z'Z)^{-1}Z'R)$
Deviation from regression	$(m-1)(n-p-1)$	By subtraction
Total	$(n-1)(m-1)$	Interaction S.S.

The sum of squares due to heterogeneity in regressions can be tested against the sum of squares for deviations from regression. The deviations from regression can be tested against $\hat{\sigma}_0^2$. An estimate of σ_δ^2 for (3.1.3) and (3.1.4) is given by the mean square due to deviations from regression.

It is not always easy to interpret the multiple regression meaningfully. It may be advisable to reduce the dimensions as much as possible by considering each variate separately and to reject those which do not account for sufficient variability in the interaction sum of squares.

3.2 Relationship with other work

Yates and Cochran's technique is a particular case of the above method with $p = 1$ (considered in (3.1.5) with $z_i = \bar{y}_{i.} - \bar{y}_{..}$). As there is inherent dependence of y_{ij} on $\bar{y}_{i.}$ Yates and Cochran's approach is not statistically valid as pointed out by Freeman and Perkins (1971). In the following we shall consider the consequences of this in more detail.

3.2.1 Some remarks on Yates and Cochran's approach

Let us consider the interpretation of the coefficients β_j defined in Chapter 2. Neither their absolute values nor their signs are independent of the varieties taken in the trial. To demonstrate this we shall consider the regression model discussed in this chapter. Consider the model (3.1.1) with an external variate z_i

$$y_{ij} = \mu + e_i + v_j + b_j z_i + \delta_{ij} \quad . \tag{3.2.1}$$

For the present purposes we shall not impose any conditions on b_j and z_i. The b_j are estimated by

$$\hat{b}_j = \sum_i \frac{(y_{ij} - \bar{y}_{i.})(z_i - \bar{z})}{\sum_i (z_i - \bar{z})^2} \quad . \tag{3.2.2}$$

The expected value of $\hat{b}_j$ is given by

$$E(\hat{b}_j) = b_j - \bar{b} \quad ;$$

thus $\hat{b}_j$ estimates the deviation of b_j from its mean over all the varieties taken in the trial. Thus only differences of two varieties i.e. $b_j - b_{j'}$ can be estimated independently of other varieties in the trial.

The estimates obtained by Yates and Cochran (1938) are obtained by replacing $y_{ij} - \bar{y}_{i.}$ in (3.2.2) by y_{ij} and substituting z_i equal to $\bar{y}_{i.} - \bar{y}_{..}$. This gives

$$1 + \hat{b}_j = \frac{\sum_i y_{ij}(\bar{y}_{i.} - \bar{y}_{..})}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2} \quad . \qquad (3.2.3)$$

Substituting (3.2.1) in (3.2.2) and taking expectations of the numerator and denominator independently one obtains in limiting case (when $n \to \infty$)

$$1 + \hat{b}_j \underset{n\to\infty}{\to} \frac{\sigma_e^2 + (b_j + \bar{b})\sigma_{ez} + b_j\bar{b}\sigma_z^2 + \sigma_\delta^2/m}{\sigma_e^2 + 2\bar{b}\sigma_{ez} + \bar{b}^2\sigma_z^2 + \sigma_\delta^2/m} \qquad (3.2.4)$$

where σ_{ez} is the covariance between e_i and z_i . In a particular case with z_i as e_i in the model (3.2.1) $(1 + \hat{b}_j)$ approaches

$$1 + \hat{b}_j \to \frac{1 + b_j + \dfrac{\sigma_\delta^2}{m(1 + \bar{b})\sigma_e^2}}{1 + \bar{b} + \dfrac{\sigma_\delta^2}{m(1 + \bar{b})\sigma_e^2}} \quad . \qquad (3.2.5)$$

This shows that $1 + \hat{b}_j$ is inconsistent estimator of $(1 + b_j)/(1 + \bar{b})$. When $\sigma_\delta^2/\sigma_e^2$ is small, which is very likely when the differences between sites are large, the bias will be negligible. However, this can be corrected by using the estimate of $\sigma_\delta^2/\sigma_e^2$. Again it is not possible to obtain estimate of $1 + b_j$ as they are standardised by division by $(1 + \bar{b})$. Thus the magnitude and sign are not independent of the other varieties considered in the trial. It may be possible to overcome this problem by considering a large number of varieties and assuming them to be giving a fair representation of the population. However, the ratios

$(1 + b_j)/(1 + b_{j'})$ can be obtained independently of other varieties in the trial when $\sigma_\delta^2/\sigma_e^2$ is negligible.

Putting z_i equal to e_i in (3.2.1) is one of the simplest cases. However, it is not necessary that z_i and e_i should be the same or even highly correlated. e_i can be taken as an additive effect of environment and z_i can be considered as an environmental variate which gives rise to multiplicative effect $(b_j z_i)$ with different regression coefficient (b_j) corresponding to each variety. The interpretation of estimators $1 + b_j$ in (3.2.3) very much depends on the relationship between e_i and z_i. In cases when the z_i cannot be measured independently the interpretation of the parameters $1 + b_j$ is not easy. The effect of the relationship between e_i and z_i can be seen from the expression in (3.2.4). When e_i and z_i are highly correlated (positively or negatively) this results in the case considered in (3.2.5).

When e_i and z_i are independent (3.2.4) yields

$$1 + \hat{b}_j \rightarrow \frac{1 + \rho\bar{b}b_j}{1 + \rho\bar{b}^2} , \qquad (3.2.6)$$

(after putting $\sigma_{ez} = 0$), where

$$\rho = \sigma_z^2/\sigma_e^2 .$$

In the above expressions, for simplicity, $\sigma_\delta^2/\sigma_e^2$ is assumed to be negligible. When ρ is small and negligible then $1 + \hat{b}_j$ will all be nearly equal to unity. However, when ρ is large then we should be able to estimate the relative magnitudes of $1 + b_j$ from (3.2.3). But it is very likely that it will be wrongly interpreted. Suppose z_i represents rainfall and is independent of additive effect e_i. Then the heterogeneous regression among $1 + b_j$ obtained through (3.2.3) will lead us to believe that heterogeneity is due to e_i i.e. productivity factors.

In any case correct interpretation of $1 + b_j$ obtained through (3.2.3) is neither easy nor unique. For example, one can obtain one set of b_j from (3.2.3), where different environments are obtained by applying different level of one fertilizer. Another experimenter can produce similar changes among environments by some other factor i.e. another fertilizer, irrigation level etc., and may arrive at different values of b_j. It shows that two persons working with the same set of varieties, but different environment factors may arrive at quite different values of b_j. This emphasises the need for defining the population of environments and the underlying environment factors for proper use and interpretation of these parameters. Similar criticism is applicable to estimates of $(1 + \beta_j)$ obtained in Chapter 2.

The problem of testing the invariance of β_j is the same as that of testing the invariance of loadings for different populations in Factor Analysis. This is not an easy problem to tackle but some methods are available in the literature (Lawley and Maxwell, 1971). In the following we shall discuss an easy and approximate method for testing the equality of coefficients b_j obtained from different populations. If they are found consistent only then they should be used for any predictive purposes over the factors considered.

3.2.2 Test of hypothesis for equality of $(1 + b_j)$

Let us consider an extended version of the model (1.3.1)

$$y_{ijk} = \mu + e_{ik} + v_{jk} + \eta_{ijk} + \varepsilon_{ijk} \; ; \qquad (3.2.7)$$

$$i = 1,2,\ldots,n_k \; ; \; j = 1,2,\ldots,m \; ; \; k = 1,2,\ldots,t \; ;$$

where another subscript (k) has been added to denote another factor e.g. years etc. y_{ijk}, η_{ijk} and ε_{ijk} are the mean yield, interaction and experimental error component for the variety j at the site i in the year k, respectively. e_{ik} is the effect of the site i in the year k and v_{jk} is the effect of the variety j in the year k. It is not necessary that the same sites should be represented every year. Consider the possible relationship between η_{ijk} and e_{ik} of the type given in (1.3.2) as

$$\eta_{ijk} = b_{jk} \, e_{ik} + \varepsilon'_{ijk} \, . \qquad (3.2.8)$$

For simplicity, we have used b_j of Yates and Cochran's approach considered in (3.2.3) rather than β_j of M.L approach considered in Chapter 2. Now in this section we want to test whether b_{jk} remains the same for all k i.e. the above relationship can be written as

$$\eta_{ijk} = b_j \, e_{ik} + \varepsilon'_{ijk} \quad . \qquad (3.2.9)$$

Under the null hypothesis

$$b_{jk} = b_j \;;\; (j = 1,2,\ldots,m \;;\; k = 1,2,\ldots,t) .$$

Geometrically speaking, we have to test, simultaneously, whether each of the m lines is parellel to the respective lines (corresponding to each variety) for the variation in k . For this, in the following we shall substitute (3.2.8) and (3.2.9) in (3.2.7) and obtain estimates and residual sum of squares which will give the test statistic immediately.

Substituting (3.2.8) in (3.2.7) we obtain

$$y_{ijk} = \mu + v_{jk} + (1 + b_{jk})\, e_{ik} + \delta_{ijk} \qquad (3.2.10)$$

where

$$\delta_{ijk} = \epsilon'_{ijk} + \xi_{ijk} .$$

We shall assume that the variance of δ_{ijk} remains constant for all k . Putting $e_{ik} = \bar{y}_{i.k} - \bar{y}_{..k}$ and taking it as a fixed variable, the least squares estimates of $1 + b_{jk}$ are obtained as

$$(1 + \hat{b}_{jk}) = \frac{\sum_i y_{ijk}\,(\bar{y}_{i.k} - \bar{y}_{..k})}{\sum_i (\bar{y}_{i.k} - \bar{y}_{..k})^2} \qquad (3.2.11)$$

and the residual sum of squares over all k (corresponding to the sum of squares of deviation from regression in Table 2.5.3) is given by

Deviation from regression =

$$\sum_{k,j,i} (y_{ijk} - \bar{y}_{.jk})^2 - \sum_{k,j} \frac{\left[\sum_i y_{ijk}(\bar{y}_{i.k} - \bar{y}_{..k})\right]^2}{\sum_i (\bar{y}_{i.k} - \bar{y}_{..k})^2} . \qquad (3.2.12)$$

Now under the null hypothesis ($b_{jk} = b_j$ for all k) the model in (3.2.7) can be written as

$$y_{ijk} = \mu + v_{jk} + (1 + b_j)e_{ik} + \delta_{ijk} . \qquad (3.2.13)$$

The combined estimates of b_j is given by

$$(1 + \hat{b}_j) = \frac{\sum_{k,i} y_{ijk}(\bar{y}_{i.k} - \bar{y}_{..k})}{\sum_{k,i} (\bar{y}_{i.k} - \bar{y}_{..k})^2} \qquad (3.2.14)$$

and the residual sum of squares is given by

Deviation from combined regression =

$$\sum_{k,j,i} (y_{ijk} - \bar{y}_{.jk})^2 - \sum_{j} \frac{\left[\sum_{k,i} y_{ijk}(\bar{y}_{i.k} - y_{..k})\right]^2}{\sum_{k,i} (\bar{y}_{i.k} - \bar{y}_{..k})^2} . \qquad (3.2.15)$$

Sum of squares due to deviation from parallelism obtained by subtracting (3.2.12) from (3.2.15)

$$\sum_{k,j} \frac{\left[\sum_{i} y_{ijk}(\bar{y}_{i.k} - \bar{y}_{..k})\right]^2}{\sum_{i} (\bar{y}_{i.k} - \bar{y}_{..k})^2} - \sum_{j} \frac{\left[\sum_{k,i} y_{ijk}(\bar{y}_{i.k} - \bar{y}_{..k})\right]^2}{\sum_{k,i} (\bar{y}_{i.k} - \bar{y}_{..k})^2} . \qquad (3.2.16)$$

The sum of squares due to heterogeneity among combined regressions is given by

$$\sum_{j} \frac{\left[\sum_{k,i} y_{ijk}(\bar{y}_{i.k} - \bar{y}_{..k})\right]^2}{\sum_{k,i} (\bar{y}_{i.k} - \bar{y}_{..k})^2} - m \sum_{k,i} (\bar{y}_{i.k} - \bar{y}_{..k})^2 . \qquad (3.2.17)$$

This must be noted here that

$$\sum_{j} b_{jk} = 0 \text{ for all } k \; ; \quad \sum_{j} b_j = 0 \quad .$$

The sum of squares due to deviation from parallelism on $(k-1)(m-1)$ D.F. can be tested against the sum of squares due to deviation from regression on $\sum_{k} (m-1)(n_k - 2)$ D.F. as obtained in (3.2.12). A worked out example has been given in Section 3.3.

3.2.3 Some further discussions

Hardwick and Wood (1972) used the model in (3.0.1) after orthogonalisation of variates z_{ik} so that the resultant components are independent. They suggested the estimation of parameters by using these independent components as environmental variates. This is of great advantage when large numbers of external variables are available but the interpretation of the orthogonal combination of variables is not easy.

Sometimes information on disease for each plot within each replication is available and this can be used in analysis of covariance, possibly, with different slopes for each variety. However, in large scale trials it is not uncommon to take such disease records on only one replication. Within site analysis for such covariates is not possible. Assuming that there is not much variation in the covariates from replication to replication (within sites) these data can be used, say, x_{ij} corresponding to each y_{ij} . Now the analysis of covariance can be used and possible differences between slopes corresponding to different varieties can be tested. In some cases the differences in slopes may account for a significant portion of the interaction sum of squares and this may be interpreted accordingly. This type of analysis does not give rise to any new statistical problem.

3.3 An example

The data given in Table 2.5.1 have been used to examine the effect of altitude, rainfall and sunshine on the relative performance of varieties, using the method given in Section 3.1. Analysis has been done for each variate individually; Table 3.3.1 gives the AOV table and Table 3.3.2 the regression coefficients and their standard errors.

TABLE 3.3.1

ANALYSIS OF INTERACTION FOR EXTERNAL VARIABLES

	Altitude		Rainfall		Sunshine	
Source	D.F.	M.S.	D.F.	M.S.	D.F.	M.S.
Heterogeneity among Regression	7	0.116	7	0.655*	7	0.081
Deviation from Regression	126	0.140*	91	0.100*	91	0.145*
Total	133	0.140	98	0.140	98	0.140
$\hat{\sigma}_0^2$	0.025		0.025		0.025	

TABLE 3.3.2

REGRESSION COEFFICIENTS FOR EXTERNAL VARIABLES

Variety Names	$\hat{b}_j$ (Alt.)	$\hat{b}_j$ (Rain)	$\hat{b}_j$ (Sun)
Ymer	-0.370	-0.225	-0.236
Gerkra	-0.042	-0.004	-0.005
Goldfield	0.506	0.739*	0.515
Imber	-0.054	0.959*	-0.054
Maris Mink	-0.531	-1.229*	-0.404
Mazurka	0.108	0.018	-0.074
Pegasus	0.343	0.744*	0.274
Universe	0.040	-1.002*	-0.016
S.E.	0.350	0.296	0.350
S.E. of difference	0.529	0.447	0.529

There is no evidence that the altitude and sunshine contribute to variety-environment interaction and we shall not consider them any further. Rainfall accounts for about 34% of the interaction S.S. The most affected varieties are Maris Mink and Universe which show markedly decreased yields when rainfall is high. This effect does not appear to be due to lodging. Both varieties are short strawed and late maturing. The negative regression is rather puzzling. Rainfall is negatively correlated ($r = -0.702$) with the average yield of all the varieties and this may explain the relative sizes of $(1 + \hat{\beta}_j)$. Thus the conclusion can be summed up as follows. Varieties Maris Mink and Universe do better than average on dry sites and worse than average on wet sites. In general mean yields are largest in experiments with low rainfall. However, more data may be needed to confirm this.

A possible practical application of these results is as follows. We consider two varieties Maris Mink and Goldfield with similar mean yields but different regression coefficient with rainfall. The following equations* give the estimated yield at the site i :

Maris Mink : $\hat{y}_{i5} = 5.57 + e_i - 1.229\, z_i$

Goldfield : $\hat{y}_{i3} = 5.57 + e_i + 0.739\, z_i$

$$\hat{y}_{i5} - \hat{y}_{i3} = -1.968\, z_i \;;$$

$$V(\hat{y}_{i5} - \hat{y}_{i3}/i) = 0.20\,(0.05 + z_i^2)\,.$$

---

* In these calculations we have assumed that the estimated regression coefficients are applicable to the whole of Scotland. Some bias may, however, result from the use of data from only fifteen sites out of twenty (See Table 2.5.1).

Considering the differences on particular sites, when $z_i > 0$ then Goldfield is preferable and when $z_i < 0$ then Maris Mink is preferable. If one has to choose between these two varieties then for the East and North, where on the average rainfall is smaller than in West, one will choose Maris Mink; but for the West one will choose Goldfield.

In the following we shall consider an example for testing the consistency of $(1 + b_j)$ calculated for three years from the data of the trial on winter wheat conducted in Scotland. The mean of three years are given in Table 3.3.3. The means for 1969 and 1970 are based on the data from 15 sites whereas means of 1971 are based on the data from 16 sites.

TABLE 3.3.3

WHEAT VARIETY MEAN YIELDS (GRAINS t/ha)

Years Varieties	1969	1970	1971
Cappelle	5.33	4.95	5.01
Bouquet	5.39	5.06	5.43
Cama	5.62	5.33	5.12
Maris Beacon	5.88	5.49	5.87
Maris Settler	5.65	5.22	4.82
Tommy	5.43	5.12	5.23

The individual year's analyses have been given in Table 3.3.4.

TABLE 3.3.4

AOV OF WINTER WHEAT FOR THREE YEARS (t/ha)

	1969		1970		1971	
Source	D.F.	M.S.	D.F.	M.S.	D.F.	M.S.
Varieties	5	0.644	5	0.574	5	2.182
Sites	14	8.319	14	15.084	15	6.620
Het. in regression	5	0.098	5	0.132	5	0.157
Dev. from reg.	65	0.068	65	0.079	70	0.252
Total	89	1.400	89	2.478	95	1.354

The separate and combined estimates of regression coefficients, as calculated by the method discussed in Section 3.2.2, are given in Table 3.3.5.

TABLE 3.3.5

ESTIMATES OF REGRESSION OF VARIETY YIELDS ON SITE MEANS

Years j	1	2	3	4	5	6
1969 $(1 + \hat{b}_{j1})$	0.991	0.871	1.073	1.055	1.007	1.003
1970 $(1 + \hat{b}_{j2})$	0.965	0.931	1.107	1.013	1.014	0.970
1971 $(1 + \hat{b}_{j3})$	0.971	1.078	0.915	1.052	0.864	1.119
Combined $(1 + \hat{b}_j)$	0.973	0.949	1.053	1.033	0.977	1.014

Table 3.3.6 gives the combined analysis over three years. The sum of squares have been calculated by using the expressions of Section 3.2.2 .

TABLE 3.3.6

COMBINED AOV OF WINTER WHEAT OVER THREE YEARS (t/ha)

Source	D.F.	S.S.	M.S.
Years	2	6.637	
Varieties	5	11.639	2.328
Years x Varieties	10	5.362	0.536
Sites (W. years)	43	426.955	
Het. Combined regression	5	0.564	0.113
Deviation from parallelism	10	1.368	0.137
Deviation from regression	200	27.834	0.139
Total	275	480.359	

In the above table sum of squares due to deviation from parallelism is not significant showing that the regression of variety yields on the site means remain consistent over three years. Sum of squares due to heterogeneity among regression in Tables 3.3.4 and 3.3.6 show that the regression coefficients do not differ significantly from zero. Therefore there is nothing much to choose among these varieties on the basis of their regression coefficients.

CHAPTER 4

ESTIMATION AND TEST OF HYPOTHESIS OF STABILITY VARIANCES

4.0 Introduction and development of previous work

The attempt to explain variety-environmental interaction in terms of β_j and b_j, as discussed in previous chapters, may not be very successful when only a small fraction of the interaction sum of squares can be attributed to heterogeneity among the regressions. In any case it is of considerable interest and importance to establish a measure of variability of performance of individual varieties over a population of environments. We now consider the estimation of σ_j^2 as defined in (1.3.5). We have

$$\sigma_j^2 = \underset{i}{E}(\epsilon_{ij}^2) = \underset{i}{E}(\eta_{ij}^2 + \xi_{ij}^2) = \sigma_j'^2 + \sigma_o^2 .$$

Thus, σ_j^2 can be considered as the sum of two components, viz. a within environmental variance σ_o^2 and the interaction variance $\sigma_j'^2$. We will call σ_j^2 the 'stability variance' of variety j. A variety is said to be 'stable' if its stability variance σ_j^2 is equal to the within environmental variance σ_o^2, i.e. if $\sigma_j'^2 = 0$. Large values of σ_j^2 indicate instability.

Estimation of σ_j^2 is analogous to estimating heterogeneous error variances in a two-way classification when variances change in one way, considered by Ehrenberg (1950) and later by Russell and Bradley (1958). Rao (1970, 1972) has generalised the above procedure for any classification and also gave some optimum properties of the estimators. Such estimates are available only when $m \geq 3$. Baker (1969) has calculated a very similar parameter and his approach will be discussed in more detail in Section 4.3.

The problem of estimation of stability variances σ_j^2 when $m \geq 3$ will be discussed in Section 4.1. The method of estimation of σ_j^2 discussed in Section 4.1 does not work when $m = 2$ and a method for this particular case will be discussed very briefly in Section 4.2. Section 4.1 also deals with the estimation of stability variances for an extended model with external variables.

The problem of testing hypotheses will be discussed in Section 4.2. Tests for equality of stability variances differ considerably for $m = 2$ and $m \geq 3$ and will be dealt with separately. We also consider a test of the stability of a single variety $(\sigma_j'^2 = 0)$. Section 4.3 discusses the relative merits of estimated stability variances and the regressions coefficients considered in previous chapters. This section also deals with the relevance of earlier work. In Section 4.4 a worked example will be considered.

4.1 Estimation of stability variances

In this section we shall consider two methods of estimation. The method in Section 4.1.1 gives an unbiased estimate of σ_j^2 and is easy to calculate. The method in Section 4.1.3 gives maximum likelihood (ML) estimates.

4.1.1 Unbiased estimation of σ_j^2 in two-way classification model

Here we shall give some results obtained by Enhrenberg (1950). We shall not discuss them in any great detail. The unbiased estimates of σ_j^2 will be denoted by $\hat{\sigma}_j^2$; they are given by

$$\hat{\sigma}_j^2 = \left[m(m-1) \sum_i r_{ij}^2 - \sum_{i,j} r_{ij}^2\right] / (n-1)(m-1)(m-2), \qquad (4.1.1)$$

where r_{ij} is the residual corresponding to (i,j)th cell,

$$r_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..} \quad .$$

Alternatively $\hat{\sigma}_j^2$ can be written as

$$\hat{\sigma}_j^2 = \left[m(m-1) \sum_i (u_{ij} - \bar{u}_{.j})^2 - \sum_{i,j} (u_{ij} - \bar{u}_{.j})^2\right] / (n-1)(m-1)(m-2), \qquad (4.1.2)$$

where

$$u_{ij} = y_{ij} - \bar{y}_{i.} \;\; ; \;\; \bar{u}_{.j} = \sum_i u_{ij}/n \quad .$$

$\hat{\sigma}_j^2$ are obtained as linear combination of squares of residuals r_{ij} ; therefore, they are independent of μ, v_j and σ_e^2 . It is not difficult to verify that they are unbiased estimators of σ_j^2 . Rao (1970) has proved that they have minimum average (over all j) variance among all possible quadratic unbiased estimators (MINQUE) of σ_j^2 .

It is not difficult to see from (4.1.1) that the mean of the $\hat{\sigma}_j^2$ is the same as the mean square due to interaction. Therefore, by multiplying each $\hat{\sigma}_j^2$ by $(m-1)(n-1)/m$ we shall obtain m components of interaction, one corresponding to each genotype. These estimates are not independent. As they are differences of two sums of squares, they can be negative, but negative estimates of variances are not uncommon in variance component problems.

The variance of $\hat{\sigma}_j^2$ is not only a function of σ_j^2 but also of the variances of other varieties in the trial.

$$V(\hat{\sigma}_j^2) = \frac{2}{(n-1)} \sigma_j^4 + \frac{4}{(n-1)(m-1)^2} \sigma_j^2 \sum_{j' \neq j} \sigma_{j'}^2 + \frac{4}{(m-2)^2} \sum_{j', j'' \neq j} \sum \sigma_{j'}^2 \sigma_{j''}^2 .$$

When m is large the variances of $\hat{\sigma}_j^2$ can be approximated by

$$V(\hat{\sigma}_j^2) \approx \frac{2\sigma_j^4}{(n-1)} . \qquad (4.1.3)$$

Under the assumption that all σ_j^2 are equal to σ^2, say, the estimate of σ^2 is given by the interaction mean square $\hat{\sigma}^2$, say.

As far as the relative variabilities of varieties from the average variability are concerned it may be advisable to calculate a dimensionless index given by

$$\hat{\sigma}_j^2 / \hat{\sigma}^2 .$$

When all σ_j^2 are equal the mean and variance of this index (approximately) are given by

$$E(\hat{\sigma}_j^2/\hat{\sigma}^2) = 1 ; \quad V(\hat{\sigma}_j^2/\hat{\sigma}^2) = \frac{2(m-1)^2}{(m-2)(\overline{n-1}\;\overline{m-1} + 2)} .$$

If m is large then

$$V(\hat{\sigma}_j^2/\hat{\sigma}^2) \simeq \frac{2(m-1)}{(n-1)(m-2)} \simeq \frac{2}{n-1} .$$

These expressions can be used to test the departure of individual index from unity. These indexes are correlated as they must add to m.

4.1.2 Further extension of model

To make further progress in the interpretation of instability, we shall consider the model (3.1.1) incorporating variation due to external variables. At present we shall consider the case of only one external variable ($p = 1$), but the method can be easily extended when information on more than one external variable is available. We shall omit one subscript of b's and z's of the model (3.1.1) and denote them by b_j ($j = 1, 2, \ldots, m$) and z_i ($i = 1, 2, \ldots, n$). As before, without any loss of generality it can be assumed that,

$$\sum_j b_j = \sum_i z_i = 0 ; \quad \sum_i z_i^2 = 1 .$$

The model can now be written as

$$y_{ij} = \mu + e_i + v_j + b_j z_i + \delta_{ij} ; \qquad (4.1.4)$$

we assume that

$$V(\delta_{ij}) = \theta_j^2 ; \quad (j = 1, 2, \ldots, m).$$

and proceed to estimate θ_j^2. The usual estimator of b_j, by the method of unweighted least squares as given in (3.1.5), can be obtained as

$$\hat{b}_j = \sum_i (u_{ij} - \bar{u}_{.j}) z_i \tag{4.1.5}$$

where

$$u_{ij} = y_{ij} - \bar{y}_{i.} .$$

Using the method of Section 4.1.1, unbiased estimates of θ_j^2 for the model in (4.1.4) can be obtained as $\hat{\theta}_j^2$ given by

$$\hat{\theta}_j^2 = \left[m(m-1) \sum_i r_{ij}'^2 - \sum_{i,j} r_{ij}'^2 \right] / (n-2)(m-1)(m-2) \tag{4.1.6}$$

where

$$r'_{ij} = u_{ij} - \bar{u}_{.j} - \hat{b}_j z_i .$$

The estimators obtained in (4.1.6) are quadratic (in y's) estimators of θ_j^2 and have the properties of MINQUE estimators. When m is large, the variance of $\hat{\theta}_j^2$ can be approximated by

$$V(\hat{\theta}_j^2) = 2\theta_j^4/(n-2) . \tag{4.1.7}$$

4.1.3 Maximum likelihood estimate of stability variance

For estimation of σ_j^2 the method described in Section 4.1.1 will be adequate. Later in this chapter we will also discuss the likelihood ratio test for the equality of all σ_j^2. For that purpose we shall need maximum likelihood estimates of σ_j^2's. Russell and Bradley (1958) gave a method, based on contrasts of y_{ij} for $m = 3$. They obtained explicit estimators of σ_j^2 and the estimators are the same as considered in Section 4.1.1. In this section we shall work with contrasts and generalise the method of Russell and Bradley (1958) for $m \geq 3$. No explicit solution of σ_j^2 is available when $m > 3$ but solutions can be obtained by iteration. It must be noted here that this method does not give maximum likelihood (ML) estimates based on all the data. ML estimates of all the parameters do not exist when environment and variety effects both are fixed. The estimates considered here are based on the contrasts and therefore independent of e_i's and v_j's or of their population parameters.

A suitable set of orthogonal row-column contrasts x_{lk} ($l = 1, 2, \ldots, n-1$; $k = 1, 2, \ldots, m-1$) is given by the elements of an $(n-1)(m-1)$ matrix X, where

$$X' = L_m Y' L_n' = (X_1, X_2, \ldots, X_{n-1}) ;$$

where

$$Y' = (Y_1, Y_2, \ldots, Y_n) .$$

Here the matrix L_q is a matrix of order $(q-1) \times q$ such that

$$L_q 1 = 0 \; ; \quad L_q L_q' = I \, .$$

It can be shown that

$$E(X) = 0 \; ; \; V(X_l) = H \; ; \; \mathrm{Cov}(X_l, X_{l'}) = 0 \text{ for } (l \neq l') \, ,$$

where

$$H = L_m \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{bmatrix} L'_m = L_m \Sigma L'_m \, .$$

and Σ is a diagonal matrix with $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$ as its elements.

The joint distribution of x_{lk} is given by

$$f(x) = (2\pi)^{-(n-1)(m-1)/2} |H^{-1}|^{(n-1)/2} \exp\left[-\tfrac{1}{2} \sum_{l=1}^{n-1} X_l' H^{-1} X_l\right] . \qquad (4.1.8)$$

Let $\hat{\sigma}_j^2$ be the estimate of σ_j^2 obtained by maximising $\log f(x)$. Also let $\log L(\Omega)$ and $\hat{H}$ be the expression of $\log f(x)$ and H obtained by substituting $\hat{\sigma}_j^2$ for σ_j^2.

Now if L_m used is such that

$$L_m = (l_{ij})$$

where

$$l_{ij} = 1/\sqrt{i(i+1)} \qquad j \leq i$$

$$= -i/\sqrt{i(i+1)} \qquad j = i+1$$

$$= 0 \qquad j > i+1$$

then

$$H = PDD_2D'P$$

where

$$P = \text{Diag}(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{6}}, \ldots, \frac{1}{\sqrt{m(m-1)}})$$

$$D_2 = \text{Diag}(\sigma_2^2, \sigma_3^2, \ldots, \sigma_m^2) + \sigma_1^2 \, 11'$$

and D is a lower triangular matrix of order $(m-1)$ such that

$$d_{ii} = i$$

$$d_{ij} = -1 \qquad j < i$$

$$= 0 \qquad j > i .$$

H^{-1} can be easily calculated by

$$H^{-1} = P^{-1}D'^{-1}D_2^{-1}D^{-1}P^{-1}$$

as all the inverses are easily obtainable. The elements of D^{-1} are such that

$$D^{-1} = (d^{ij})$$

$$d^{ii} = 1/i$$

$$d^{ij} = 1/j(i+1) \qquad j < i$$

$$d^{ij} = 0 \qquad j > i \ .$$

Now

$$-\frac{2}{n-1} \log L(\Omega) = \log|H| + \operatorname{tr}(H^{-1}T) + \text{constant} \qquad (4.1.9)$$

where T is given by

$$T = \frac{X'X}{(n-1)} \ .$$

Starting with some good initial estimates of σ_j^2, (4.1.9) should be minimised iteratively. For this purpose the initial estimates can be taken as $\hat{\sigma}_j^2$ obtained in Section 4.1.1.

It must be noted here that the ML estimator of the common variance $\sigma^2(\sigma_j^2 = \sigma2$ for all j) is given by $\hat{\sigma}^2$, the mean square of variety environment interaction.

4.1.4 <u>Relative merits of the two estimates</u>

When $m = 3$ the estimates $\hat{\sigma}_j^2$ and $\hat{\hat{\sigma}}_j^2$ are identical. However, it is difficult to compare the efficiencies in general. MINQUE estimators are easily obtainable and the method can be easily extended to the case of non-orthogonal data. When m is large the ML estimates are difficult to obtain even by a high speed computer. Thus for most purposes unbiased estimators $\hat{\sigma}_j^2$ will be adequate at least when n is moderate.

When e_i are random variables another estimate of σ_j^2 and σ_e^2 can be obtained using the likelihood function of S as defined in Section 2.1. However, we have not considered this method at present as we are not particularly interested in estimating σ_e^2. The estimates for σ_j^2 by both likelihood methods are not likely to differ very much, at least when n is large.

4.2 Test of hypothesis

The test considered here for $m = 2$ and $m \geq 3$ differ markedly and are dealt with separately.

Morgan (1939) and Pitman (1939) have considered the test of significance for a difference between two variances from a bi-variate normal population, without making any assumption about the correlation coefficient. Grubbs (1948) considered the problem of estimating the precision of instruments, when only one observation is taken on the subjects for each instrument. Maloney and Rastogi (1970), and Jaech (1971) considered some tests of hypotheses for the comparison of the precision of two instruments. A generalised version of Pitman's (1939) result given by Shukla (1973), is discussed in Section 4.2.1.

For $m \geq 3$, Han (1969) has given an exact method. However, the power of his tests depends upon the variability among e_i. A better method (Shukla, 1972b) will be given in Section 4.2.2. A likelihood ratio test which is a generalisation of Russell and Bradley (1958) will be discussed in Section 4.2.3.

4.2.1 Test of hypothesis for two variances

4.2.1.1 Formulation and notation

Consider the model (1.3.1) with two columns

$$y_{ij} = \tau_i + v_j + \epsilon_{ij} ; \quad (i = 1, 2, \ldots, n : \; j = 1, 2) \qquad (4.2.1)$$

where

$$\tau_i = \mu + e_i ; \quad V(\tau_i) = \sigma_e^2 ; \quad V(\epsilon_{ij}) = \sigma_j^2 ; \quad V(y_{ij}) = \sigma_e^2 + \sigma_j^2 .$$

$$\mathrm{Cov}(y_{ij}, y_{ij'}) = \sigma_e^2 \text{ for } j \neq j' .$$

Denote the sample variances and covariance of y_1 and y_2 by

$$s_{11} = \sum_i (y_{i1} - \bar{y}_{.1})^2/(n-1); \quad s_{22} = \sum_i (y_{i2} - \bar{y}_{.2})^2/(n-1)$$

$$s_{12} = \sum_i (y_{i1} - \bar{y}_{.1})(y_{i2} - \bar{y}_{.2})/(n-1) .$$

It is easy to show that

$$E(s_{11}) = \sigma_e^2 + \sigma_1^2$$

$$E(s_{22}) = \sigma_e^2 + \sigma_2^2$$

$$E(s_{12}) = \sigma_e^2 \ .$$

Grubbs (1948) estimated σ_1^2, σ_2^2 and σ_e^2, respectively, by

$$\hat{\sigma}_1^2 = s_{11} - s_{12} \ ; \ \hat{\sigma}_2^2 = s_{22} - s_{12} \ ; \ \hat{\sigma}_e^2 = s_{12} \ . \qquad (4.2.2)$$

4.2.1.2 <u>Test of hypothesis</u>

The present work is concerned only with the null hypothesis

$$H_0 : \sigma_1^2 = k\sigma_2^2$$

where k is a specified constant $(k \geq 0)$. This will be tested against

$$H_A : \sigma_1^2 \neq k\sigma_2^2 \ .$$

Consider

$$g_i = y_{i1} - y_{i2} \ ; \ f_i = y_{i1} + ky_{i2} \ .$$

The variances and covariance of g and f are given by

$$\begin{aligned} V(g) &= \sigma_1^2 + \sigma_2^2 \\ V(f) &= \sigma_1^2 + k^2\sigma_2^2 + \sigma_e^2\overset{2k}{\cancel{(1+k)}}^2 \qquad (4.2.3) \\ Cov(g,f) &= \sigma_1^2 - k\sigma_2^2 \ . \end{aligned}$$

Denote the population correlation coefficient between g and f by ρ ; then

$$\rho = \frac{\sigma_1^2 - k\sigma_2^2}{[(\sigma_1^2 + \sigma_2^2)(\sigma_1^2 + k^2\sigma_2^2 + \sigma_e^2(1+k)^2)]^{\frac{1}{2}}} \quad . \qquad (4.2.4)$$

A necessary and sufficient condition for ρ to be zero is that $\sigma_1^2 = k\sigma_2^2$; this suggests that the test of H_0 is equivalent to a test of g and f having zero correlation coefficient. Denote the sample correlation coefficient between g and f by r ; this can be expressed in terms of s_{11}, s_{22} and s_{12} as follows. Write the sample variances and covariance of g and f as

$$s_{gg} = s_{11} + s_{22} - 2s_{12}$$

$$s_{ff} = s_{11} + k^2 s_{22} + 2ks_{12}$$

$$s_{gf} = s_{11} - ks_{22} + (k-1)s_{12} \quad .$$

Then, by definition

$$r = \frac{s_{gf}}{(s_{gg}s_{ff})^{\frac{1}{2}}} = \frac{s_{11} - ks_{22} + (k-1)s_{12}}{[(s_{11} + s_{22} - 2s_{12})(s_{11} + k^2 s_{22} + 2ks_{12})]^{\frac{1}{2}}} \,. \qquad (4.2.5)$$

Thus under the null hypothesis H_0 , the statistic t_0 given by

$$t_0 = r\sqrt{(n-2)/(1-r^2)} \qquad (4.2.6)$$

will have a Student's t-distribution with (n-2) D.F. .

4.2.1.3 Likelihood ratio test

It can be easily seen that the test statistic used in (4.2.6) is the same as that derived from the likelihood ratio test. For simplicity we shall work with the transformed variates g and f instead of y_1 and y_2 ; g and f have a bi-variate normal distribution with variances and covariance as given in (4.2.3). Let us denote the maximum value of the likelihood under the alternative hypothesis H_A by $L(\Omega)$, where

$$\log L(\Omega) = -\frac{n}{2} - \frac{n}{2}\log(s_{gg}s_{ff} - s_{gf}^2) - n\log(2\pi)$$

$$= -\frac{n}{2} - \frac{n}{2}\log(s_{11}s_{22} - s_{12}^2) - n\log(1+k) - n\log(2\pi) \ . \qquad (4.2.7)$$

Under H_0, g and f are independently distributed. Denote the maximum value of the likelihood under H_0 by $L(\omega)$, and put $s_{gf}^2 = 0$, Since g and f are independent then (4.2.7) becomes

$$\log L(\omega) = -\frac{n}{2} - \frac{n}{2}\log(s_{gg}s_{ff}) - n\log(2\pi)$$

$$= -\frac{n}{2} - \frac{n}{2}\log[(s_{11} + s_{22} - 2s_{12})(s_{11} + k^2 s_{22} + 2ks_{12})]$$

$$- n\log(2\pi) \ .$$

Thus

$$\log\lambda = \log L(\omega) - \log L(\Omega)$$

$$= \frac{n}{2}\log\left[\frac{s_{11}s_{22} - s_{12}^2}{(s_{11} + s_{22} - 2s_{12})(s_{11} + k^2 s_{22} + 2ks_{12})}\right]$$

$$+ n \log(1 + k) .$$

But from (4.2.5)

$$1 - r^2 = \frac{(1 + k)^2 (s_{11}s_{22} - s_{12}^2)}{(s_{11} + s_{22} - 2s_{12})(s_{11} + k^2 s_{22} + 2ks_{12})} ,$$

thus giving the value of

$$\log \lambda = \frac{n}{2} \log(1 - r^2) .$$

4.2.1.4 Confidence intervals for relative precision

The statistic considered in (4.2.6) can be used to obtain a confidence interval for k . Let us denote the Student's -t value at $\alpha/2$ probability level with (n-2) D.F. by $t_{\alpha/2}$. Then the $(1 - \alpha)\%$ confidence limits for k can be obtained by substituting the value of $t_{\alpha/2}$ in place of t_0 and substituting the value of the ratio $r^2/(1-r^2)$ in terms of s_{11}, s_{22}, s_{12} and k .

The quadratic equation in k reduces to

$$k^2(a^2 - P) - 2k(ab + P) + (b^2 - P) = 0 \qquad (4.2.8)$$

where

$$a = s_{22} - s_{12}$$

$$b = s_{11} - s_{12}$$

$$P = t^2_{\alpha/2}\,(s_{11}s_{22} - s_{12}^2)/(n - 2)\ .$$

The solution of (4.2.8) gives K_L and K_U as lower and upper limits,

$$K_L = \frac{b - \sqrt{P}}{a + \sqrt{P}}\ ; \quad K_U = \frac{b + \sqrt{P}}{a - \sqrt{P}}$$

such that

$$P(K_L \leq k \leq K_U) = 1 - \alpha\ .$$

Looking at the expressions for the limits it is obvious that when $a \leq \sqrt{P}$ or $b \leq -\sqrt{P}$, these limits can be exclusive. If a is nearly equal to $\sqrt{P}$ these limits can be infinitely large. If $b \leq \sqrt{P}$ then the lower limit is negative.

4.2.1.5 Comparison with other works

Maloney and Rastogi (1970) considered a particular case of H_0 with $k = 1$. Another particular case of $k = 0$, testing the hypothesis that $\sigma_1^2 = 0$, was considered by Maloney and Rastogi (1970) and by Jaech (1971), and the result in this section gives an exact distribution of the test statistic they considered. Moreover, the results given here enable us to give an exact confidence limit. The disadvantage of the above test is that the power depends upon σ_e^2 : the larger σ_e^2 , the smaller is the power of the test. This has been discussed in more detail in Maloney and Rastogi (1970). If the data from more than two varieties (or instruments) are available then a test can be obtained whose power is independent of σ_e^2 . This is considered in the next section. When $\sigma_e^2 = 0$ (one item being measured repeatedly) then the obvious test for $\sigma_1^2 = k\sigma_2^2$ is to consider the ratio of s_{gg} and s_{ff} and test by an F-test, where $g_i = y_{i1}$ and $f_i = \sqrt{k}\, y_{i2}$.

4.2.2 Test for equality for more than two variances

4.2.2.1 Notation and method

Consider the model (1.3.1) and as before define a $(q - 1) \times q$ matrix L_q such that

$$L_q 1 = 0 \; ; \; L_q L_q' = I \; ; \qquad (4.2.9)$$

where 1 is the vector of unit elements. Now a vector C_i giving the values of $m - 1$ orthogonal contrasts among the observations in the row i can be obtained as follows;

$$C_i' = Y_i' L'_m = (c_{i1}, c_{i2}, \ldots, c_{i(m-1)}), \; (i = 1, 2, \ldots, n) \qquad (4.2.10)$$

where

$$Y_i' = (y_{i1}, y_{i2}, \ldots, y_{im}) .$$

It is easy to show that

$$E(C_i) = L_m V \; ; \; V(C_i) = L_m \Sigma L'_m = H \; , \qquad (4.2.11)$$

where

$$V' = (v_1, v_2, \ldots, v_m)$$

and Σ is a diagonal matrix of $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$ as defined before.

Under the null hypothesis H_0, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2 = \sigma^2$, say, $V(C_i) = \sigma^2 I$.

We test H_0 by (i) estimating H by T defined below and (ii) applying the test of sphericity described by Mauchly (1940). The matrix T has elements $t_{kk'}$ given by

$$t_{kk'} = \frac{1}{(n-1)} \sum_i (c_{ik} - \bar{c}_{.k})(c_{ik'} - \bar{c}_{.k'}); \quad (k, k' = 1, 2, \ldots, m-1) \qquad (4.2.12)$$

The test statistic λ is defined by

$$\lambda = \frac{|T|^{n/2}}{\left[\frac{tr(T)}{m'}\right]^{nm'/2}} \qquad (4.2.13)$$

where $m' = m-1$.

Under the null hypothesis of equality of variances the value of λ should be equal to 1 apart from sampling error.

It is not difficult to show that the statistic λ is invariant for all possible sets of orthogonal contrasts. For, consider any other set of orthogonal contrasts $M_m Y_i$; where M_m is given by

$$M_m = DL_m$$

and D is an orthogonal matrix. Hence

$$DD' = I \text{ and } M_m M'_m = I .$$

Now let λ_1 be the value of λ when M_m is substituted for L_m in equation (4.2.10). This gives from (4.2.13)

$$\lambda_1^{2/n} = \frac{|DTD'|}{\left[\frac{tr(DTD')}{m'}\right]^{m'}} = \frac{|T|}{\left[\frac{tr(T)}{m'}\right]^{m'}} = \lambda^{2/n} .$$

Hence λ is independent of the choise of orthogonal contrasts. Also λ is obviously not affected by any changes in μ, e_i and v_j .

4.2.2.2 <u>Distribution of λ</u>

Moments of the statistic λ under the null hypothesis are given by Anderson (1958, p. 262). For simplicity of notation we shall put $n' = n-1$, and $W = \lambda^{2/n}$. The q^{th} moment of W is given by

$$E(W^q) = \frac{m^{qm}\Gamma(\frac{mm'}{2})}{\Gamma(\frac{mm'}{2} + mq)} \prod_{i=1}^{m} \frac{\Gamma(\frac{n' + 1 - i}{2} + 1)}{\Gamma(\frac{n' + 1 - i}{2})} .$$

Box (1949) has discussed a general method for approximating the distribution of a LR criterion by the first few terms in an infinite series of χ^2's. Anderson (1958) has applied this method to the sphericity test and the results can be put in the following theorem.

<u>THEOREM</u>: Under the null hypothesis of equality of all variances $-2\nu \log \lambda$ is approximately distributed as χ^2 with $\frac{m(n+1)}{2} - 1$ D.F., where ν is given by

$$\nu = 1 - \frac{2m'^2 + m' + 2}{6m'(n-1)} .$$

This result is adequate for large n. A more accurate result (correct to order n^{-2}) can be obtained by including three terms of the infinite series. We then have

$$\Pr\left[-2\nu \log \lambda \leq \eta\right] = \Pr\left[\chi_f^2 \leq \eta\right] + \omega_2\left[\Pr(\chi_{f+4}^2 \leq \eta) - \Pr(\chi_f^2 \leq \eta)\right] + O(n^{-3})$$

where

$$\omega_2 = \frac{(m'+2)(m'-1)(m'-2)(2m'^3 + 6m'^2 + 3m' + 2)}{288\, m'^2 n'^2 \nu^2} , \text{ and}$$

$$f = \frac{m'(m'-1)}{2} - 1 .$$

4.2.2.3 Test of the equality of two column variances

Suppose that H_0 has been shown to be unacceptable. We may now wish to compare a particular pair of column variances, for example σ_1^2 and σ_2^2. The null hypothesis is as follows:

$$H_o^* : \sigma_1^2 = \sigma_2^2 \ .$$

To test the hypothesis we first define the contrasts

$$c_{i1} = (y_{i1} - y_{i2})/\sqrt{2} \ ; \ c_{i2} = (y_{i1} + y_{i2} - 2\bar{y}_{i(m-2)})/\sqrt{6} \ ;$$

where

$$\bar{y}_{i(m-2)} = \sum_{j=3}^{m} y_{ij}/(m-2) \ .$$

The expected values of c_{i1} and c_{i2} are given by

$$E(c_{i1}) = (v_1 - v_2)/\sqrt{2} \ ; \ E(c_{i2}) = (v_1 + v_2 - 2\bar{v}_{m-2})/\sqrt{6} \ ,$$

and the variances and covariance are given by

$$V(c_{i1}) = (\sigma_1^2 + \sigma_2^2)/2 \ ; \ V(c_{i2}) = (\sigma_1^2 + \sigma_2^2 + \frac{4\bar{\sigma}^2_{m-2}}{m-2})/6$$

$$\mathrm{Cov}\ (c_{i1}, c_{i2}) = (\sigma_1^2 - \sigma_2^2)/\sqrt{12}$$

where

$$\bar{\sigma}^2_{m-2} = \sum_{j=3}^{m} \sigma_j^2/(m-2) \ ; \ \bar{v}_{m-2} = \sum_{j=3}^{m} v_j/(m-2) \ .$$

Under the null hypothesis H_o^* the correlation between c_{i1} and c_{i2} is zero. Hence we can test H_o^* by comparing

$$\frac{r^*}{\sqrt{(1-r^{*2})}}\ (n-2)^{\frac{1}{2}} \qquad (4.2.14)$$

with the Student's t-distribution on (n-2) D.F. Here r^* is the sample correlation coefficient between c_{i1} and c_{i2} .

Note, however, that $V(c_{i2})$ and hence the power of the test depends on $\bar{\sigma}^2_{m-2}$, i.e. on the variances in columns other than the two under consideration. Some improvement in the test may be possible if we replace $\bar{y}_{i(m-2)}$ by another mean with a smaller variance. For example, if we know a priori that $\sigma^2_{j'}$ is large we might replace $\bar{y}_{i(m-2)}$ by the mean for columns 3, 4, ..., m; excluding column j' . The power of the test is, however, independent of μ, e_i and v_j . In cases when e_i's are very variable the above test may provide a more powerful test than the one discussed in Section 4.2.1.

4.2.3 A likelihood ratio test

In Section 4.1.3 we have discussed the method of ML for the estimation of σ^2_j's . This method immediately yields the likelihood ratio test.

Let $\log L(\Omega)$ be the value of log-likelihood obtained from (4.1.8) after substituting the ML estimate of σ^2_j i.e. $\hat{\sigma}^2_j$. Under the null hypothesis H_0 the ML estimator of the common variance σ^2 is simply $\hat{\sigma}^2$ i.e. the mean square due to interaction. Let $\log L(\omega)$ the value of log likelihood under H_0 obtained by substituting $\hat{\sigma}^2$ for all σ^2_j in

(4.1.8) and taking the logs. Let λ' be the likelihood ratio. Then

$$\log \lambda' = \log L(\omega) - \log L(\Omega) .$$

This reduces to

$$\log \lambda' = \frac{(n-1)(m-1)}{2} \left[- \log \hat{\sigma}^2 + \frac{1}{(m-1)} \log|\hat{H}| + \frac{1}{(m-1)} \operatorname{tr}(\hat{H}^{-1}T) -1 \right] . \quad (4.2.15)$$

The quantity $-2 \log \lambda'$ is asymptotically distributed as χ^2 with $(m-1)$ D.F. under H_0 .

It is interesting to observe that in the particular case $m = 3$ the statistics λ of (4.2.13) and λ' of (4.2.15) are the same for large n i.e. the sphericity test and the LR test are asymptotically equivalent when $m = 3$.

This result can be shown by substituting

$$L_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix}$$

in equation (4.2.11).

Russell and Bradley (1958) showed that

$$-2 \log \lambda' = -(n-1) \, [\log (\hat{\sigma}_1^2 \, \hat{\sigma}_2^2 + \hat{\sigma}_1^2 \, \hat{\sigma}_3^2 + \hat{\sigma}_2^2 \, \hat{\sigma}_3^2) - 2 \log \hat{\sigma}^2 - \log 3], \quad (4.2.16)$$

where the $\hat{\sigma}_j^2$ are obtained by substituting $m = 3$ in (4.1.1).

$\hat{\sigma}^2_j$ can also be expressed in terms of the elements of T given in (4.2.12)

$$\hat{\sigma}^2_1 = t_{11} + \sqrt{3}t_{12} ,$$

$$\hat{\sigma}^2_2 = t_{11} - \sqrt{3}t_{12} ,$$

$$\hat{\sigma}^2_3 = (3t_{22} - t_{11})/2 .$$

Substituting in equation (4.2.16) we obtain

$$- 2 \log \lambda' = - (n-1) [\log (t_{11}t_{12} - t^2_{12}) - 2 \log \hat{\sigma}^2] . \qquad (4.2.17)$$

This expression is the same as $- 2 \log \lambda$ obtained from equation (4.2.13) apart from the multiplicative factor of n in place of n-1 . The discrepancy is unimportant when n is large.

4.2.4 Comparison of the two tests

It is not clear how the tests proposed in Sections 4.2.2 and 4.2.3 compare in general. Looking at the numbers of D.F. it appears that the LR test may always give a more powerful test for $m > 3$, but in general the value of $- 2 \log \lambda$ will be larger than $- 2 \log \lambda'$. The test proposed in Section 4.2.2 is easier to obtain. It appears that for small m the test discussed in Section 4.2.2 should be adequate.

The test proposed in Section 4.2.2 appears to give satisfactory results, as compared with the L.R. test, for two examples considered in (i) Section 4.4.3, and (ii) Shukla (1972b). The calculations are relatively easier than for the L.R. test. For these two examples the M.L. estimates converge quickly but in some cases convergence can be very slow.

The test in Section 4.2.2 is also very sensitive to the changes in covariances. We have assumed that ϵ_{ij} and $\epsilon_{ij'}(j \neq j')$ are independent. However, the test is not affected when ϵ_{ij} have the same covariance. In the variety trials the same environmental factors may produce variabilities of different magnitudes in different varieties (as considered in Chapters 2 and 3) and thus ϵ_{ij} will have heterogeneous variances as well as heterogeneous covariances. Under these circumstances the procedure considered in Section 4.2.2 provides a simultaneous test for the homogeneity of variances and covariances. The conventional AOV tests are applicable only if variances and covariances are homogeneous. When this test fails to accept the null hypothesis then Hotelling's T^2 statistic should be used for testing the equality of all the means simultaneously as suggested by Graybill (1954). Differences between two variety means should be tested by using Student's paired t-test.

4.2.5 Test for the stability of varieties

When all the varieties are not equally variable for their interaction variances then it is of interest to examine whether for some varieties the interaction variance $\sigma_j'^2$ is negligibly small. An exact test is difficult to obtain but an approximate test can be easily obtained using the expression for the variance of $\hat{\sigma}_j^2$ given in (4.1.3). When m is large,

$(n-1)\ \hat{\sigma}^2_j/\sigma^2_j$ will be distributed approximately as χ^2 on $(n-1)$ D.F. ; thus under the hypothesis $\sigma'^2_j = 0$ the statistic F^* given by

$$F^* = \hat{\sigma}^2_j/\hat{\sigma}^2_0 \tag{4.2.18}$$

will have approximately a F-distribution on $(n-1)$ and n_0 D.F. This result is due to Johnson (1962).

Following similar methods for the extended model considered in (4.1.4) the equality of $\hat{\theta}^2_j$ with $\hat{\sigma}^2_0$ can be tested by the ratio

$$\hat{\theta}^2_j/\hat{\sigma}^2_0\ . \tag{4.2.19}$$

This ratio approximately follows the F-distribution on $(n-2)$ and n_0 D.F. If for some of the varieties the test in (4.2.19) gives an insignificant result while the corresponding statistic in (4.2.18) gives a significant result, it may be inferred that the instability is due to an effect of external variables Z_i. Such information may be of some use to a breeder.

4.3 Relationship with other works

4.3.1 Estimation of stability variances

Baker (1969) estimated the parameters σ^2_j by

$$\sum_i r^2_{ij}/(n-1)$$

after putting constraints on η_{ij} i.e. $\sum\limits_j \eta_{ij} = 0$ for all i . In the method proposed in Section 4.1.1 no such constraint is necessary. In our notation $\sum_i r_{ij}^2/(n-1)$ estimates

$$\left[(m-2)\sigma_j^2 + \bar{\sigma}^2\right]/m$$

where

$$\bar{\sigma}^2 = \sum_j \sigma_j^2/m \ .$$

4.3.2 Test of hypothesis

A more powerful test for comparing any two varieties, usable only when data from more than two varieties are available, has been described in Section 4.2.2. Very recently a similar method has been proposed by Grubbs (1973) in the context of instruments. Johnson (1962) also suggested an approximate test of H_0^* based on the ratio

$$F = \frac{\sum_i (y_{i1} - \bar{y}_{i.} - \bar{y}_{.1} + \bar{y}_{..})^2}{\sum_i (y_{i2} - \bar{y}_{i.} - \bar{y}_{.2} + \bar{y}_{..})^2} \ .$$

Han (1969) considered a simple test of the equality of all σ_j^2's. He considered the multiple correlation coefficient of $\bar{y}_{i.}$ on $y_{ij} - \bar{y}_{i.}$ $(j = 2, 3, \ldots, m)$ R, say: under H_0, R should be zero. The usual test based on an F-statistic can be used for this purpose. Shukla (1972b) and McIntyre (1972) have pointed out that the power of the above test depends on σ_e^2 as does that of the test described in Section 4.2.1. This is undesirable as persons working with different population of e_i

may arrive at different results. The power of the test considered in Section 4.2.2 does not depend on σ^2_e. It is also independent of any assumption on whether e_i are random or fixed effects. When e_i are very variable (σ^2_e large) the test considered in Section 4.2.2 will be more powerful than Han's (1969) test.

4.3.3 Relationship between regression and component of interaction approach

To examine the relationship between the regression and component of interaction approach we shall consider the model (3.2.1). Putting $V(\delta_{ij}) = \sigma^2_{\delta j}$ and taking the expectations of the numerator and denominator independently one obtains

$$\hat{b}_j \to \frac{(b_j - \bar{b})\sigma_{ez} + \bar{b}(b_j - \bar{b})\sigma^2{}_z + (\sigma^2_{\delta j} - \bar{\sigma}^2_\delta)/m}{\sigma^2_e + 2\bar{b}\,\sigma_{ez} + \bar{b}^2\sigma^2_z + \bar{\sigma}^2_\delta/m} \qquad (4.3.1)$$

where

$$\bar{\sigma}^2_\delta = \sum_j \sigma^2_{\delta j}/m .$$

Consider few particular cases of (4.3.1). If one puts $z_i = e_i$

$$\hat{b}_j \to \frac{b_j - \bar{b} + \dfrac{(\sigma^2_{\delta j} - \bar{\sigma}^2_\delta)}{m(1 + \bar{b})\,\sigma^2_e}}{1 + \bar{b} + \dfrac{\bar{\sigma}^2_\delta}{m(1 + \bar{b})\,\sigma^2_e}} . \qquad (4.3.2)$$

Thus $\hat{b}_j$ are also affected by the heterogeneity among $\sigma^2_{\delta j}$. If σ^2_e is large its effect will be negligible and one obtains the same result as in (3.2.5). But of σ_e^2 is relatively small then $\hat{b}_j$ will be affected by deviations $(\sigma^2_{\delta j} - \bar{\sigma}^2_{\delta})$ and thus magnitude and sign both are affected. If $(b_j - \bar{b})$ and $(\sigma^2_{\delta j} - \bar{\sigma}^2_{\delta})$ are of opposite sign then this may give misleading results.

A similar situation arises when $\sigma_{ez} = 0$ and σ^2_e is relatively small the expression in (4.3.1) shows that it may give misleading results. The case when $\sigma^2_{\delta j}$ are homogeneous has been considered in the previous section.

Now using the model (3.2.1) $\hat{\sigma}^2_j$ can be estimated by using (4.1.1) and this gives

$$E(\hat{\sigma}^2_j) = \frac{m}{m-2}\left[(b_j - \bar{b})^2 - \frac{\sum_j (b_j - \bar{b})^2}{m(m-1)}\right]\sigma^2_z + \sigma^2_{\delta j} . \qquad (4.3.3)$$

If m is reasonably large then the first term in r.h.s. of (4.3.3) is always positive and thus σ^2_j gives the sum of two heterogeneity components. As far as measure of variability is concerned σ^2_j seems to give a better measure than b_j as later gives only a measure of linear dependence and that too can sometimes give misleading results as discussed above.

There is a clear cut advantage in using b_j as a measure of variability (when z_i and e_i are highly correlated) that it gives the direction as well as magnitude. σ^2_j gives only magnitude as it is free of sign. Use of both measures may be more informative and $\sigma^2_{\delta j}$ can also be estimated by the method discussed in Section 4.1.2.

4.3.4 Effect of unequal covariances on the estimates of stability parameters

In Section 4.1 we have assumed that

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{ij'}) = 0 \quad \text{for } j \neq j' \ ;$$

in the present section we shall consider the effect of unequal covariances on the estimators $\hat{\sigma}^2_j$. Let us assume that

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{ij'}) = \sigma_{jj'} \ .$$

Using the estimator $\hat{\sigma}^2_j$ as given in (4.1.1) we obtain

$$E(\hat{\sigma}^2_j) = \frac{m}{m-2} \left[\sigma_{jj} - 2\bar{\sigma}_{j.} + \bar{\sigma}_{..} - \frac{\Sigma\sigma_{jj}/m - \bar{\sigma}_{..}}{m-1}\right]$$

where

$$\bar{\sigma}_{j.} = \sum_{j'} \sigma_{jj'}/m \ ; \quad \bar{\sigma}_{..} = \sum_{j,j'} \sigma_{jj'}/m^2 \ .$$

Thus the estimators $\hat{\sigma}^2_j$ are affected by unequal amounts depending upon $\bar{\sigma}_{j.}$.

It is not possible to estimate all the $\sigma_{jj'}$ when the e_i are unknown but it is possible to estimate linear combinations of $\sigma_{jj'}$.

Now consider a few particular cases of interest:

i) $\sigma_{jj'} = 0$ for $j \neq j'$.

Then

$$E(\hat{\sigma}^2_j) = \sigma_{jj} \ ;$$

as expected. This case has already been discussed in Section 4.1.

ii) $\sigma_{jj'} = c$ for $j \neq j'$.

$$E(\hat{\sigma}_j^2) = \sigma_{jj} - c ;$$

this means that all $\hat{\sigma}_j^2$ are equally biased by an amount c.

iii) $\eta_{ij} = b_j z_i$.

In this case

$$\sigma_{jj} = b_j^2 \sigma_z^2 + \sigma_\xi^2$$

$$\sigma_{jj'} = b_j b_{j'} \sigma_z^2 \qquad \text{for } j \neq j' .$$

Under these assumptions

$$E(\hat{\sigma}_j^2) = \frac{m}{m-2} \left[(b_j - \bar{b})^2 - \frac{\Sigma(b_j - \bar{b})^2}{m(m-1)} \right] \sigma_z^2 + \sigma_\xi^2 .$$

This has also been considered in the previous section. Now suppose that one of the varieties, say variety 1, is not affected by the unknown variate $z(b_1 = 0)$, whilst the other varieties are equally affected ($b_2 = b_3 = \dots = b_m = b$). Then using the above expressions we have

$$E(\hat{\sigma}_1^2) = b^2 \sigma_z^2 + \sigma_\xi^2$$

$$E(\hat{\sigma}_2^2) = E(\hat{\sigma}_3^2) = \dots\dots = E(\hat{\sigma}_m^2) = \sigma_\xi^2 .$$

This will certainly lead us to believe that the first variety is more variable than the others. In general the effect of unequal covariances will not be as extreme as in this particular case. When a large number of varieties are considered the $\bar{\sigma}_{j.}$ are likely to be approximately equal and each $\hat{\sigma}_j^2$ will thus be biased by an equal amount. When unequal covariances among the η_{ij} are suspected then the methods discussed in Chapter 5 will be more appropriate. When the z are known the methods discussed in Section 4.1.2 should be applied.

4.3.5 Relative merits of different concepts of stability

In this chapter we have called a variety stable if its yield differs from site to site by an additive constant denoting the site effect and an experimental error. This has been expressed mathematically by writing

$$\sigma_j^2 = \sigma_0^2 . \tag{4.3.4}$$

Another possible way of defining a stable variety is that its yield does not differ from site to site apart from an experimental error. Expressing this mathematically by using the notation of the previous section, a variety is called stable if

$$\sigma_{jj} = \sigma_0^2 ; \tag{4.3.5}$$

where σ_{jj} is the variance of the jth variety yield over the population of sites. This definition of stability in (4.3.5) is similar to that of Tai (1971).

According to the definition in (4.3.4) a stable variety will have different yields at different sites but it will differ in such a way that its deviation from the average of all the varieties in that the site will remain consistent from site to site. Thus a farmer using such a variety is assured of obtaining a consistently higher (or lower) yield than the average of all the varieties at his farm. This difference will depend on the varieties included in the trial but when η_{ij}'s are independent then σ_j^2 is independent of the varieties included in the trial. When η_{ij}'s are not independent then σ_j^2 depends on other varieties and this has already been discussed in the preceding section. Sometimes it may be possible to study the causes giving rise to instability and if some environmental factors are found responsible then this information can be used in recommendation of varieties for specific purposes.

According to the definition in (4.3.5) the interaction component η_{ij} works in such way that at high productive sites it pulls down the yield towards the average, and at low productive sites it raises the yield towards the average. In practice it may be difficult to find a variety with constant yield. However, it is possible that some varieties may have relatively larger σ_{jj} than the others. If a variety with relatively small σ_{jj} is found and has a high average yield then the farmer is assured of getting a high yield whatever the productivity of his site may be. Again σ_{jj} can also give misleading results. For example a variety can lodge at high productive sites in such a way that its yield remains constant from site to site. According to this definition of stability such a variety will be more stable than the other varieties which do not lodge. If a variety with relatively small σ_{jj} is found but its average yield is not far below the other varieties then it is likely that such a variety will give higher yield than others at low productive sites. This will correspond to a relatively smaller β_j for the model considered in Chapter 2.

It may be of some interest to compare the relative variabilities σ_{jj} for different varieties. For testing the equality of any two σ_{jj} the test given in Section 4.2 can be used. The test remains valid even if η_{ij}'s are correlated. This test can be generalised for more than two varieties by considering the multiple correlation coefficient of $\bar{y}_{i.}$ on $y_{i1}-\bar{y}_{i.}$, $y_{i2}-\bar{y}_{i.}$, ..., $y_{i\overline{m-1}}-\bar{y}_{i.}$, under the assumption that η_{ij}'s have equal covariances. This has been considered by Han (1969). When covariances are unequal then no simple test of equality of all σ_{jj} is available and it will be more appropriate to test the equality of all σ_{jj} and all $\sigma_{jj'}, (j \neq j')$ by the method proposed by Wilks (1946).

4.4 An example

4.4.1 Estimation of σ_j^2

The data of Table 2.5.1 will now be used to demonstrate the different methods described in this section. Estimates of σ_j^2 obtained by both methods described in Section 4.1 are given in Table 4.4.1 under the heading $\hat{\sigma}_j^2$ and $\hat{\hat{\sigma}}_j^2$. Estimates of σ_{jj} , denoted by s_{jj} , are also given.

TABLE 4.4.1

STABILITY VARIANCES

	Variety No.	1	2	3	4	5	6	7	8
MINQUE	$\hat{\sigma}^2_j$	0.112*	0.029	0.090*	0.186*	0.336*	0.026	0.080*	0.254*
Max. Lik.	$\hat{\hat{\sigma}}^2_j$	0.086	0.048	0.099	0.120	0.358	0.033	0.048	0.276
	$\hat{\sigma}^2_j$ *	0.103	0.047	0.075	0.117	0.224	0.022	0.063	0.123
	s_{jj}	1.311	1.248	1.241	1.261	1.802	1.249	0.964	1.978

$\hat{\sigma}^2 = 0.139$; $\hat{\sigma}^2_0 = 0.025$

The estimates $\hat{\sigma}^2_j$ and $\hat{\hat{\sigma}}^2_j$ do not differ much; σ^2_j's marked with asterisks are found significant against $\hat{\sigma}^2_0$ showing that most of the varieties have significant σ^2_j .

4.4.2 Test of equality of pair of variances

In this section we use the method described in Section 4.2.1. Two varieties Ymer (1) and Maris Mink (5) have been chosen for this purpose. We are interested in testing whether $\sigma^2_1 = \sigma^2_5$, a particular case of Section 4.2.1 with k = 1. We have

$$s_{11} = 1.3114 ,$$

$$s_{55} = 1.8023 ,$$

$$s_{15} = 1.2891 ,$$

---

* These values are based on the fifteen sites for which rainfall data were available (See Table 2.5.1).

$$r = \frac{s_{11} - s_{55}}{[(s_{11} + s_{55} - 2s_{15})(s_{11} + s_{55} + 2s_{15})]^{\frac{1}{2}}} = -0.2812 ,$$

$$t = \left| r\sqrt{\frac{(n-2)}{1-r^2}} \right| = 0.2812 \sqrt{\frac{18}{1-0.0791}} = 1.2432 .$$

This value of t is not significant at 5% level, so that the hypothesis that $\sigma_1^2 = \sigma_5^2$ is not rejected.

To calculate confidence limits we require

$$a = s_{55} - s_{15} = 0.5132 = \hat{\sigma}_5^2$$

$$b = s_{11} - s_{15} = 0.0223 = \hat{\sigma}_1^2$$

$$P = t^2_{\alpha/2} \; (s_{11}s_{55} - s_{15}^2)/(n-2) = 0.1721 .$$

The estimate of k is given by

$$\hat{k} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_5^2} = 0.0434 .$$

Hence lower and upper limits are given by

$$K_L = \frac{b - \sqrt{P}}{a + \sqrt{P}} = -0.4230 \; ; \; K_U = \frac{b + \sqrt{P}}{a - \sqrt{P}} = 4.4421 .$$

As there are big differences among sites the above test is not very powerful. We have also used the method described in Section 4.2.2.

For this purpose take

$$g = y_1 - y_5$$

$$f = y_1 + y_5 - 2\bar{y}_{(6)}$$

where $\bar{y}_{(6)}$ is the mean of the remaining six varieties on a particular site. The correlation-coefficient between g and f should be zero when $\sigma_1^2 = \sigma_5^2$.

$$s_{gg} = 0.5355 \; ; \; s_{ff} = 0.3653 \; ; \; s_{gf} = -0.2246$$

$$r_{gf} = \frac{-0.2246}{\sqrt{0.5355 \times 0.3653}} = -0.508$$

$$t = \left| r_{gf}\sqrt{\frac{18}{1-r_{gf}^2}} \right| = 2.502^* .$$

This gives a significant value showing that σ_1^2 and σ_5^2 differ at 5% level.

Using the approximate method suggested by Johnson (1962) one obtains

$$Q = \frac{\sum_i (y_{i5} - \bar{y}_{i.} - \bar{y}_{.5} + \bar{y}_{..})^2}{\sum_i (y_{i1} - \bar{y}_{i.} - \bar{y}_{.1} + \bar{y}_{..})^2} = \frac{5.1166}{1.925} = 2.66^* .$$

This should be tested by $F_{(19,\ 19)}$. The calculated Q is significant.

The first test, which utilises information from only two columns, is obviously not very sensitive because of large differences among sites (large σ_e^2). The second test, which utilises information from the remaining columns to eliminate the differences among sites, gives a more powerful test. Using information from more columns does not necessarily increase the power of the test as has been already discussed in Section 4.2.2.

The test suggested by Johnson (1962) removes the variability due to sites but introduces a correlation between the numerator and denominator of the test statistic Q. The distribution of Q then only approximately follows the F-distribution. However, when m is large the correlation should be small and the test adequate.

4.4.3 Testing the equality of all variances

We now illustrate the method of Section 4.2.2. The multipliers of contrasts chosen are given by the matrix L_8 such that

$$L_8 = (l_{ij}); \quad i = 1, 2, \ldots, 7 \; ; \quad j = 1, 2, \ldots, 8 \; ;$$

$$l_{ij} = 1/\sqrt{i(i+1)} \, , \quad i \leq j$$

$$l_{ij} = -i/\sqrt{i(i+1)} \, , \quad j = i + 1$$

$$l_{ij} = 0 \; . \qquad j > i + 1$$

The estimate of variance-covariance matrix of the contrasts is given by T, where

$$T = \frac{1}{19}\begin{bmatrix} 0.8437 & & & & & & \\ 0.7710 & 2.3335 & & & & & \\ -0.6267 & -0.0198 & 2.1625 & & & & \\ 0.9178 & 0.5822 & -1.6089 & 6.4440 & & & \\ -0.0928 & -0.0073 & 0.2793 & -0.3754 & 0.9042 & & \\ -0.2132 & 0.0810 & 0.6755 & -1.9767 & 0.4340 & 1.2795 & \\ 0.9561 & 0.4008 & -2.2547 & 3.4433 & -0.0054 & -1.1943 & 4.5152 \end{bmatrix}$$

$$|T| = 6.276 \times 19^{-7} ,$$

$$tr(T) = 0.9728 ,$$

$$\lambda = \frac{|T|^{10}}{\frac{tr(T)}{7}^{70}} ,$$

$$\log \lambda = -70 \log 19 + 10 \log 6.276 - 70 \log 0.139$$

$$= -50.113$$

$$\nu = 0.866$$

$$-2\nu \log \lambda = 86.796 .$$

As the tabulated 5% value of χ^2_{27} is 40.113 there is a strong evidence of heterogeneity of variances.

For the likelihood ratio test described in Section 4.2.2 the value of log (Ω) after substituting the values of $\hat{\hat{\sigma}}^2_j$ one obtains

$$\log (\Omega) = 86.253 .$$

Under H_0 the estimate of σ^2 is given by the interaction mean square in the conventional analysis of variance. We have

$$\hat{\sigma}^2 = 0.139 ,$$

$$-\frac{2}{n-1} \log L(\omega) = 7 \log \hat{\sigma}^2 + 7 ,$$

so that

$$\log L(\omega) = 64.246 .$$

The calculated value of

$$-2 \log \lambda' = 44.014$$

and this should be tested against χ^2 on 7 D.F. ($\chi^2_7 = 14.067$ at 5%). There is evidence of heterogeneity among variances.

To demonstrate the application of these results in practical cases we have considered the results of three varieties i.e. Universe, Maris Mink and Mazurka.

Variety Name	Mean (t/ha)	$\hat{\sigma}^2_j$
Universe	5.82	0.254
Maris Mink	5.57	0.336
Mazurka	5.72	0.026

From the above results it is apparent that Universe should always be preferable to Maris Mink because the former has higher mean and smaller stability variance. It is not easy to choose between Mazurka and Universe. Mazurka seems to be a very stable variety as indicated by small $\hat{\sigma}_j^2$ whereas Universe is associated with a larger $\hat{\sigma}_j^2$. The larger stability variance is partly due to its dependence on the productivity of sites as indicated by a large value of $1 + \beta_j$ (Table 2.5.2) and on the rainfall* (Table 3.3.2). Mazurka is not affected by these variations in addition to an additive effect which is common to all the varieties. Thus, for general purposes Mazurka seems to be a more reliable variety but Universe should be preferred in the regions of average and high productivity and lower rainfall.

The general procedure and conclusion can now be summarised as follows. When a variety with a high stability variance (like Universe and Maris Mink in this example) is found then its possible relationship with productivity and external environmental factors (like rainfall in this example) should be investigated by the methods described in Chapters 2 and 3. If such dependence is found then it may be possible to accept the variety for specific purposes. In the absence of a relationship with external factors varieties with high stability variance will usually only be recommended if they have some special feature, such as resistance to disease, high malting quality etc.

---

* The $\hat{\sigma}_j^2$ of Table 4.4.1 give the stability variances after eliminating the effects of rainfall, using the technique discussed in Section 4.1.2. There is a considerable reduction in the variances for Imber (4), Maris Mink (5) and Universe (8) but Maris Mink remains the most variable and Mazurka is least variable variety.

CHAPTER 5

APPLICATION OF EMPIRICAL METHODS IN THE DATA ANALYSIS AND USE OF GRAPHICAL TECHNIQUES

5.0 Introduction

In the analysis of experiments involving qualitative factors the first step usually is to start with a simple model, and if possible to test its goodness of fit. Usually the simplest model is an additive model. Under certain mild conditions it is easy to obtain efficient estimates of the parameters of an additive model and test certain hypotheses about them. Interpretation of the parameters of an additive model is relatively easy. When an estimate of the experimental error is available it is possible to test the departure of the data from the simple additive model. When the departure from the additive model is considerable there are several courses open for the data analyst. The first is just to stop at this stage, as any further analysis of the departure from additivity, which shall be named as interaction, does not add much to the understanding of the nature of the effects of factors under consideration. In certain circumstances these interactions are taken as error to test the hypothesis about the additive effects as suggested by many authors e.g. Fisher (1949, Section 65). The second course is to transform the data with the help of non-linear, monotonic transformation of observations. Sometimes transformed data can be represented adequately by an additive model; the analyses and interpretation is then straightforward. In general, this procedure does not help in understanding the nature of effects and interactions in the scale (of measurement) in which experimenter is most interested and usually this scale is one in which observations are recorded, or a simple transformation of them. A very good account of the theory of transformation is given by Box and Cox (1964). Usually transformations are used to make the statistical analysis simple rather than for understanding the actual nature of interactions. In many cases interactions are such that it is not possible to make them fit into an additive model by a simple transformation of the observations.

The third course is to extend the additive model to take into account the non-additive and non-linear terms. In particular we shall consider the use of multiplicative models. The first attempt towards this direction was made by Fisher and Mackenzie (1923), who considered the responses to different levels of manure on different varieties of potatoes. They found the departure from an additive model was not significant, though a multiplicative model gave a slightly better fit. As the calculations involved in the efficient estimation of parameters were heavy, these methods did not come into common practice. Tukey (1949) suggested a simple test of significance for a particular type of multiplicative term. If this multiplicative term is found significant he suggested methods for transforming the data to fit into an additive model. Mandel (1961) extended Tukey's test for a more generalised type of non-additivity.

Williams (1952), following the method of Fisher and Mackenzie (1923) discussed the estimation of the multiplicative parameters. Tukey

(1962) suggested the use of Factor Analysis methods. Gollob (1968) and Mandel (1969; 1971) using similar empirical methods discussed the application in factorial experiments in a great detail. Again starting with a matrix of observation Y of order n × m the hypothetical model can be written as follows

$$y_{ij} = \sum_{l=1}^{s} \phi_{il} \theta_{jl} + \varepsilon_{ij} \quad . \tag{5.0.1}$$

The above model can be taken as a generalised version of the additive model (1.3.1). In this model μ, e_i, v_j and η_{ij} are combined and expressed as a sum of s product of row terms ϕ_{il} and column terms θ_{jl}. A particular case of this with s = 1 was considered by Fisher and Mackenzie (1923). The main difference between the model (5.0.1) and that considered in Factor Analysis is that in the latter, rows are assumed to be a random sample from an infinite population. Consequently ϕ is treated as a random variable with an unknown mean and variance. Applications with fixed rows have been considered by Whittle (1953) and Anderson and Rubin (1956). Whether rows are taken as fixed or random variables, the final estimates are very similar. When ϕ_{il} are taken as random variables then, strictly speaking, it is not possible to estimate ϕ_{il}. However, this can be done by first considering ϕ_{il} as random variables and estimating θ_{jl} and then substituting the estimated value of θ_{jl} and estimating ϕ_{il} as if they are fixed parameters. This is analogous to the methods used in Factor Analysis for estimating loadings and factors scores where ϕ_{il} corresponds to the factor scores and θ_{jl} to the factor loadings. There are many indeterminacies in the model (5.0.1) and for the estimation of the parameters it will be necessary to put some constraints.

The use of a multiplicative model (5.0.1) rather than an additive model sometimes gives a great insight in studying the nature of interaction. As we shall see later that multiplicative models help considerably in reducing the large amount of data into smaller number of statistics. These statistics may reveal certain structure (if existing) which may not be otherwise be apparent.

In the following we shall consider a more general version of the model (5.0.1) which will enable us to use the well known results of canonical analysis.

5.1 Description of the model

For present purpose we shall assume that the environments are controlled and they are given as quantitative measure by a matrix Z of order nxp . The treatment of qualitative levels will be considered as a particular case of this general approach. As before we shall represent elements y_{ij} of the matrix Y by different models as they will be used later for different purposes. We shall consider two cases of the model

$$y_{ij} = \mu + v_j + b_{1j}z_{i1} + b_{2j}z_{i2} + \dots + b_{pj}z_{ip} + \epsilon'_{ij} \qquad (5.1.1)$$

which can also be written as

$$Y'_i = \mu 1' + V' + Z'_i B_C + \epsilon'_i \quad (5.1.2)$$

The other case of interest is to extract another parameter e_i from B such that

$$Y'_i = (\mu + e_i)1' + V' + Z'_i B_{RC} + \epsilon'_i \quad (5.1.3)$$

where

$$Y'_i = (y_{i1}, y_{i2}, \ldots, y_{im})$$

$$Z'_i = (z_{i1}, z_{i2}, \ldots, z_{ip}) \;.$$

The regression coefficients B in (5.1.2) and (5.1.3) are not the same and to distinguish them from each other they are denoted by B_C and B_{RC}. To express the model in (5.1.1) by the combination of multiplicative terms we express B such that

$$B = \lambda_1 \phi_1 \theta'_1 + \lambda_2 \phi_2 \theta'_2 + \cdots + \lambda_s \phi_s \theta'_s \quad (5.1.4)$$

where $\lambda_1, \lambda_2, \ldots, \lambda_s$ are scalar multipliers and ϕ_l and θ_l $(l = 1, 2, \ldots, s)$ are column vectors of order p and m, respectively; s is the rank of matrix B ($s \leq$ smaller of m and p). These parameters have been arranged in such a way that most of the variation in y_{ij} is explained by the first term in (5.1.4), the next highest contribution comes from the second term, and so on. The model in (5.1.1) can also be written as

$$Y'_i = \mu 1' + V' + \lambda_1 q_{i1} \theta'_1 + \lambda_2 q_{i2} \theta'_2 + \dots \lambda_s q_{is} \theta'_s + \epsilon'_i \qquad (5.1.5)$$

where

$$q_{il} = Z'_i \phi_l = \sum_{k=1}^{p} z_{ik} \phi_{kl} \; ; \; l = 1, 2, \dots, s.$$

The q_i are weighted linear functions of the z_{ik}. We suppose that the first new external variate q_{i1} accounts for the largest amount of variation in y_{ij} ; q_{i2} accounts for the next largest and so on.

In the same way using the structure (5.1.4) one can write (5.1.3) as

$$Y'_i = (\mu + e_i)1' + V' + \lambda_1 q_{i1} \theta_1' + \lambda_2 q_{i2} \theta'_2 + \dots + \lambda_s q_{is} \theta'_s + \epsilon'_i \qquad (5.1.6)$$

5.2 Estimation of the parameters

For estimation purposes we shall consider models (5.1.5) and (5.1.6), separately. First consider the model in (5.1.5). As in Chapter 3 we take z_{ij} as deviations from their means and thus it follows that

$$\sum_i q_{il} = 0 \qquad (5.2.1)$$

for all l.

As we are interested in the parameters which occur in multiplicative terms in (5.1.4) we shall work only with a matrix Y_C of deviations of y_{ij} from the column means. This has the same effect as estimating μ and v_j by the usual estimators

$$\hat{\mu} = \bar{y}_{..} \; ; \; \hat{v}_j = \bar{y}_{.j} - \bar{y}_{..} \; .$$

Let Y'_{iC} denote the i^{th} row of Y_C then

$$Y'_{iC} = \lambda_1 q_{i1} \theta'_1 + \lambda_2 q_{i2} \theta'_2 + \dots + \lambda_s q_{is} \theta'_s + \epsilon'_{iC} \qquad (5.2.2)$$

Thus for estimation purposes we impose the following additional conditions to determine the parameters uniquely.

$$\theta'\theta = \underset{m\times m}{I} \; ; \; Q'Q = \phi'(Z'Z)\phi = \underset{s\times s}{I} \qquad (5.2.3)$$

where

$$\theta' = (\theta_1, \theta_2, \dots, \theta_s) \; : \; Q = (q_{il}) \; .$$

The conditions on θ and Q are such that they are orthogonal among themselves. These conditions have been introduced only to give an unique solution. It is possible to estimate the parameters under any other adequate number of conditions so as to give an unique solution.

Now multiplying Y'_{iC} by the column vector θ_1 and using the conditions in (5.2.3) one obtains

$$Y'_{iC}\theta_1 = \lambda_1 q_{i1} + \epsilon'_{iC}\theta_1 \; . \qquad (5.2.4)$$

We have now the problem of calculating estimates of θ_1 and ϕ_1 with maximum covariance (or correlation under the conditions in (5.2.3)). For this we can either use the well known results of canonical correlation analysis or apply least squares theory directly to y_{ij} under the constraints in (5.2.3). In the following we have used the former. $\hat{\theta}_l$ is the l^{th} latent vector of the matrix

$$Y'_c Z(Z'Z)^{-1}Z'Y_c \tag{5.2.5}$$

and $\hat{\lambda}_l^2$ is the corresponding latent root. $\hat{\phi}_l$ can be easily obtained as

$$\hat{\phi}_l = \frac{(Z'Z)^{-1}Z'Y_c\hat{\theta}_l}{\hat{\lambda}_l} \; . \tag{5.2.6}$$

$\hat{\lambda}_l^2$ is the sum of squares accounted by each term and the total sum of squares due to all λ's i.e. $\Sigma\lambda^2_l$ is the same as sum of square due to the regression coefficients b_{kj}.

$$\sum_{l=1}^{s} \hat{\lambda}_l^2 = Tr(Y'_c Z(Z'Z)^{-1}Z'Y_c)$$

and

$$\hat{B}_c = \hat{\lambda}_1\hat{\phi}_1\hat{\theta}_1' + \hat{\lambda}_2\hat{\phi}_2\hat{\theta}_2' + \dots\dots + \hat{\lambda}_s\hat{\phi}_s\hat{\theta}_s' \; .$$

Now consider the extended model (5.1.3) where an additive environmental term has been added. We have

$$Y_i' = (\mu + e_i)1' + V' + \lambda_1 q_{i1}\theta_1' + \lambda_2 q_{i2}\theta_2' + \dots + \lambda_s q_{is}\theta_s' + \epsilon_i' \, . \tag{5.2.7}$$

Again using the usual estimates of e_i i.e.

$$\hat{e}_i = \bar{y}_{i.} - \bar{y}_{..}$$

puts additional constraint on θ's i.e.

$$\sum_j \theta_{jl} = 0$$

for all l.

Following the same argument as before and denoting the matrix Y_{RC} by

$$Y_{RC} = (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})$$

the estimate of the parameters can be obtained by substituting Y_{RC} instead of Y_C in (5.2.5) and (5.2.6). The sum of square due to all the terms λ's is

$$\sum_{l=1}^{s-1} \lambda_l^2 = \mathrm{Tr}(Y'_{RC}Z(Z'Z)^{-1}Z'Y_{RC})$$

and this is same as the sum of square due to heterogeneity in regression as given in Table 3.1. The rank s will be smaller by 1 than in the case of model (5.1.5) because of the additional constraint.

5.3 Test of hypothesis

The problem now arises of how many terms in the model (5.1.5) and (5.1.6) should be included to explain the data adequately. For simplicity of interpretation it would obviously be convenient if most of the variation were accounted for by the first two or three terms. The problem amounts to testing for the number of non-zero λ's . In the following we shall use the results obtained by Bartlett (1951) in connection with canonical correlation.

If r, the number of within environment replicates, were large and λ_l^2 zero, the corresponding $\hat{\lambda}_l^2$ would be asymptotically distributed as

$$\hat{\lambda}_l^2 \sim \sigma_0^2 \chi^2_{(m + p + 1 - 2l)}$$

in case of model (5.1.5). The approximate test for each latent root λ_l^2 is

$$F^* = \frac{\hat{\lambda}_l^2}{\hat{\sigma}_0^2(m + p + 1 - 2l)} \sim F_{(m + p + 1 - 2l, n_0)} .$$

When model (5.1.6) is used, m in the above expression should be replaced by $m - 1$.

However, in general r is not large and the above test will give too many significant terms. Possibly a better procedure is to proceed sequentially. First test the whole regression sum of squares, and if this is found significant, extract the first root and test the remainder, and so on as is usually done in regression analysis for testing the significance of an additional term. The cumulative D.F. corresponding to residuals at each stage are given in Table 5.3.

Table 5.3

Degrees of freedom for the sum of squares for test of significance of number of terms in multiplicative model

	Model (5.1.5)		Model (5.1.6)	
l	D.F.	S.S.	D.F.	S.S.
0	mp	Reg.S.S.	(m-1)p	Reg.S.S.
1	mp-(m+p)+1	Reg.S.S.$-\hat{\lambda}_1^2$	(m-1)p-(m+p-1)+1	Reg.S.S.$-\hat{\lambda}_1^2$
2	mp-2(m+p)+2	Reg.S.S.$-\hat{\lambda}_1^2-\hat{\lambda}_2^2$	(m-1)p-2(m+p-1)+2	Reg.S.S.$-\hat{\lambda}_1^2-\hat{\lambda}_2^2$
.		.	.	.
.		.	.	.
s	mp-s(m+p)+s	Reg.S.S.$-\sum_{l}\hat{\lambda}_l^2$	(m-1)p-s(m+p-1)+s	Reg.S.S.$-\sum_{l}\hat{\lambda}_l^2$

5.4 Interpretation of the parameters and the relationship with the works of other authors

In the model (5.1.5) we have estimated s orthogonal external variates (q_i) as linear combinations of the original variates (z_i) in such a way so as to account for variation in $y_{ij} - \bar{y}_{.j}$ in ascending order. This may be a useful technique to reduce the dimensions when there are many external variables. However, the interpretation and test of significance are difficult and in general may not give any advantage over regression technique. The same is true of model (5.1.6).

A particular case of the above general approach can be considered when no information on the external variables is available. In the general approach considered in the preceding section we have made no assumption about z_i. For this we shall consider a case when Z is a matrix of dummy variables of order $n \times (n - 1)$ such that

$$Z'Z = I \text{ and } ZZ' = I - \frac{11'}{n} .$$

This matrix can be constructed in infinitely many ways. By substituting in (5.2.5), the $\hat{\theta}_1$ are obtained as latent vectors of $Y_C'Y_C$. $\hat{\theta}_1$ is proportional to $(1 + \beta)$ considered in Chapter 2. $\hat{\phi}_1$ can be calculated using the results in (5.2.6) as

$$\hat{\phi}_1 = \frac{Z'Y_C\hat{\theta}_1}{\hat{\lambda}_1} .$$

The measure of the environment is given by

$$Q_1 = Z\hat{\phi}_1 = \frac{Y_C\hat{\theta}_1}{\hat{\lambda}_1} .$$

The result is the same as given by E_i in Section 2.3. This gives a measure of the effect of environment in the absence of any knowledge of external variables. The process can be further extended corresponding to the latent root $\hat{\lambda}_2^2$ and so on but we shall consider an alternative approach through model (5.1.6).

Now considering the model (5.1.6) and using Z as a matrix of dummy variable one obtains $\hat{\theta}_1$ as latent vector of matrix $Y_{RC}'Y_{RC}$ and the Q_1 given by

$$\hat{Q}_1 = \frac{Y_{RC}\hat{\theta}_1}{\hat{\lambda}_1} .$$

Alternatively Q_1 can be obtained as a latent vector of the matrix $Y_{RC}Y'_{RC}$ and the θ can be obtained from Q_1 by using relationship

$$\hat{\theta}_1 = \frac{Y'_{RC}\hat{Q}_1}{\hat{\lambda}_1} .$$

It is advantageous to use only the matrix of smaller dimensions and by the above relationships Q_1 and θ_1 can be calculated.

This particular case has been considered in great detail by Gollob (1968) and Mandel (1969, 1971). They have obtained the results directly without considering the canonical correlation approach. Gollob has also developed the test of hypothesis as given in Table 5.3 by heuristic methods. Here same results have been obtained by using canonical correlation approach and the methods suggested in Section 5.3 may give more satisfactory results. Mandel (1971) considered tests by Monte-Carlo studies when an estimate of $\hat{\sigma}_0^2$ is not available. However, in the present application a good estimate of σ_0^2 should always be available.

Mandel (1969) has considered an example in a controlled experiment where rows are quantitative levels of treatments. He has not used this information while using his method. By considering a more generalised approach in the preceding sections such information can be used in matrix Z. In this matrix one can include the desired function of the initial external values of the variables. However, such methods may be no better than simple multiple regression.

When the interaction term (η_{ij}) is such that most of it is accounted for by the multiplicative term of the row and column effects, as considered by Tukey (1949), then it will be found that θ_{j1} and q_{i1} will be proportional to v_j and e_i, respectively. This can be examined by plotting θ_{j1} and q_{i1} against v_j and e_i. In some cases the q_i may be found proportional to some external site variables, then the coefficients θ_j may be interpreted as the regression coefficients on the external variables as described in Chapter 3.

θ and Q also give coefficients of orthogonal contrasts among rows and columns, respectively, and are such that they account for the interaction sum of squares in the descending order. This property has been used in the next section in the graphical representation of distances between varieties and environments.

Hardwick and Wood (1972) also considered the general model of (5.2.7) with only one term. For estimation purposes they followed the method given by Mandel (1969, 1971). The method given here gives an extended version of the model by showing the relationship with the canonical correlation approach.

It is not always necessary to consider dummy variables with $p = n - 1$. When one has to examine the effect of qualitative factors such as soil type, sub-region etc. then p is likely to be considerably smaller than $n - 1$. By suitably choosing dummy variables the general method can be used to give a quantitative measure of effects and interactions of such qualitative factors.

5.5 Application of graphical techniques

5.5.1 Analysis of interaction

The parameters estimated in this chapter can also be used to represent data in a considerably reduced dimension and thus making graphical representation easier. For this we shall define a few useful statistics and consider their calculation and graphical representation.

Firstly, we shall consider the interaction of the pair of varieties (j,j') with all the environments in trial. The interaction sum of squares between a pair of varieties (j,j') and environments is denoted by $IV_{jj'}$, where

$$IV_{jj'} = \frac{1}{2} \sum_i (y_{ij} - \bar{y}_{.j} - y_{ij'} + \bar{y}_{.j'})^2 . \qquad (5.5.1)$$

Now consider the relationship (5.2.7). Substituting the estimates of λ, q and θ we obtain identities

$$y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..} = \lambda_1 q_{i1} \theta_{j1} + \lambda_2 q_{i2} \theta_{j2} + \cdots + \lambda_s q_{is} \theta_{js} ;$$

$$y_{ij'} - \bar{y}_{i.} - \bar{y}_{.j'} + \bar{y}_{..} = \lambda_1 q_{i1} \theta_{j'1} + \lambda_2 q_{i2} \theta_{j'2} + \cdots + \lambda_s q_{is} \theta_{j's} .$$

By taking the differences, squaring, adding (over i) and using the property that $\sum_i q^2_{il} = 1$; $\sum_i q_{il} q_{il'} = 0$ for $l \neq l'$, we obtain

$$IV_{jj'} = \tfrac{1}{2}\left[\lambda_1^2(\theta_{j1} - \theta_{j'1})^2 + \lambda_2^2(\theta_{j1} - \theta_{j2'})^2 + \dots + \lambda_s^2(\theta_{js} - \theta_{j's})^2\right] \quad (5.5.2)$$

The above expression shows that the contribution of the last few terms of (5.5.2) is small if the corresponding λ^2's are small and corresponding values of $(\theta_{j1} - \theta_{j'1})^2$ are not very large. Each variety can be represented by a co-ordinate $V_j(\lambda_1\theta_{j1}, \lambda_2\theta_{j2}, \dots, \lambda_s\theta_{js})$ in s-dimensional space and square of distance between two points gives twice the interaction sum of squares. The square of the distance of each point from the origin gives the contribution of the j^{th} variety to the interaction

$$\sum_i (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \ .$$

When most of the interaction S.S. is accounted for by the first two latent roots then a plot on a graph may be a good way of showing the pattern in interaction. This plot may reveal certain varieties which contribute most to the interaction. Alternatively, certain groups of varieties which clump together contribute very little to the interaction and most of the interaction may be due to between group comparisons. This may be a more efficient method of grouping varieties (on the basis of interaction) than the method suggested by Perkins and Jinks (1968) and Finlay and Wilkinson (1963).

Finlay and Wilkinson classified varieties on the basis of the regression coefficients $(1 + \beta_j)$ given in Chapter 2. Perkins and Jinks (1968) suggested a method of classification based on a further dimension by observing the sign of correlation coefficients of residuals after fitting $(1 + \beta_j)$ in the model described in (2.1.0). Here we extend these methods to as many dimensions as thought to be necessary.

A similar interpretation can be given to q_{il}. The interaction sum of squares for each pair of environments is given by

$$IE_{ii'} = \frac{1}{2} \sum_j (y_{ij} - \bar{y}_{i.} - y_{i'j} + \bar{y}_{i'.})^2$$

and can alternatively be written as

$$IE_{ii'} = \frac{1}{2} \left[\lambda_1^2 (q_{i1} - q_{i'1})^2 + \lambda_2^2 (q_{i2} - q_{i'2})^2 + \dots + \lambda_s^2 (q_{is} - q_{i's})^2 \right].$$

Each environment can be represented by a co-ordinate $E_i = (\lambda_1 q_{i1}, \lambda_2 q_{i2}, \dots, \lambda_s q_{is})$ in a s-dimensional space. Points E_i at a greater distance from the origin (0, 0, ..., 0) contribute more to the interaction. If some points lie far away from the origin the data for the corresponding environments should be examined carefully. This may reveal certain environments behaving quite differently from others. Reasons for such behaviour should be investigated as far as possible.

The points clumping together may form a group of environments which contribute little to interaction (within group) and most of the interaction may come from the contribution between groups. This type of information may be useful when the same site (environments) are used year after year for variety trials. If these groups remain consistent then only one member from each group may provide most of the information about interactions. However, the members within a group may not remain consistent if substantial year × site and year × variety interactions are present.

These methods are very sensitive to outliers. Once a point is found lying far away from the origin, this should be examined carefully. If there is sufficient reason to doubt the data from a particular environment, it should be rejected or corrected and the data reanalysed.

5.5.2 Distance between environments

For the model (5.1.5), we have shown that q_{il} ($i = 1, 2, \ldots, n$; $l = 1, 2, \ldots, s$) gives a weighted measure of environment. Define the distance between two environments as the total sum of squares due to two environments and interaction given by

$$DE_{ii'} = \frac{n(\bar{y}_{i.} - \bar{y}_{i'.})^2}{2} + \sum_j \frac{(y_{ij} - \bar{y}_{i.} - y_{i'j} + \bar{y}_{i'.})^2}{2} = \sum_j \frac{(y_{ij} - y_{i'j})^2}{2} .$$

Alternatively this can be written

$$DE_{ii'} = \tfrac{1}{2}\left[\lambda_1^2(q_{i1} - q_{i'1})^2 + \lambda_2^2(q_{i2} - q_{i'2})^2 + \cdots + \lambda_s^2(q_{is} - q_{i's})^2 \right]$$

where q_i are given by

$$Q_l = Z \phi_l ,$$

using the expression of model (5.1.5). Each environment can be represented by a co-ordinate in a s-dimensional space $E_i(\lambda_1 q_{i1}, \lambda_2 q_{i2}, \ldots, \lambda_s q_{is})$. If the first few λ^2's take into account most of the variation, the relative position of each environment can be represented geometrically. In the case of variety trials the geometrical representation of environments may not be very useful as most of the variation may be due to productivity differences. Of course, there is a possibility of investigating the relationship between productivity of sites and other environmental factors.

5.5.3 Distance between varieties

In the preceding section we considered a concept of distance between two environments. In this section we shall discuss the distance between a pair of varieties (j, j'); the difference between the mean values can be tested against the within environmental error or variety × environment interaction depending upon the object of the experimenter (Fisher, 1949, section 65; Cochran and Cox, 1957). The interaction sum of squares between a pair of varieties can be tested against the within environmental error, to determine whether the differences between a pair of varieties differ from environment to environment. It may be of interest to combine the above two measures in one and define the square of distance $DV_{jj'}$ given by

$$DV_{jj'} = \frac{n(\bar{y}_{.j} - \bar{y}_{.j'})^2}{2} + \sum_i \frac{(y_{ij} - \bar{y}_{.j} - y_{ij'} + \bar{y}_{.j'})^2}{2} = \sum_i \frac{(y_{ij} - y_{ij'})^2}{2}$$

Now if there is no difference between two varieties j and j' (e.g. if the same variety is taken twice), then they should not differ from each other apart from experimental error, and

$$E\,\frac{(DV_{jj'})}{n} = \sigma_0^2 \ .$$

This measure $DV_{jj'}$ gives a combined measure (mean and interaction) based on the yield of varieties in different environments. This measure may be more effective in classifying large number of varieties than one based on only interaction (IV_{jj}'). For geometrical representation we use a method similar to that used before; this time we work with the matrix

$$Y_R'Y_R$$

where Y_R is obtained from Y after deducting the row means i.e.

$$Y_R = (y_{ij} - \bar{y}_{i.}) \ .$$

Denote the latent root of this matrix by $\lambda_1^2, \lambda_2^2, \dots, \lambda_s^2$, and the latent vectors by $\theta_{j1}, \theta_{j2}, \dots, \theta_{js}$ $(j = 1, 2, \dots, m)$. It can be shown as before that

$$DV_{jj'} = \tfrac{1}{2}\left[\lambda_1^2(\theta_{j1} - \theta_{j'1})^2 + \lambda_2^2(\theta_{j2} - \theta_{j'2})^2 + \dots + \lambda_s^2(\theta_{js} - \theta_{j's})^2\right]$$

and

$$\Sigma\lambda_i^2 = \sum_{ij}(y_{ij} - \bar{y}_{i.})^2 = \text{Variety S.S.} + (V \times E)\text{S.S.}$$

As before each variety can be represented by a co-ordinate $V_j(\lambda_1\theta_{j1}, \lambda_2\theta_{j2}, \dots, \lambda_s\theta_{js})$, and the relative position of each variety in relation to the others can be geometrically represented. If some varieties clump together then they are very similar yielding varieties in all the environments considered. If the environments included in the analysis represent the universe it may be advisable to pick only one variety from each group and reject the others.

If a variety lies far away from the others it may happen for one of two reasons: (i) only means differ and differences between means remains consistent (no interaction), (ii) the means may or may not differ but differences vary (i.e. interaction present). When the former is true (no interaction) $IV_{jj'}$ will be small. This can be seen from the geometrical representation discussed in Section 5.5.1. Latter case will be true when $IV_{jj'}$ is large. In this case further investigation may be necessary to examine if some varieties do better in some environment and worse in others.

The above type of analysis will be found useful only if a large number of varieties are taken in a large number of environments. The experimenter may be interested in grouping the varieties on the basis of their yields in all the environments. A geometrical picture may make the problem of investigation of relationship between varieties and environments much easier.

The concept of distance and classification used here are similar to that in Numerical Taxonomy. The analysis of $Y'Y$ and YY' are also analogous to Q and R techniques commonly used in Numerical Taxonomy. Many other classification techniques of Numerical Taxonomy can also be usefully employed for the present purposes.

5.6 Discussion and relationship with the work described in the previous chapters

We start again with the model (1.3.1):

$$y_{ij} = \mu + e_i + v_j + \eta_{ij} + \xi_{ij} \,. \tag{5.6.1}$$

The variance-covariance matrix of y_{ij} when e_i and η_{ij} are independent and η_{ij} are independent among themselves is given by Σ, where

$$\Sigma = \begin{bmatrix} \sigma_e^2 + \sigma^2 & \sigma_e^2 & \cdots & \sigma_e^2 \\ \sigma_e^2 & \ddots & & \\ & & \ddots & \\ \sigma_e^2 & \cdots & & \sigma_e^2 + \sigma^2 \end{bmatrix} = \sigma_e^2 \, 11' + \sigma^2 I \,. \tag{5.6.2}$$

This corresponds to the case of Chapter 4 with all $\sigma_j^2 = \sigma^2$. Sometimes a simple constraint $\sum\limits_j \eta_{ij} = 0$ is imposed which produces a symmetrical structure of covariances. The above structure gives the usual covariance matrix in the analysis of variance of the mixed model.

If the η_{ij} are such that they are independent of e_i and among themselves but have unequal variances then the structure of Σ is of the type.

$$\Sigma = \begin{bmatrix} \sigma_e^2 + \sigma_1^2 & \sigma_e^2 & \cdots & \sigma_e^2 \\ \sigma_e^2 & \ddots & & \\ & & \ddots & \\ \sigma_e^2 & \cdots & & \sigma_e^2 + \sigma_m^2 \end{bmatrix} = \sigma_e^2 \, 11' + \mathrm{Diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2) \,. \tag{5.6.3}$$

This case has been considered in Chapter 4 and the method of estimation and testing of hypothesis have been discussed there.

The η_{ij} may sometimes have unequal variances and unequal covariances. This may be due to the influence of some environmental factors (possibly unknown).

The simplest case is that of linear dependence on an unknown factor, say, x_i .

$$\eta_{ij} = \beta_j x_i + \epsilon'_{ij} . \qquad (5.6.4)$$

Assume that

$$V(\delta_{ij} = \epsilon'_{ij} + \xi_{ij}) = \sigma_\delta^2 \ ; \quad \text{cov} \ (\delta_{ij}, \delta_{ij}') = 0 \ .$$

When x_i is taken to be e_i , the additive environmental factor, we obtain the variance-covariance matrix Σ , where

$$\Sigma = \sigma_e^2 \ (1 + \beta)(1 + \beta)' + \sigma_\delta^2 \ I \qquad (5.6.5)$$

and

$$(1 + \beta)' = (1 + \beta_1, 1 + \beta_2, \ldots, 1 + \beta_m) .$$

Estimation of the parameters and test of hypothesis of this structure has been discussed in Chapter 2.

Further extension of the type (5.6.4) is possible when the ϵ'_{ij} are found correlated and of unequal variances. It can be assumed that the structure is given by

$$\eta_{ij} = \beta_{1j}e_{i1} + \beta_{2j}e_{i2} + \beta_{3j}e_{i3} + \cdots + \beta_{sj}e_{is} + \xi_{ij} \tag{5.6.6}$$

such that unknown factors $e_{i1}, e_{i2}, \ldots, e_{is}$ are independent of each other. In the above model $\beta_{1j} = \beta_j$ and $e_{i1} = e_i$. However, this assumption of independence of e_i is not necessary and if there is some good reason for assigning any other variance-covariance matrix, it can be done. Interpretation of e_i may not be easy. Estimation has been discussed in Section 5.2. The structure of Σ with η_{ij} as in (5.6.6) is given by

$$\Sigma = \sigma_e^2 (1 + \beta)(1 + \beta)' + \sigma_{e2}^2 B_2 B_2' + \sigma_{e3}^2 B_3 B'_3 + \cdots + \sigma_{es}^2 B_s B_s' + \sigma_0^2 I, \tag{5.6.7}$$

where $\sigma_{e2}^2, \sigma_{e3}^2, \ldots, \sigma_{es}^2$ are the variances of $e_2, e_3, \ldots, e_s$, respectively and

$$B_k' = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{km}); \; (k = 2, 3, \ldots, s).$$

Usually the effect of the additive environmental component e_i is large and it is also of interest to consider the variation among deviations of y_{ij}'s from the site mean $\bar{y}_i$.

$$y_{ij} - \bar{y}_{i.} = v_j - \bar{v} + \eta_{ij} - \bar{\eta}_{i.} + \xi_{ij} - \bar{\xi}_{i.} \tag{5.6.8}$$

Substituting the structure η_{ij} as given in (5.6.6) we obtain

$$y_{ij} - \bar{y}_{i.} = v_j - \bar{v} + (\beta_{1j} - \bar{\beta}_1)e_{i1} + (\beta_{2j} - \bar{\beta}_2)e_{i2} + \cdots + (\beta_{sj} - \bar{\beta}_s)e_{is} + \xi_{ij} - \xi_{i.}. \tag{5.6.9}$$

Let us denote the variance-covariance of $y_{ij} - \bar{y}_{i.}$ by Σ', which is given by

$$\Sigma' = \sigma^2_{e1} B^*_1 B^{*\prime}_1 + \sigma^2_{e2} B^*_2 B^{*\prime}_1 + \ldots\ldots + \sigma^2_{es} B^*_s B^{*\prime}_s + \sigma^2_o \left(I - \frac{11'}{m}\right) \quad (5.6.10)$$

where

$$B^{*\prime}_k = (\beta_{k1} - \bar{\beta}_k, \beta_{k2} - \beta_k, \ldots, \beta_{km} - \bar{\beta}_k) \; ; \; (k = 1,2,\ldots,s).$$

Σ' has the rank $(m-1)$ (if $n > m-1$). The structure in (5.6.9) is only a hypothetical one and the interpretation of the parameters is not easy. We have used these parameters for graphical representation. In the null case when the η_{ij} are independent with equal variance then Σ' reduces to

$$\Sigma' = (\sigma^2 + \sigma^2_o)\left(I - \frac{11'}{m}\right) . \quad (5.6.11)$$

The first $(m - 1)$ latent roots Σ' under the null hypothesis are $(\sigma^2 + \sigma^2_o)$ and the last one is 0 . The latent vectors are given by the contrasts orthogonal to (1, 1,...., 1). Therefore, if the first $(m - 1)$ latent vectors of the matrix $Y'_{RC}\, Y_{RC}$ are found equal then there is no advantage in looking for the structure of type (5.6.9).

5.7 An example

The data given at the end of the Chapter 2 have been used to provide an example.

5.7.1 Analysis of interaction

As discussed in Section 5.5.1 we shall find out the latent roots and vectors of matrix $Y'_{RC} Y_{RC}$ and the first three are given in Table 5.7.1.

Table 5.7.1

Latent roots and vectors of $Y'_{RC}Y_{RC}$

l	λ_l^2	% variation	θ_{1l}	θ_{2l}	θ_{3l}	θ_{4l}	θ_{5l}	θ_{6l}	θ_{7l}	θ_{8l}
1	10.87	58.83	0.239	0.028	0.028	0.453	-0.601	0.081	0.299	-0.528
2	2.57	13.90	0.655	0.275	-0.641	-0.177	-0.097	-0.029	-0.139	0.153
3	2.52	13.66	-0.078	0.202	0.305	-0.386	-0.668	-0.007	0.136	0.496

These three roots account for more than 86% of the interaction sum of squares. As mentioned earlier these vectors can be considered as coefficients of orthogonal contrasts accounting for maximum interaction sum of squares in descending order. Looking at the coefficients of the first vector it is apparent that the most of the interaction is accounted for by comparison between two groups of varieties (5,8) and (1,4,7).

Graph 5.7.1 shows the plot of these varieties corresponding to first and second, and to first and third component, respectively.

The points plotted are $\lambda_1 \theta_{j1}$ $(j = 1, 2, \ldots, 8)$. The square of the distance of these points from the origin $(0, 0, \ldots, 0)$ gives the contribution of each variety to the interaction. It is apparent that varieties 5 and 8 contribute most of the interaction. This has already been indicated by larger values of $\hat{\sigma}_5^2$ and $\hat{\sigma}_8^2$ in Table 4.4.1. There does not appear to be any group of varieties which contribute little to the interaction (within group). There is some indication that the varieties 1, 4 and 7 tend to behave similarly. Varieties 5 and 8 are in contrast with varieties 4 and 7.

To examine, if there is any grouping among sites as suggested by the interactions, the scores (q_{i1}) have been calculated by

$$q_{i1} = \sum_j \frac{(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})\theta_{j1}}{\lambda_1}$$

and are given in Table 5.7.2. Points $\lambda_1 q_{i1}$ have been plotted in Graph 5.7.2 corresponding to first and second, and to first and third, component respectively.

Table 5.7.2

Interaction scores for sites

l	λ_l^2	% variation	1	2	3	4	5	6
1	10.87	58.83	0.140	0.288	0.242	0.132	0.300	-0.024
2	2.57	13.90	0.209	-0.180	-0.198	0.259	-0.047	-0.574
3	2.52	13.66	-0.123	-0.148	0.100	-0.034	-0.049	-0.139

l	7	8	9	10	11	12	13
1	0.123	0.176	0.146	-0.076	-0.142	0.213	0.145
2	0.301	0.154	-0.067	0.327	0.312	-0.136	0.126
3	-0.015	-0.001	0.337	-0.100	0.001	0.120	0.101

l	14	15	16	17	18	19	20
1	-0.442	-0.106	-0.046	-0.471	-0.272	-0.244	-0.082
2	0.061	0.062	-0.241	-0.109	-0.095	0.037	-0.201
3	-0.478	0.080	-0.046	0.273	0.490	0.101	-0.470

Looking at the scatter of the points there appears to be some indication of groups of points clumping together. Sites 1-6 come from the East of Scotland, 7-13 come from North of Scotland and 14-20 come from West of Scotland. It appears that sites within North and East do not contribute much to the interaction. There appears to be a larger scatter among sites in West thus contribution to interaction is large. There is some indication that relative performance of varieties is similar in the North and East. Most of the contribution to the interaction comes from differential performance between the North and East at one end and the West at the other. It is not clear why sites in the West perform so differently from those in the North and East. Looking into details it appears that there is greater incidence of diseases, lodging and leaning in the sites taken in West than in North and East. A climatic factor like rainfall might have contributed to the grouping of sites in West. We have seen in Chapter 3 that rainfall can contribute significantly towards interaction. Rainfall is generally higher in the West than in the North. No data were available from sites from the East but it is hoped that their average will not be greater than that of the North. Looking at residuals it appears that for each of the varieties 4, 5 and 8 results in the North and East contrast with those in the West. Table 5.7.3 gives the break up of sum of squares for three regions showing the interaction pattern as discussed above.

Table 5.7.3

Analysis of variance of winter wheat

Source	D.F.	S.S.	M.S.
Sites (S)	19	191.650	10.087
Varieties (V)	7	2.615	0.374
V × S	133	18.470	0.138
V × (N + E vs W)	7	6.755	0.965
V × (N vs E)	7	1.096	0.157
Remainder	119	10.619	0.089
$\hat{\sigma}_0^2$			0.025

5.7.2 Distance between varieties

Until now we have considered the performance of varieties as deviations from their respective means. In the present section we will use the method described in Section 5.5.3 to analyse the total sum of squares of differences between each pair of varieties. For this we calculate the latent roots of matrix $Y_R'Y_R$ and the co-ordinates of each variety are given in Table 5.7.4. The first three latent roots account for more than 85% of the variation and we shall only consider them.

Table 5.7.4

Distance co-ordinates of varieties

1	λ_1^2	% variation	1	2	3	4	5	6	7	8
1	12.17	57.70	0.951	0.209	0.062	1.623	-1.807	-0.006	1.030	-2.062
2	3.28	15.55	0.436	0.100	-0.484	0.341	1.288	-0.571	-0.300	-0.811
3	2.57	12.19	-0.994	-0.495	0.897	0.412	0.394	0.037	0.168	-0.419

Points corresponding to the first and second components, and to the first and third components are plotted in Graph 5.7.3. There does not appear to be any clear cut grouping but there is an indication of varieties (4,7), (1,2) and (3,6) lying closer to each other. The varieties within each pair perform similarly over all sites. Varieties 5 and 8 lie far away from each other and from the rest. This indicates that each perform differently from all other varieties. The distance of a point from the origin (0, 0, ..., 0) is the square root of the sum of squares of deviations from the site means.

5.7.3 Examination of residuals

In Section 5.2 we have calculated certain statistics, based on the residuals, for studying some possible patterns among varieties and sites, separately. It may be instructive to look for any pattern among the residuals for sites and varieties simultaneously and this can be done by examining the two-way table of residuals. When the dimensions of this table are small this may not create any special problem but when large numbers of sites and varieties are considered then it may be difficult to look for any pattern among the residuals. The parameters calculated in Section 5.2 can be used to rearrange the table (by appropriate arrangement of rows and columns) in such a way that the residuals which are similar (on the average) lie together. After rearranging the table of residuals in this way any possible relationship with characteristics of sites and varieties can be examined easily and may be helpful in recognising any possible pattern among the residuals.

For this we shall consider the data given in Chapter 2 and arrange the varieties (columns) and sites (rows) on the basis of magnitudes of θ_{j1} and q_{i1} and the Graphs 5.7.1 and 5.7.2 in such a way that points lying close to each other are nearer to each other in this table. More precise ordering of sites and varieties can be obtained by using some distance algorithm but this is not very necessary in the case we are considering and the above approximate method may be adequate for this purpose. The rearranged rows and columns with their corresponding residuals are given in Table 5.7.5. This table also gives some other characteristics available for each site with variety and site means.

This may be noted here that the total of each row and column is zero. The varieties and sites which behave contrastingly (with respect to their means) lie near to two opposite extremes of the table. For example, Imber does worse than its mean at the sites near the top of the table whereas Maris Mink does better than its mean near the top of the table. This performance is reversed near the bottom of the table. Thus, the largest contribution to the interaction comes from the interaction of the varieties (near to extreme left and right of the table) with the sites (near the top and bottom of the table). Most of the sites at the bottom of the table belong to the West of Scotland. Looking in the Table 2.5.1 it is found that Imber does worse than Maris Mink in the North and East whereas Imber is better than Maris Mink in the West. Similar conclusions can be drawn about the varieties and sites which lie close to the extremes. Residuals corresponding to the varieties and sites lying near the centre are small and interaction between them contribute little to the total interaction sum of squares.

With the help of this table it may be easier to examine for any possible relationship of interaction with the characteristic of sites and varieties. We have already seen that most of the sites in the top and middle of the table belong to the North and East, and most of the sites near the bottom belong to the West, thus showing the contrasting behaviour of sites of North and East to that of West. There is no variation in soil type among the sites and thus it cannot be examined for interactions. There does not appear to be any ordering among the other site variates like drainage, farm type, sunshine and altitude except that of rainfall. It appears that sites of West have higher rainfall than North and East and most of them lie in the bottom of the table. We have already seen from the analysis of these data in Chapter 3 that the varieties Universe and Maris

Mink have significant negative regression whereas Imber and Pegasus have significant positive regression on rainfall and this may partly explain the contrasting behaviour of North, East and West, particularly, for these four varieties.

In cases when the residuals are proportional to the site and variety effects (as considered by Tukey, 1949 by a multiplicative term) the ordering of varieties and sites should be according to their means. In the present data there does not appear to be any evidence of this type of relationship. A more general type of relationship, as described in Chapter 2, may not be obvious by just looking at this table.

When rows and columns are arranged in such a way then the residuals in contiguous cells will be homogeneous (on the average). In some cases certain cells may have very different residuals from the neighbouring cells. When such cases are found the yield of the corresponding variety at that site should be closely examined for any possible discrepancy. For example, the residuals corresponding to variety Goldfield at sites 6, 16 and 20 are very different than the neighbouring residuals. Now examining the behaviour of the variety Goldfield at sites 6, 16 and 20 it appears that for the first two sites all the varieties, except Goldfield, are affected by lodging. In these two sites Goldfield outyields the other varieties but in general it is not very high yielding variety. This suggests that Goldfield is a lodging resistance variety and may do better than others in areas where lodging is a common occurrence. However, no such reason can be assigned for its better performance at site 20. As there is no incidence of lodging on the other sites the above remarks about Goldfield cannot be conclusive, and further observations should be made before such conclusions are confirmed.

The above method of ordering and arranging residuals may be of great help in recognising any pattern among residuals and their possible relationship with the site and variety characteristics, particularly, when the dimension of the matrix is large. This may reduce the amount of work involved in detailed analysis, of the type considered in Chapter 3, by suggesting to consider only those external variables which appear to have some relationship with the interactions such as rainfall in this case.

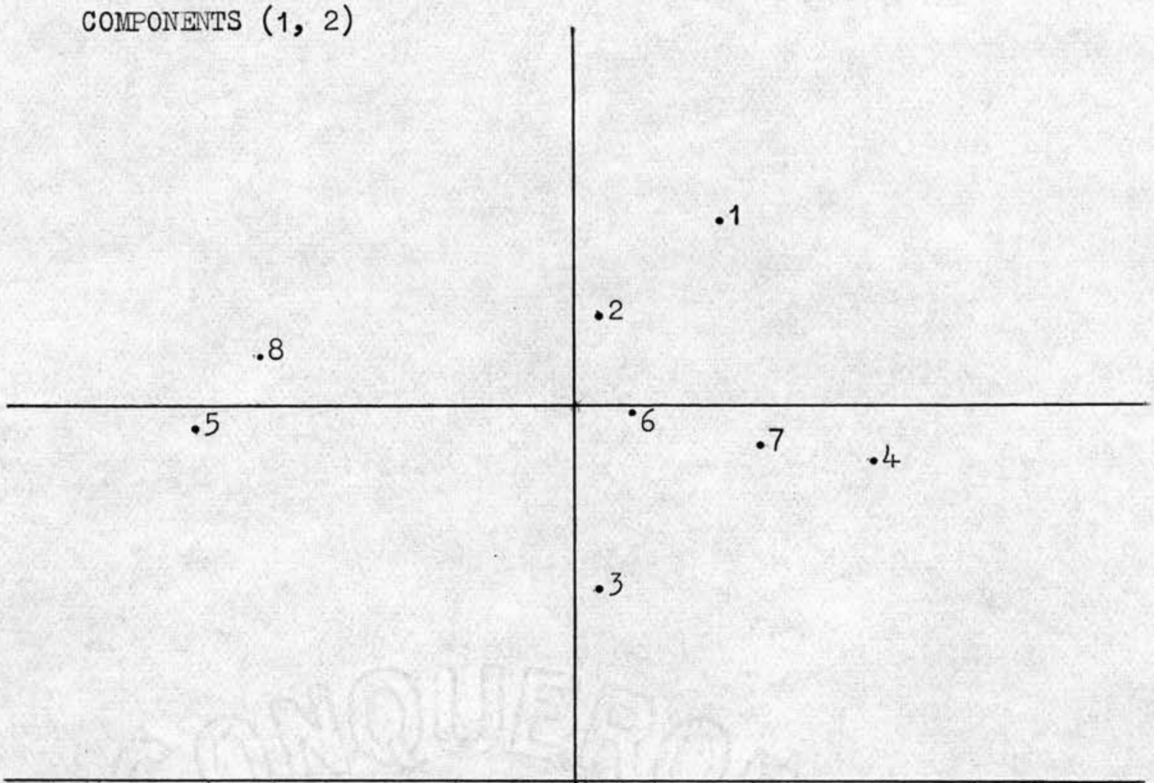
TABLE 5.7.5

RESIDUAL OF VARIETY × SITE TABLE 2.5.1

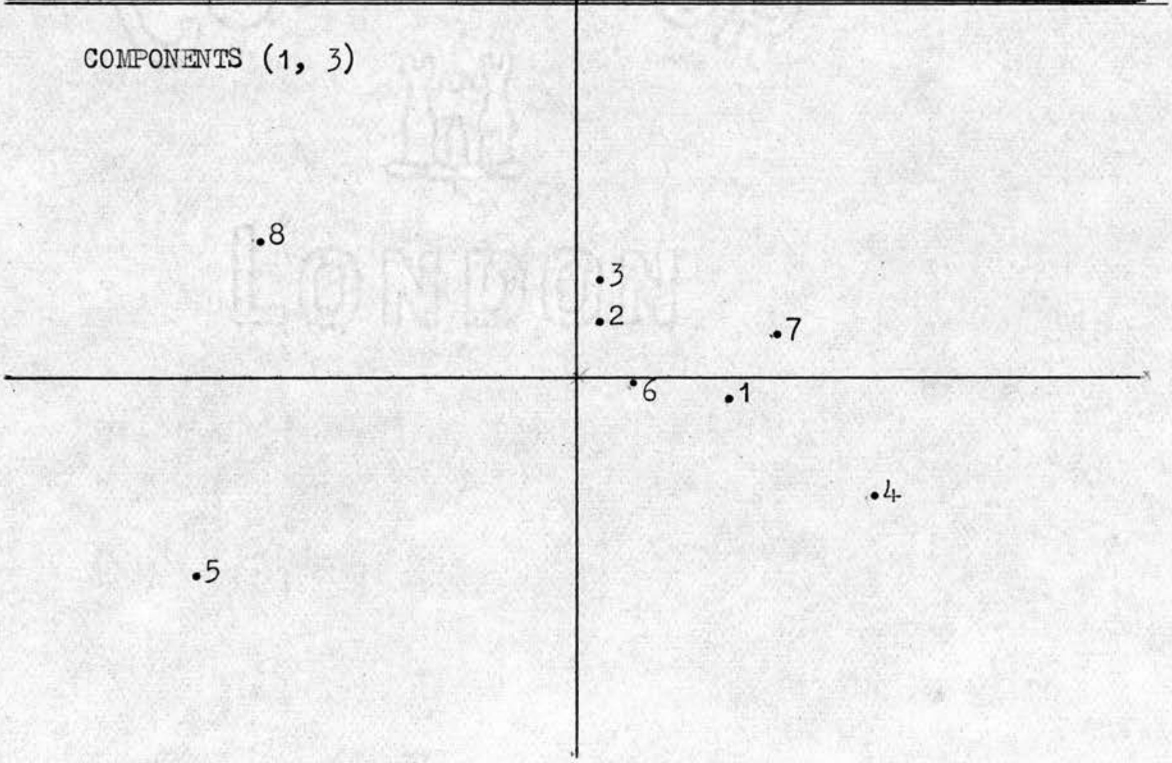
Drainage	Farm	Sun hrs	Rain mm.	Alt. m.	College	Site No.	VARIETIES								Mean Yield
							Imber 4	Pegasus 7	Ymer 1	Mazurka 6	Gerkra 2	Gold-field 3	Universe 8	Maris Mink 5	
Free	Crops			61	E	5	-.54	.01	-.32	.03	-.24	-.11	.61	.57	6.12
Imp.	Crops			97	E	2	-.29	-.19	-.61	-.02	-.05	.05	.72	.39	7.19
Free	Crops			58	E	3	-.24	-.16	-.37	-.47	.11	.23	.21	.69	4.11
Imp.	Ley	959	234	21	N	12	-.42	-.09	-.20	-.06	-.11	.15	.09	.64	5.13
Free	Crops	740	212	90	N	8	-.30	-.41	.18	-.12	.00	.01	.32	.31	5.91
Free	Crops	893	309	27	N	9	-.07	-.47	-.00	.03	-.09	.11	-.15	.65	6.38
Imp.	Crops	973	212	76	N	13	-.24	-.15	.04	.08	-.09	-.22	.21	.37	5.70
Imp.	Crops			61	E	1	-.24	-.27	.05	-.12	.23	-.16	.41	.09	6.95
Free	Crops			46	E	4	-.28	-.06	.19	-.20	.06	-.27	.36	.21	5.43
Free	Crops	901	205	30	N	7	-.22	-.28	.19	.03	.13	-.33	.33	.15	6.63
Free	Ley		492	180	E	6	.22	.16	-.63	.21	-.42	.55	.08	-.17	5.89
Free	Ley	955	426	30	W	16	.27	-.13	-.17	-.01	-.13	.35	-.01	-.16	7.03
Imp.	Crops	939	481	180	W	20	-.38	.30	-.21	.08	.33	.52	-.09	-.54	2.79
Free	Crops	955	426	15	W	15	.29	.00	.18	-.13	.03	-.05	-.18	-.15	5.78
Imp.	Ley	887	257	61	N	10	-.20	.15	.30	.26	.22	-.38	-.07	-.29	5.36
Free	Crops	846	233	137	N	11	.24	.04	.44	-.07	.10	-.33	-.07	-.35	5.90
Free	Ley	849	343	34	W	19	.46	.40	.09	.47	-.30	-.37	-.31	-.45	4.13
Free	Crops	834	502	45	W	18	.75	.18	-.00	.11	.11	-.20	-.95	.00	4.71
Free	Ley	976	528	40	W	14	.40	.41	.44	-.08	.28	.34	-.39	-4.39	5.96
Free	Grass	923	477	105	W	17	.80	.57	.42	-.04	-.18	.11	-1.10	-.57	4.24
				Mean Yield of Varieties			5.44	5.49	5.44	5.72	5.49	5.57	5.82	5.57	5.57

GRAPH 5.7.1 INTERACTION DISTANCE BETWEEN VARIETIES

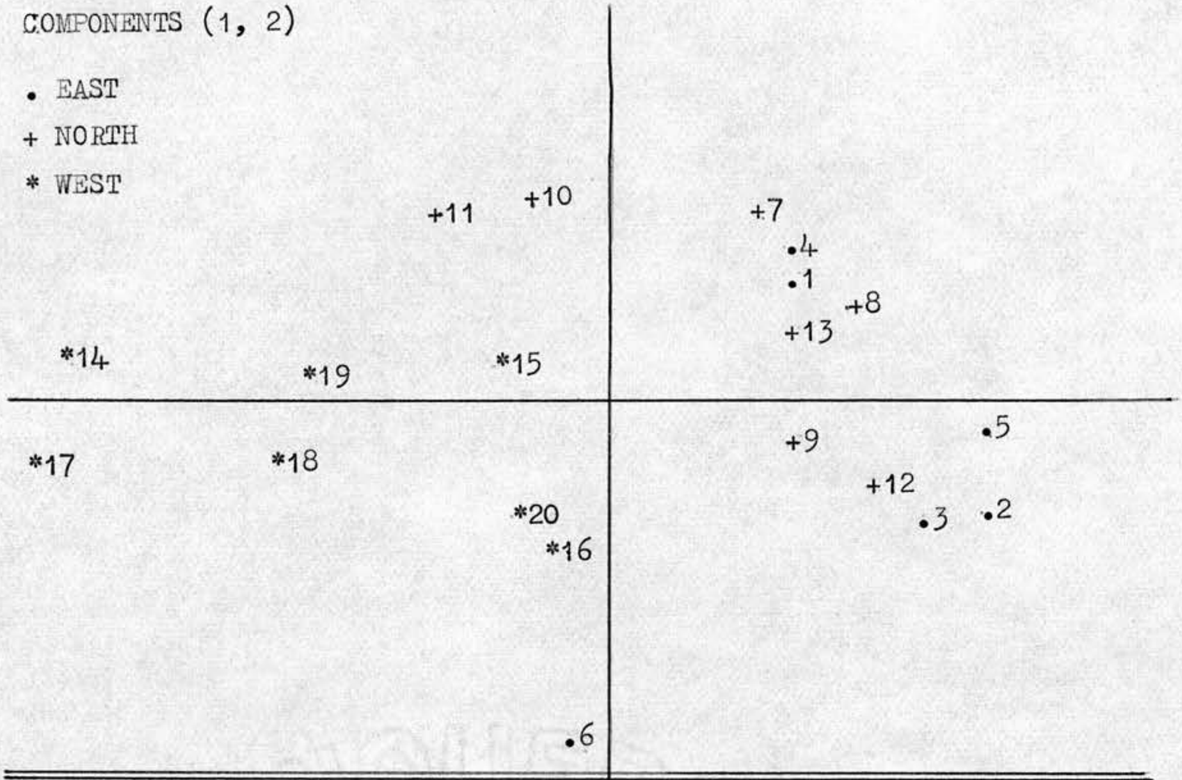
COMPONENTS (1, 2)

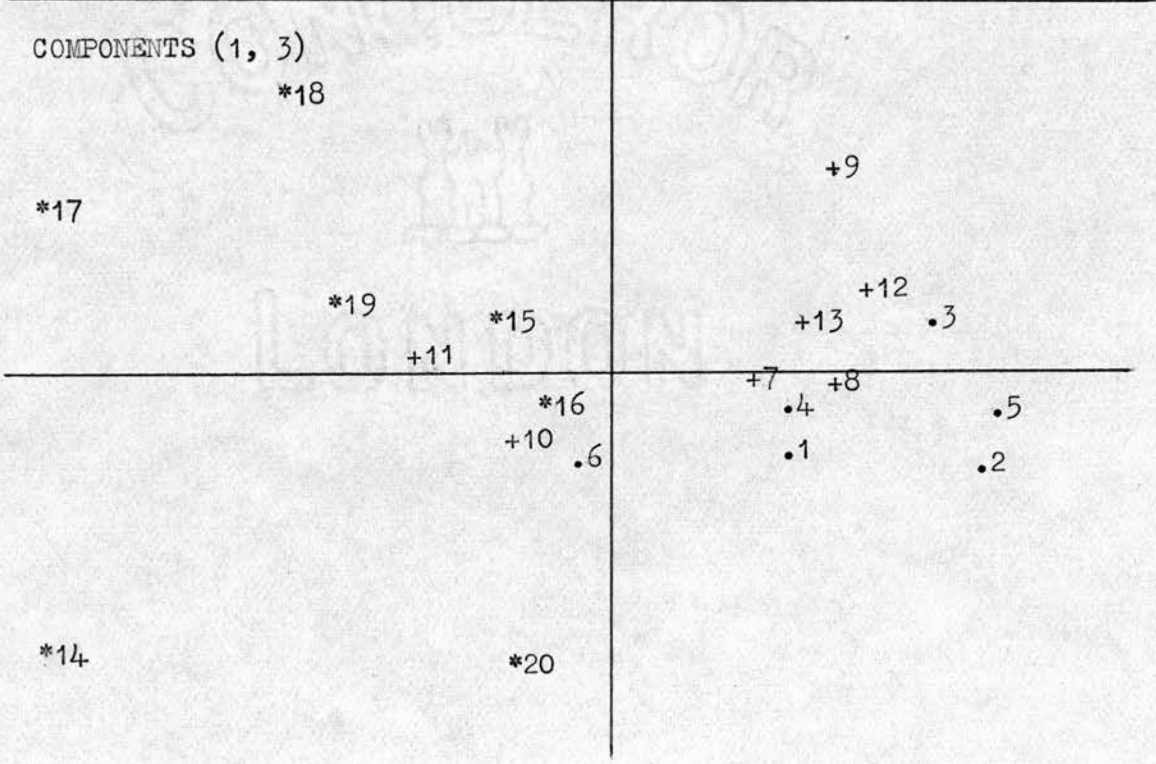


COMPONENTS (1, 3)

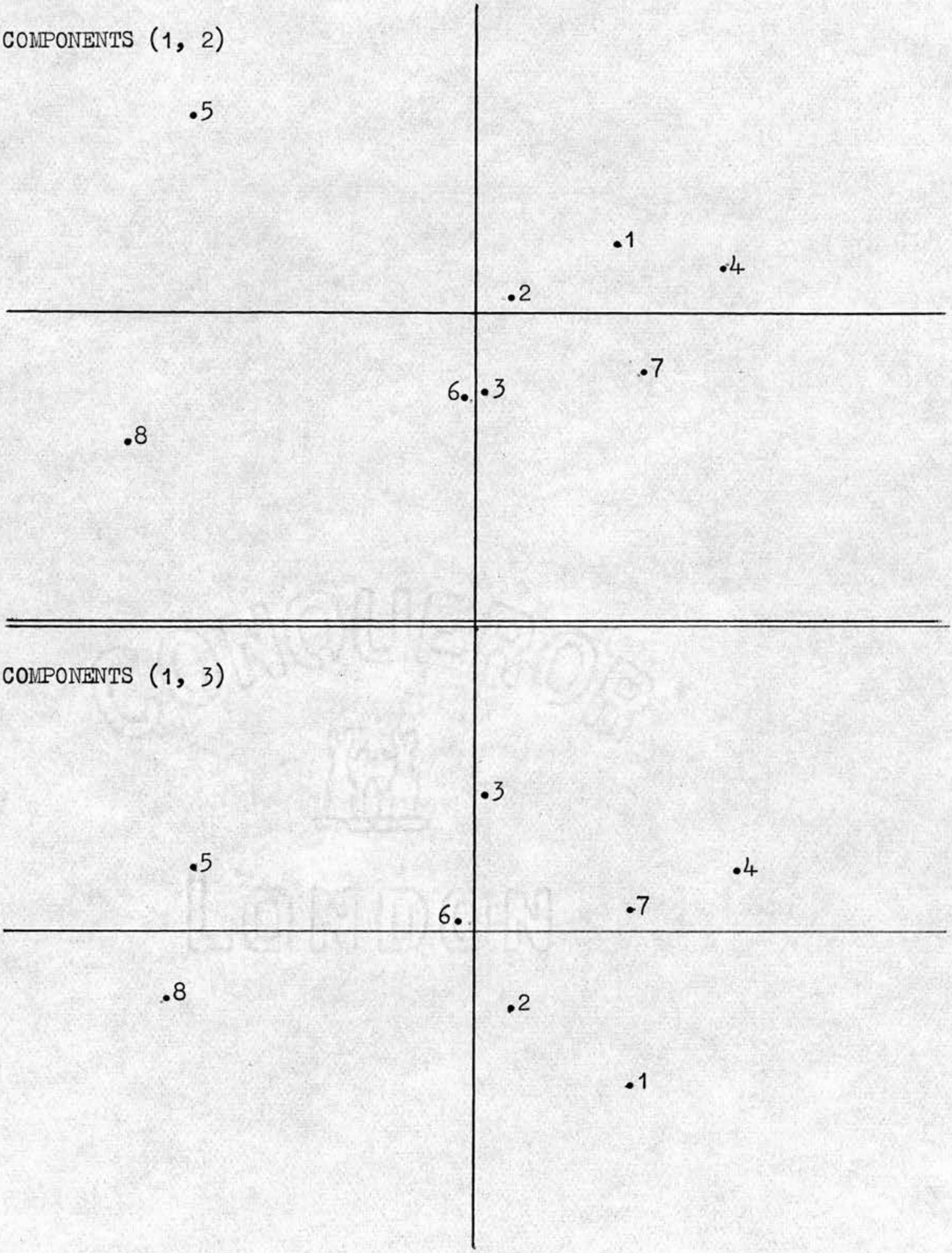


GRAPH 5.7.2 INTERACTION DISTANCE BETWEEN SITES





GRAPH 5.7.3 DISTANCE BETWEEN VARIETIES



CHAPTER 6

OPTIMUM ALLOCATION OF RESOURCES IN PLANNING OF EXPERIMENTS

6.1 Introduction

In the preceding chapters we have discussed the analysis and interpretation of interaction between varieties and environments. There the word environment referred to sites within years, but, if data from several years are available we can also use the word to refer to years. However, the inclusion of climatic factors will only be of limited value because, in general, they cannot be predicted in advance. Sometimes it is possible to characterize years by some climatic factors such as rainfall, sunshine, disease etc. Now the differential responses of the varieties to these environmental factors can be exploited usefully by recommending varieties according to their suitability to respective environmental factors.

When variety-year interactions are present it may be of interest to find a variety which is likely to do better than others, on the average over several years. For this purpose, as mentioned earlier, it may be necessary to test varieties within years (over many sites) and over years. The usual practice is to compare each variety with the standard for three to four years; if some are found better than the standard, they are recommended and others are rejected. It seems reasonable that if a variety has performed very badly in comparison with the standard, it may be rejected before the completion of a fixed period. Similarly if a variety is found much superior to the standard in the first year or two, there is a temptation to recommend it before the completion of a fixed period. There is always an risk of a superior variety being rejected and inferior variety being recommended. This type of risk is large when we take trials for a small number of years and the year variety interaction is large. As the trials are conducted at yearly intervals, we get an opportunity of deciding whether or not to continue a variety in future trials, on the basis of all the information available up to that time.

In this chapter we shall consider a two-stage procedure for deciding the optimum amount of experimentation so as to maximise the expected net gain, which will be defined precisely later. We shall assume that some information about the difference between variety yields, under consideration, is available from the first one or two years of experimentation. Three sources of variation will be considered i.e. (1) within site experimental error with variance σ_ξ^2, (2) variety-site interaction with variance σ_{vs}^2 and (3) variety-year interaction with variance σ_{vy}^2. We shall assume that good estimates of σ_ξ^2, σ_{vs}^2 and σ_{vy}^2 are available in advance. We shall try to obtain the optimum amount of (1) within site replication r, (2) number of sites within year n, and (3) the number of years of further experimentation t; for choosing one of the two varieties for recommendation. For this purpose a linear cost function will be considered. We shall consider some variation of the model and also give some approximate results for a completely sequential approach.

6.2 Development of previous work

A first attempt towards determining the optimum allocation of experimental resources, when the results are applied on a given area, was made by Yates (1952). Yates's study was followed up by Grundy et al (1956). The latter considered a two-stage procedure. Some information on the treatment differences is available from the first-stage and a question is asked whether to (i) accept (or reject) one of the treatments immediately on the basis of existing information from the first stage, or (ii) to conduct some more experiments before taking any final decision. Grundy et al (1956) also considered the optimum amount of further experimentation required. Raiffa and Schlaifer (1961) also discuss the same problem and obtain the same results by an alternative approach.

In this chapter we have used the procedure described in Grundy et al (1956) and extended it to the case of three sources of variations as mentioned above. We have considered variations of the model which may be more appropriate in the case of variety trials and also considered a completely sequential approach.

6.3 Formulation of the problem

Let us denote by θ, the difference between the yields per unit area of two varieties between which choice is to be made. If θ is established to be positive one variety is preferred and if θ is established to be negative other variety is preferred. The variety with the higher yield is chosen for recommendation and the other is rejected. However, in practice, yield is not the only criterion for preferring a variety over the other and many other considerations such as quality, disease resistance characteristics are also very important. For simplicity we shall assume that the scale on which θ is measured takes account of such factors.

6.3.1 Variance of difference between variety means

Let us assume that the yield of the variety j, at the site i, in the year k, for the replication l is given by Y_{ijkl} and can be represented as

$$Y_{ijkl} = \mu + v_j + y_k + s_{i(k)} + (vs)_{ijk} + (vy)_{jk} + \xi_{ijkl} ; \qquad (6.3.1)$$

$$i = 1, 2, \ldots, n; \quad j = 1, 2; \quad k = 1, 2, \ldots, t; \quad l = 1, 2, \ldots, r;$$

where μ and v_j are as defined earlier; y_k is the additive effect of the year k; $s_{i(k)}$ is the additive effect of the site i in the year k; $(vs)_{ijk}$ is the interaction component for the site i (of the year k) and the variety j; ξ_{ijkl} is the within site components of error associated with Y_{ijkl}. Here we have assumed that the sites are selected randomly every year and Y_{ijkl} is normally distributed. The variance components are denoted by

$$V\left[(vs)_{ijk}\right] = \sigma_{vs}^2 ; \quad V\left[(vy)_{jk}\right] = \sigma_{vy}^2 ; \quad V(\xi_{ijkl}) = \sigma_\xi^2 .$$

Then,

$$E(\bar{Y}_{.1..} - \bar{Y}_{.2..}) = v_1 - v_2 ;$$

$$V(\bar{Y}_{.1..} - \bar{Y}_{.2..}) = \frac{2}{t}\left(\sigma_{vy}^2 + \frac{\sigma_{vs}^2}{n} + \frac{\sigma_{\xi}^2}{nr}\right) . \qquad (6.3.2)$$

6.3.2 Cost of experimentation

Let C denote the total cost of experimentation every year as measured in a monetary unit. Let us assume that it can be expressed as a sum of three components

$$C = C_o + nC_1^{*} + nrC_2 ; \qquad (6.3.3)$$

where C_o is the fixed cost of extending the experiment for one more year; C_2 is the cost of taking an extra replication within each site. For keeping discussions simple and for practical convenience we shall assume that the number of sites (n) and the number of replications (r) remain constant throughout the trial period. In the following we shall consider two cases.

6.3.2.1 Minimum variance within year for a fixed cost

Let us assume that the total cost of each year's experimentation is fixed and is given by C_F. For a given C_F best choice of n and r is given by minimising the variance of the mean difference within each year subject to the condition that

$$C_F = C_o + nC_1 + nrC_2 .$$

* C_1 is the cost of taking an extra site.

This can be done easily by using the method of Lagrange multipliers. For this we minimise

$$F = 2\left(\frac{\sigma_{vs}^2}{n} + \frac{\sigma_\varepsilon^2}{nr}\right) + \alpha(C_o + nC_1 + nrC_2 - C_F) \ ;$$

where α is a Lagrange multiplier. This immediately gives

$$r_F = \sqrt{\frac{C_1\,\sigma_\varepsilon^2}{C_2\,\sigma_{vs}^2}} \ ; \quad n_F = \frac{C_F - C_o}{C_1 + r_F C_2} \ . \qquad (6.3.4)$$

6.3.2.2 Minimum variance over years

Let us assume that the total cost of experimentation over t years is fixed and is given by C_S. For a given C_S the best choice of n and r is given by minimising the mean difference over t years subject to the condition that

$$C_S = t(C_o + nC_1 + nrC_2) \ .$$

This can be done easily by using the method of Lagrange multipliers. For this minimise

$$F = \frac{2}{t}\left[\sigma_{vy}^2 + \frac{1}{n}\left(\sigma_{vs}^2 + \frac{\sigma_\varepsilon^2}{r}\right)\right] + \alpha(tC_o + ntC_1 + ntrC_2 - C_S) \ ;$$

where α is a Lagrange multiplier. This immediately gives

$$r_S = \sqrt{\frac{C_1\sigma_\varepsilon^2}{C_2\sigma_{vs}^2}} \ ; \quad n_S = \sqrt{\frac{C_o(\sigma_{vs}^2 + \sigma_\varepsilon^2/r_S)}{(C_1 + r_S C_2)\sigma_{vy}^2}} \ ; \quad t_S = \frac{C_S}{(C_o + n_S C_1 + n_S r_S C_2)} \ . \qquad (6.3.6)$$

It is apparent from the expressions of r_S and n_S that they are not affected in any way by the variations in C_S. For fixed r_F (or r_S) and C_F, n is fixed for the case considered in Section 6.3.2.1. The values of r and n as given in (6.3.6) are also obtained by an alternative approach and will be discussed briefly in Section 6.4.

For keeping discussions general we shall use C for the cost of experimentation per year and it should be replaced by C_F or C_S corresponding to two cases considered here.

6.3.3 Definition of gain function

Let us assume that the true difference between the first and the second variety under consideration is θ ($\theta = v_1 - v_2$). Suppose K is the price of additional produce to be obtained for a unit increase in θ, from the whole area (over years) on which this variety is to be grown. Suppose we have an estimate d_1 of θ available from the first stage of the trial based on t_1 years, then

$$E(d_1) = \theta \; ; \; V(d_1) = \frac{1}{ht_1} \; ; \qquad (6.3.7)$$

where

$$\frac{1}{h} = 2 \left[\sigma_{vy}^2 + \frac{1}{n}\left(\sigma_{vs}^2 + \frac{\sigma_\varepsilon^2}{r}\right)\right] .$$

Suppose we conduct the trial for a further t_2 years; on n sites within each year; and on r replications within each site. Let the mean difference (mean over all rnt_2 observations) between two varieties is denoted by d_2. Then d_2 provides another independent estimate of θ such that

$$d_2 = \bar{Y}_{.1..} - \bar{Y}_{.2..} = \frac{1}{nrt_2} \left[\sum_{i,k,l} (Y_{i1kl} - Y_{i2kl}) \right] ;$$

$$E(d_2) = \theta ;$$

$$V(d_2) = \frac{1}{ht_2} .$$

The combined estimate of θ is given by

$$\hat{\theta} = \frac{t_1 d_1 + t_2 d_2}{t_1 + t_2} . \qquad (6.3.8)$$

Now the decision for choosing one variety or the other will be based on the sign of $\hat{\theta}$. Let us assume that for a given value of d_1, t_1 and θ the probability of taking a correct decision is P i.e. $\hat{\theta} \geq 0$ when $\theta \geq 0$ (or $1-P$ if $\hat{\theta} < 0$ when $\theta < 0$).

By taking a correct decision we expect to gain (on the average) a quantity,

$$\begin{aligned} \text{Gain} &= K.P.\,\theta \qquad (\theta \geq 0) \\ &= -K.(1-P).\,\theta \qquad (\theta < 0) . \end{aligned} \qquad (6.3.9)$$

The probability P, that $\hat{\theta}$ is positive i.e. $d_2 \geq \frac{-t_1 d_1}{t_2}$, by using (6.3.7), is given by

$$P = \Phi[(t_1 d_1 + t_2 \theta)(h/t_2)^{\frac{1}{2}}] \qquad (6.3.10)$$

where

$$\Phi(X) = \int_{-\infty}^{X} \phi\,(u)du$$

and

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) .$$

As the expression in (6.3.9) contains θ which is unknown we shall integrate it over its fiducial distribution based on d_1 and call it integral gain (I.G.);

$$\begin{aligned} \text{I.G.} &= K(ht_1)^{\frac{1}{2}} \left[\int_0^{\infty} P\theta - \int_{-\infty}^{0} (1-P)\theta\right] \phi\left[(\theta - d_1)(ht_1)^{\frac{1}{2}}\right] d\theta \\ &= \frac{K}{\sqrt{ht_0}} L(D^*) + K|d_1| + \frac{K}{\sqrt{ht_1}} L(d_1\sqrt{ht_1}) \; ; \end{aligned} \qquad (6.3.11)$$

where

$$L(u) = \phi(u) - u\left[1 - \Phi(u)\right]$$

$$D^* = |d_1|(ht_o)^{\frac{1}{2}}$$

$$\frac{1}{t_o} = \frac{1}{t_1} - \frac{1}{t_1 + t_2} \quad .$$

Now the cost of conducting trials for another t_2 years is given by t_2C. The last two expressions in (6.3.11) are independent of t_2 and it is more convenient to work with 'net gain' which is obtained by subtracting (i) the cost of extra experimentation and (ii) the gain by taking immediate decision $(t = 0)$ based on d_1, from the expression of I.G. in (6.3.11).

$$\begin{aligned} N.G. &= I.G. - t_2C - K|d_1| - \frac{K}{\sqrt{ht_1}}\, L(d_1\sqrt{ht_1}) \\ &= \frac{K}{\sqrt{ht_o}}\, L(D^*) - t_2C \quad . \end{aligned} \qquad (6.3.12)$$

Now define the ratios

$$\rho = t_2/t_1$$

$$D' = |d_1|\ (ht_1)^{\frac{1}{2}}$$

$$\lambda = K(t_1{}^3h)^{-\frac{1}{2}}/C$$

$$\beta = \left(\frac{\rho}{\rho+1}\right)^{\frac{1}{2}} \quad ;$$

and denote by G the ratio

$$G = \frac{N.G.}{t_1 C} = \lambda \, \beta \, L(D'/\beta) - \rho \quad ; \qquad (6.3.13)$$

where G expresses the net gain in the dimensionless form. In the next section we shall obtain optimum t_2 by maximising G.

6.4 Results

The expression of G is the same as obtained by Grundy et al (1956) and Raiffa and Schlaifer (1961; p.115). In the following we shall use their results. Let us denote G of (6.3.13) by $G(\rho, D', \lambda)$ as it is a function of these three parameters. Using the conventional method of finding a maximum by differentiating G with respect to ρ we obtain equation

$$\frac{dG}{d\rho} = \frac{1}{2} \lambda \, \rho^{-1/2}(1 + \rho)^{-3/2} \, \phi(D'/\beta) - 1 = 0 . \qquad (6.4.1)$$

Suppose ρ^o is the root of the equation (6.4.1) which corresponds to a local maximum of $G(\rho, D', \lambda)$ if one exists: then the following properties of ρ^o are obtained:

If $D' = 0$, a local maximum always exists and ρ^o is the unique root of (6.4.1).

If $D' > 0$, a local maximum may or may not exist. If one does exist, the maximiser ρ^o is the larger root of (6.4.1) but $G(\rho^o, D', \lambda)$ may or may not be greater than $G(0, D', \lambda) = 0$. A zero value of t_2 corresponds to a case when no further experimentation is required. Thus for using the results one has to obtain ρ^0 by solving equation (6.4.1) and making sure that $G(\rho^0, D', \lambda) > 0$ by substituting in (6.3.13). The following table gives the optimum value of t_2/t_1 for different values of D' and λ as read from the Chart 1 given in Raiffa and Schlaifer (1961).

TABLE 6.4.1

OPTIMUM NUMBER OF YEARS OF FURTHER EXPERIMENTATION RELATIVE TO THE NUMBER OF YEARS OF THE FIRST-STAGE EXPERIMENTS (t_2/t_1)

D' $\lambda^{1/3}$	0.0	0.2	0.4	0.6	0.8	1.0	1.2
1.0	0.036	-	-	-	-	-	-
1.5	0.236	0.202	-	-	-	-	-
2.0	0.620	0.580	0.480	-	-	-	-
2.5	1.078	1.050	0.938	0.781	-	-	-
3.0	1.620	1.575	1.476	1.287	0.990	-	-
3.5	2.205	2.168	2.058	1.838	1.531	1.164	-
4.0	2.848	2.800	2.672	2.400	2.112	1.680	-
4.5	3.524	3.443	3.341	3.038	2.693	2.228	1.721
5.0	4.250	4.200	4.000	3.688	3.325	3.050	2.250

For given values of K, d_1, C and t_1 one can calculate λ and D' and then obtain the optimum value of t_2 from the above table. If optimum t_2 is near zero then no further trials are needed and the decision on the choice of variety should be taken on the sign of d_1. The value of G (6.3.13) should be calculated and if it is positive corresponding to $\rho = \rho^o$ only then trials for further t_2 years should be carried out and then the decision should be based on the sign of $\hat{\theta}$.

The optimum value of r remains the same for both the cases considered in Section 6.3.2. The optimum value of t_2 obtained here is such that this maximises the net gain for a fixed precision of the first-stage information $(1/ht_1)$. As the precision of the first-stage information decreases the optimum value of t_2 increases and so is the gain by taking extra trials.

r_S and n_S obtained in Section 6.3.2.2 are such that they maximise the N.G. (6.3.12) for the second-stage of experimentation. This can be easily verified by maximising (6.3.12) with respect to r, n and t_2 in such a way that $V(d_1)$ remains constant.

6.5 Some further extensions of the model

The gain function considered in Section 6.3.3 does not take into account the loss (or gain) incurred during the years of experimentation. This can be easily explained as follows. Suppose the first variety is a new variety and the second is the standard variety which is already in use. If the new variety is better than the standard and this may be suggested by the first-stage of experimentation then we may be losing an additional amount for not recommending it just after the first-stage. For this purpose we shall assume that the price of the additional produce per year for a unit increase in θ is K_1. For simplicity we shall assume that K_1 remains constant over years and the recommended variety is grown for T years. Then

$$K = K_1 . T \quad .$$

The expressions for gain can now be written as

$$\begin{aligned} \text{Gain} &= K_1 T P \theta - K_1 t_2 \theta \quad (\theta \geq 0) \\ &= -K_1 T(1-P)\theta - K_1 t_2 \theta . \quad (\theta < 0) \quad . \end{aligned} \tag{6.5.1}$$

Integrating this with respect to θ and subtracting the cost of additional experimentation, as in Section 6.3, and taking the deviations from gain corresponding to $t_2 = 0$ obtain

$$G = \frac{N.G.}{t_1 C} = \lambda \beta L(D'/\beta) - \frac{\rho K_1 d_1}{C} - \rho \quad . \tag{6.5.2}$$

For finding the optimum value of t_2 we differentiate this with respect to ρ as before and obtain the local maximum. The equation for the optimum solution is given by

$$\frac{1}{2} \lambda \rho^{-1/2}(1+\rho)^{-3/2} \phi(D'/\beta) - \frac{K_1 d_1}{C} = 1 \quad . \tag{6.5.3}$$

This can be re-written in the form of (6.4.1) as

$$\frac{1}{2} \lambda_1 \rho^{-1/2}(1+\rho)^{-3/2} \phi(D'/\beta) = 1 \tag{6.5.4}$$

where

$$\lambda_1 = \frac{K(t_1^3 h)^{-1/2}}{(C + K_1 d_1)} \quad .$$

Again no explicit solution of equation (6.5.4) exists but the solution can be obtained by using Table 6.4.1 and replacing λ by λ_1 as given above. The effect of considering this model is to increase the cost of experimentation from C to $C + K_1 d_1$ and this reduces the optimum amount of further experimentation.

6.6 Approximate results for a completely sequential approach

As the trials are conducted at yearly intervals and there is a time lag of one year between the trials, it will be more desirable to adopt a completely sequential approach of stopping trials and taking decision. The trials should continue for more years so long as they are expected to increase the net gain. No further trials should be conducted if they are expected to decrease the net gain.

For this we shall use the results of Sections 6.3 and 6.4 for finding the approximate boundary for stopping further experiments. Exact results for a completely sequential approach are very difficult to obtain and the approximate results given here should be adequate for the present purposes.

We shall consider the gain function discussed in Section 6.3.3. We have seen in Section 6.4 that the decision for taking trials for more years is based on t_1 and d_1 for given K, C and h. We shall assume that h remains constant throughout the trial period. Now for given K, C, t_1 and d_1 it is possible to find from the expressions of G in (6.3.13) whether the trials for additional years are expected to increase or decrease the N.G. As discussed before the function G is such that for $D' = 0$, G increases monotonically with ρ and then decreases monotonically after attaining a maximum. This is because the previous information does not prefer one variety to the other and it is always advantageous to take some more trials. For $D' \neq 0$ as ρ increases from zero G decreases in the beginning and then increases and goes to a maximum and then starts decreasing again. If for given t_1 and d_1, $G(o, D', \lambda) \geq G(\rho^o, D', \lambda)$ then no further trials should be conducted and the decision should be based on the sign of d_1. Here ρ^o represents the value of ρ corresponding to maximum G. The function G is such that for a given t_1 if no trials are needed for $D' = D_1$ then no trials are needed for any $D' = D_2$ if $D_2 > D_1$.

For obtaining the boundary in terms of D' for each t and λ we have to calculate the value of D' for which $G(\rho^o, D' \lambda) = 0$. Here ρ^o is the optimum value of ρ. The optimum ρ for each D' and λ is given by equation (6.4.1) which can be re-written as

$$\frac{2\beta}{(1 - \beta^2)^2 \phi(D'/\beta)} = \lambda \qquad (6.6.1)$$

At the boundary net gain is zero and thus equating $G(\rho^o, D', \lambda) = 0$ we obtain

$$\frac{\beta}{(1 - \beta^2)\, L\,(D'/\beta)} = \lambda \quad . \qquad (6.6.2)$$

The boundary values of D' for each λ can be obtained from these two equations. For solving them it is more convenient to put them in the form

$$\sqrt{1 - \frac{2\, L\,(D'/\beta)}{\phi\,(D'/\beta)}} = 2\,\lambda\,\frac{L^2\,(D'/\beta)}{\phi\,(D'/\beta)} \qquad (6.6.3)$$

because

$$\beta = \sqrt{1 - \frac{2\, L\,(D'/\beta)}{\phi\,(D'/\beta)}} \quad . \qquad (6.6.4)$$

Now for each λ obtain D'/β from (6.6.3) and then obtain β from (6.6.4) and thus D' .

It is more convenient to present the results in the terms of $K/C\sqrt{h}$ and D'_t for each t , where

$$D'_t = \frac{\left|\sum_{i=1}^{t} d_i\right|}{t}\sqrt{ht} \quad .$$

Here d_i represents the mean difference between the two varieties in the ith year. Table 6.6.1 gives the result for some chosen values of $K/C\sqrt{h}$.

TABLE 6.6.1

APPROXIMATE BOUNDARIES (D'_t) FOR SEQUENTIAL PROCEDURE

t	$K/C\sqrt{h}$					
	50	100	500	1,000	5,000	10,000
1	1.10	1.30	1.80	2.00	2.50	2.65
2	0.80	1.00	1.50	1.70	2.20	2.40
3	0.60	0.80	1.30	1.50	2.00	2.20
4	0.40	0.70	1.20	1.40	1.90	2.10
5	0.35	0.60	1.10	1.30	1.80	2.00
6	0.30	0.50	1.00	1.20	1.70	1.90
7	0.25	0.40	0.90	1.10	1.60	1.80
8	0.20	0.30	0.80	1.00	1.50	1.70

Further trials should continue so long as D'_t is smaller than the values tabulated above. Experiment should stop as soon as the above boundary is crossed for the first time. The decision should be based on the sign of Σd_i. The optimum boundaries are shown in the Graph 6.6.1.

Very often these boundaries will not be crossed within the maximum time period allowed for conducting trials. In this case it is still advisable to take the decision on the sign of Σd_i but the calculations of N.G. for the optimum number of further trial may give a good guidance about the expected loss involved in not conducting any more trials.

6.7 Discussion of the results

The main difficulty in applying the approach discussed in this chapter is this that K is never known in advance. From the past experience it may be possible to guess the approximate value of K. It is apparent from the Table 6.6.1 that when K is large the boundary is not too sensitive for slight variations in K. However, it is hoped that this approach will provide more satisfactory procedure than the testing of hypothesis approach which assigns probability levels quite arbitrarily. Results of Table 6.6.1 can be used to calculate the appropriate probability levels (strictly speaking they are not valid in this case) and this suggests that when K is small the corresponding level of test of significance is large and when K is large the level of significance is small. This procedure gives an approximate idea of choosing the appropriate significance level after taking into account the economic importance of the decision.

As far as the choise of n and r is concerned we have considered two methods. The optimum values of r obtained by both the methods remains the same and therefore it should be chosen. For practical convenience it may not be possible to continue trials indefinitely and there may be a limit over which it may not be desirable to go. If we allocate the cost equally over the maximum number of years fixed in advance then C_F can be obtained and so the optimum value of n. The results of Sections 6.3 - 6.6 can be used for finding the optimum number of further trials and the sequential boundary.

As far as the choice between the two gain functions (discussed in Sections 6.3 and 6.5) is concerned the gain function considered in Section 6.3 is more appropriate when the main interest lies in comparing two new varieties which are not in use. When a new variety is compared with the standard variety of which latter is already in use, and if new variety appears to be superior ($d_1 > 0$) to the standard one then the gain function considered in Section 6.5 appears to be more appropriate.

When the interaction between variety and year is large then the assumption of d_1 being normally distributed when based on a small number of years may not be valid. Small departure from normality may not affect the results very much. When large interactions are present then taking trials for a small number of years may not be adequate as will be suggested by the optimum boundaries, particularly when θ is small. When the interactions are small then the results of Section 6.3.2.2 suggest that large number of sites should be selected within years and it is advisable to take n_S sites and fix the maximum number of years as t_S obtained in Section 6.3.2.2.

6.8 An example

The data used in Chapter 3 (Table 3.3.3) on Wheat variety trials have been used here to give an example using the results of this chapter. The estimates of σ^2_ξ, σ^2_{vs} and σ^2_{vy} are obtained as

$$\hat{\sigma}^2_\xi = 0.128 \ ; \ \hat{\sigma}^2_{vs} = 0.106 \ ; \ \hat{\sigma}^2_{vy} = 0.026 \ .$$

For the present example we shall assume they give good estimates of the corresponding parameters, though, $\hat{\sigma}^2_{vy}$ is based on few degrees of freedom. Assume that the cost is measured in pound sterling and is given by

$$C_o = 500 \; ; \; C_1 = 100 \; ; \; C_2 = 20 \; ; \; K = 300{,}000 \; ;$$

and also assume that the total fund available for conducting trials for a maximum of three years is £6,000. Then

$$C_F = 6{,}000/3 = 2{,}000$$

$$r_F = \sqrt{\frac{C_1 \sigma^2_\varepsilon}{C_2 \sigma^2_{vs}}} \simeq 2$$

$$n_F = \frac{C_F - C_o}{C_1 + r_F C_2} \simeq 10$$

$$\frac{1}{h} = 2[\sigma^2_{vy} + \frac{\sigma^2_{vs}}{10} + \frac{\sigma^2_\varepsilon}{20}] = 0.086$$

Now consider two varieties Cappell and Tommy which give the following yields in three years (from Table 3.3.3)

	1969	1970	1971	D'_1	D'_2	D'_3
Cappell	5.33	4.95	5.01			
Tommy	5.43	5.12	5.23	0.341	0.651	0.965

Taking the first year's information as the information of the first-stage we shall calculate the optimum number of years of further experimentation.

$$\lambda = \frac{300,000}{2,000}\sqrt{0.086} = 43.95 .$$

Looking in the Table 6.4.1 corresponding to $\lambda = 43.95$ and $D' = 0.34$ optimum number of further experimentation are obtained as (approximately) two. We have also calculated N.G. for different values of t using the formula

$$N.G. = \lambda \beta L (D'/\beta) C_F - t C_F$$

t	N.G. (in pound sterling)
0	0
1	11,759
2	13,406
3	13,323
4	12,247

This gain has been calculated after taking deviation from gain at $t = 0$. Gain at $t = 0$ is given by

$$K|d_1| + \frac{K}{\sqrt{h}} L(d_1 \sqrt{h}) = 52,152$$

and this should be added to N.G. obtained above.

For using the sequential procedure we calculate

$$\frac{K}{C\sqrt{h}} = 44.0$$

and looking in the Table 6.6.1 corresponding to $\frac{K}{C/h} = 50$ we find that D'_3 crosses the boundary for the first time and the trial should be stopped after three years and Tommy should be recommended for the further use. It is a coincidence that in this particular case both the approaches give the same result.

If we consider the model of Section 6.5 and taking $T = 10$ we obtain

$$K_1 = \frac{K}{T} = 30,000$$

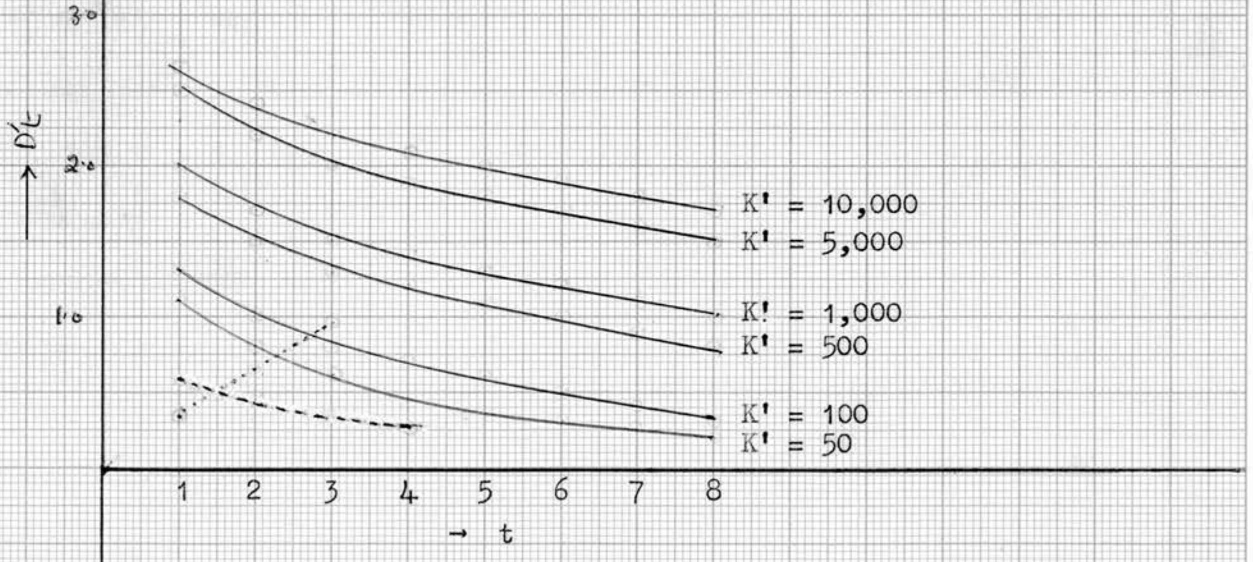
$$\lambda_1 = \frac{K}{(C_F + d_1 K_1)/h} = \frac{300,000}{3.410\ (2,000 + 3,000)} = 17.60 .$$

Again looking in the Table 6.4.1 corresponding to $\lambda = 17.60$ and $D' = 0.341$ optimum number of years of further experimentation is obtained as one.

Sequential path for this example has been given in Graph 6.6.1. Approximate optimum boundary, corresponding to the model considered in Section 6.5, has also been given for $t = 1, 2, 3, 4$ years for this example.

GRAPH 6.6.1

BOUNDARIES FOR SEQUENTIAL PROCEDURE



$K' = K/C \sqrt{h}$

——— Optimum boundary for the gain function in Section 6.3

------ Optimum boundary for the gain function in Section 6.5

•••••• Sequential path for the example in Section 6.6

CHAPTER 7

SEQUENTIAL PROCEDURE FOR COMPARISON OF VARIETIES

7.0 Introduction

To use the procedure suggested in Chapter 6, knowledge of the cost parameters K, C_0, C_1 and C_2 is essential. This information may be difficult or impossible to obtain. Moreover, the method described in Chapter 6 is limited to only two varieties. Sometimes more than one new variety is compared with the standard variety and it may be more relevant to ask if any one (or more) of the new varieties is likely to give a higher yield than the standard. When a comparison is made between the highest yielding variety and the standard, the distribution theory becomes more complicated and simple methods are no longer applicable. Some results obtained by Dunnett (1960) can be used in conjunction with the procedures described in Chapter 6. In the present chapter we shall consider an alternative approach to the following problems:

a) Suppose we have to select the highest yielding variety out of several submitted for test. We are allowed to conduct experiments sequentially. The procedure should be such that the probability that the finally selected variety differs from the true highest yielding variety (which is unknown), by an amount δ, does not exceed $1 - P^*$.

b) In some cases varieties include a standard variety and it is desired to compare the new varieties with the standard. Now it is of interest to select the highest yielding new variety if it is better than the standard.

In this chapter sequential procedures for solving the above two problems are given. Sequential procedures are defined by boundaries for each stage. Boundaries for the problem (a) are such that as soon as the cumulative difference between the highest and the lowest yielding variety crosses the boundary for the first time, the lowest yielding variety is discarded. All the varieties whose differences fall within the boundaries are passed on to the next stage and the same procedure is followed successively until all but one variety is discarded. The region within boundaries is called the continuation region. These boundaries are closed in the sense that the maximum number of stages can be chosen under certain restrictions.

For problem (b) cumulative differences between the standard and the other varieties are plotted and as soon as the differences cross the boundary the experiment is terminated and appropriate action is taken.

The boundaries are chosen to satisfy probability requirements. An important feature of these procedures is that inferior varieties can be discarded at an early stage. These boundaries tell us which varieties should be discarded at each stage. This is of particular interest in variety trials where the rejection of non-contending varieties at early stages leads to a better use of experimental resources so that new varieties can be included as soon as possible. To keep the boundaries simple some assumptions are made about the distribution of variety means.

These assumptions may look rather unrealistic but Monte Carlo studies show that the procedures are robust.

7.1 Notation and review of previous work

For simplicity we shall assume that there are k normal populations $\pi_j (j = 1, 2, \ldots, k)$. Population j has mean θ_j with a common variance σ^2. In this chapter we shall assume that σ^2 is known and using the notation of Chapter 6

$$\sigma = \sqrt{\sigma_{vy}^2 + \frac{\sigma_{vs}^2}{n} + \frac{\sigma_{\xi}^2}{nr}} \quad . \tag{7.1.1}$$

Let us assume that at least one population has a higher mean than the others and we have to select the population with the highest mean. For simplicity we shall assume that the first $k-1$ populations have mean θ_0 (unknown) and the mean of the kth population is $\theta_0 + \delta$ $(\delta > 0)$.

$$\theta_1 = \theta_2 = \ldots = \theta_{k-1} = \theta_k - \delta = \theta_0 \quad . \tag{7.1.2}$$

Now the problem is to find out which is the highest yielding variety (kth in the above notation). For the procedure described here it is necessary to specify δ and a quantity P^* $(0 < P^* < 1)$ such that the probability of choosing the superior variety exceeds P^*. The sequential procedures described here are such that if $\theta_1, \theta_2, \ldots, \theta_{k-1}$ differ among themselves in such a way that their maximum is smaller than θ_k by an amount exceeding δ then the probability of selecting the true highest yielding variety exceeds P^*. Thus, the configuration of the means given in (7.1.2) is the most unfavourable one.

The equations given in (7.1.2) can hardly be expected to hold in practice. Moreover we are interested in discriminating among those varieties which have larger differences among means. We need some protection against classifying a markedly inferior variety as a superior variety.

An immediate consequence of considering this configuration is that, whatever the real means may be, the probability of selecting a variety which differs from the highest yielding variety (unknown) by an amount δ is less than $1-P^*$. Suppose some varieties differ from the highest yielding variety by an amount ledd than δ and some by an amount more than δ. For procedures given in this chapter, the variety selected as the maximum yielding variety will differ from the true highest yielding variety by an amount less than δ with probability greater than $1-P^*$. The probability of selecting a variety whose yield differs by an amount more than δ from the true maximum will be less than $1-P^*$. Under these circumstances δ and P^* should be chosen to give the protection which the experimenter desires.

Bechhofer (1954) considered the configuration given in (7.1.2) for a fixed number of stages. Sequential procedures were also considered by Bechhofer (1958) and Bechhofer and Blumenthal (1962). These procedures are difficult to apply in practice because boundaries are difficult to calculate. Another serious drawback is that one has to go on taking observations from each of the k populations till a final decision is taken. Paulson (1964) considered a very simple sequential procedure with the special feature that the inferior varieties can be progressively discarded. The maximum number of stages can also be chosen under certain restrictions and this will be discussed later. In Section 7.1.1 we shall discuss the procedure for the problem (a). The procedure is essentially due to Paulson (1964). The solution of the problem (b) is on the similar lines and therefore we have described Paulson's method in Section 7.1.1. We have also done a small number of Monte Carlo studies for the optimum choice of the maximum number of stages. Solution of the problem (b) differs considerably from that of (a) because of the special importance of the standard variety. Paulson (1962) considered the solution such that the continuation region is open-ended. In Section 7.2.1 we modify his results by suggesting a closed sequential procedure. The solution of variation on a similar problem will also be considered in Section 7.2.2. In Section 7.3 some results of Monte Carlo studies are given. Section 7.4 is a general discussion and in Section 7.5 we apply these results to a sample data from variety trials.

In the following the word observation will refer to the mean yield of a variety at n sites with r replications within each site. The word 'stages' is used synonymously with 'years'.

7.1.1 Selection of the variety with the largest mean (Paulson's (1964) approach)

In this section we shall consider the solution of the problem (a). We shall start with k varieties, normally distributed with known variance

σ^2. The structure of σ^2 is as described in (6.3.2). To keep the notation simple we shall denote the mean yield of r replications in n sites in stage (year) t of variety j by $x_{jt}(j = 1, 2, \ldots, k;$ $t = 1, 2, \ldots, T)$. In sequential schemes T can be unbounded but here we consider plans for which T is bounded. We standardise the means x_{jt} and δ by dividing them by σ to give

$$z_{jt} = \frac{x_{jt}}{\sigma} \; ; \; \Delta = \frac{\delta}{\sigma} . \qquad (7.1.3)$$

The sequential scheme given in this section can be described as follows:

1) At each stage m and for each variety compute a quantity

$$R_{jm} = \text{Max}(Z_{1m}, Z_{2m}, \ldots, Z_{km}) - Z_{jm}; \quad m = 1, 2, \ldots, T \qquad (7.1.4)$$

where

$$Z_{jm} = \sum_{t=1}^{m} z_{jt} . \qquad (7.1.5)$$

2) Discard all varieties j from further trial for which

$$R_{jm} \geq c - dm \text{ for some } m . \qquad (7.1.6)$$

The parameters $c(> 0)$ and $d(\geq 0)$ are calculated in such a way that Prob [kth variety is selected/$\theta_k - \max(\theta_1, \ldots, \theta_{k-1}) \geq \delta] \geq P^*$. Here $1 - P^*$ is the probability of selecting a variety as the maximum which differs from the true highest yielding variety by an amount δ.

3) Continue the trial for one more year with the remaining varieties and follow the above procedure. Continue the trial until all but one variety is rejected.

The parameters c and d define the sequential boundary and are chosen so as to satisfy the probability requirements. In the following we shall proceed to obtain parameters c and d .

It is not difficult to see that this procedure of discarding the varieties, if adopted at each stage, will satisfy the probability requirement. Let us denote the parametric configuration given in (7.1.2) by H_0 . Let T_0 be the smallest integer $\geq {}^c/d$. After T_0 stages the experimentation stops and the best variety is chosen accordingly. If the variety k is misclassified, then at least for one inferior variety $j \neq k$ and for some stage $m \leq T_0$

$$Z_{jm} - Z_{km} \geq c - dm \text{ .}$$

Now

$$\text{Prob}[\max_{1\leq j\leq k-1} Z_{jm} - Z_{km} \geq c - dm \text{ for some } m/H_0]$$

$$\leq \sum_{\substack{j=1 \\ j\neq k}}^{k-1} P[Z_{jm} - Z_{km} \geq c - dm \text{ for some } m/H_0] \text{ .} \qquad (7.1.7)$$

Now we shall use the results of a known Lemma; (Bechhofer <u>et al</u>, 1968, pp.164)

<u>Lemma</u>: If U_j is a sequence of independent and identically distributed

random variables with a negative expectation then for $c > 0$

$$\text{Prob}\Big[\sum_{j=1}^{m} U_j > c \text{ for some } m < \infty\Big] \leq e^{-h_o c} ; \tag{7.1.8}$$

where h_o is the non-zero root of $E(e^{hU}) = 1$.

Applying this Lemma and using (7.1.7) we obtain

$$\text{Prob}[\text{incorrect decision}/H_0] \leq (k - 1)\, e^{-h_o c} , \tag{7.1.9}$$

where h_o is the non-zero root of

$$E\left[e^{h(z_{jt} - z_{kt} + d)}\right] = 1 \tag{7.1.10}$$

such that $d < \Delta$. By a proper choice of c and d the probability in (7.1.9) can be made to satisfy

$$(k - 1)e^{-h_o c} \leq 1 - P^* . \tag{7.1.11}$$

The solution of (7.1.10) yields

$$h_o = \Delta - d$$

and by substituting in (7.1.11) obtain

$$c = \frac{\log\left[(k-1)/(1-P^*)\right]}{\Delta - d} . \tag{7.1.12}$$

Now all combinations of $c(> 0)$ and $d(< \Delta)$ satisfying (7.1.12) give a family of straight lines $(c - dm)$ defining boundaries for each integer $m \leq \frac{c}{d}$.

When $d = 0$ the boundary is parallel to the x-axis and if $d > 0$ then the line intersects the x-axis at $\frac{c}{d}$. It is not clear what choice of c and d is optimum in the sense of minimising the expected number of observations or stages. It is not difficult to see from the expression (7.1.12) that $\frac{c}{d}$ is minimised by taking $d = \Delta/2$. However, this choice may neither minimise the average number of observations nor the average number of stages. Paulson (1964) conjectures that $d = \Delta/4$ should be near the optimum for minimising the average number of observations. In the variety trials it may be more desirable to minimise the expected number of stages rather than the average number of observations. In Section 7.3 we shall report the results of a small number of Monte Carlo studies and see that there is not much variation in the number of stages for the choice of d near $\Delta/4$.

The inequality used in (7.1.7) is usually too crude and in practice much better probability protection is given by adopting this scheme than is indicated by P^*. Therefore it is desirable to choose a much higher value of $(1 - P^*)$ than that used in a test of significance.

This procedure can be easily extended to select the highest yielding $q(<k)$ varieties. For this we assume the configuration

$$\theta_1 = \theta_2 = \dots = \theta_{k-q} = \theta_{k-q+1} - \delta = \dots = \theta_k - \delta .$$

The above results can be easily extended and the values of c and d are given by

$$c = \frac{\log[q(k-q)/(1-P^*)]}{\Delta - d} . \qquad (7.1.13)$$

The discarding procedure follows on similar lines; at each stage rank the varieties which have not been eliminated already, in ascending order of the Z_{jm}. As soon as the difference between the qth variety from the top, say $[k - q + 1]$, and the remaining varieties of smaller rank exceeds

$$Z_{[k-q+1]m} - Z_{[j]m} \geq c - dm$$

$$[1] \leq [j] \leq [k-q]$$

discard the $[j]$th variety and continue with the remainder till only q varieties are left.

7.2.1 Selection of variety better than a standard variety

The standard (or control) variety plays a special role. We may be interested in retaining the standard variety in use for commercial purposes unless a new variety gives a reasonably higher yield than the standard variety. As before a quantity δ (or Δ) specified by the experimenter gives the minimum difference which he wants to discriminate. In this case two types of probability protection are desired i.e. (i) the standard variety is selected with a probability P_1^* if it is superior to the new varieties; (ii) the new variety is selected with a probability P_2^* if it exceeds the standard variety by an amount δ. Now we shall assume that the kth variety is the standard variety and it is specified in advance. We shall consider the following parametric configuration

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_{k-1} = \theta_k = \theta_0$$

$$H_j : \theta_1 = \theta_2 = \dots = \theta_j - \delta = \theta_{j+1} = \dots = \theta_{k-1} = \theta_k = \theta_0 \qquad (7.2.1)$$

Under H_0 the correct decision is D_k i.e. choose the kth (standard) variety with a probability greater than P_1^* .

$$P(D_k/H_0) \geq P_1^* \ .$$

Under H_j the correct decision is D_j i.e. choose the jth variety with probability greater than P_2^* .

$$P(D_j/H_j) \geq P_2^* \ .$$

As before we shall consider a scheme as follows:

1) Calculate

$$R_{jm} = Z_{jm} - Z_{km} - \frac{m\Delta}{2} \ ; \ (j = 1, 2, \dots, k-1) \ .$$

2) Stop experimentation and select D_j as soon as

$$\max_{1 \leq j \leq k-1} R_{jm} \geq c_1 - d_1 m \ .$$

3) Discard the jth variety if

$$R_{jm} \leq d_2 m - c_2 \ .$$

4) Select D_k as soon as

$$\max_{1 \leq j \leq k-1} R_{jm} \leq d_2 m - c_2 \ .$$

By using a similar argument to that of the previous section and noting that

$$E(z_{jt} - z_{kt} - \frac{\Delta}{2}/H_0) = - \Delta/2 \; ;$$

we obtain the probability of not selecting D_k when H_0 is true as

$$P \left[\underset{1<j<k-1}{\mathrm{Max}} \; Z_{jm} - Z_{km} - \frac{m\Delta}{2} \geq c_1 - d_1 m \text{ for some } m/H_0 \right]$$

$$\leq \sum_{j=1}^{k-1} P\left[Z_{jm} - Z_{km} - \frac{m\Delta}{2} \geq c_1 - d_1 m/H_0\right] \leq (k - 1)e^{-h_1 c_1}$$

where h_1 is the non-zero root of

$$E \left[e^{-h_1(z_{jt} - z_{kt} - \frac{\Delta}{2})}/H_0\right] = 1 \, .$$

The probability of not selecting D_j when H_j is true is equal to the probability that some other variety $(j' \neq j)$ is selected, and is given by

$$\sum_{\substack{j'=1 \\ j'\neq j}}^{k} P(D_{j'}/H_j) = P \left[R_{jm} \leq d_2 m - c_2 \text{ for some } m/H_j\right]$$

$$+ \sum_{\substack{j'=1 \\ j'\neq j}}^{k-1} P \left[R_{j'm} \geq c_1 - d_1 m \text{ for some } m/H_j\right]$$

$$\leq e^{-h_2 c_2} + (k - 2)e^{-h_1 c_1}$$

where h_2 is the non-zero root of

$$E\left[e^{-h_2(z_{jt} - z_{kt} - \frac{\Delta}{2})} / H_j\right] = 1 \quad .$$

Now (c_1, d_1) and (c_2, d_2) should be chosen in such a way that

$$(k - 1)e^{-h_1 c_1} \leq 1 - P_1^*$$

$$e^{-h_2 c_2} + (k - 2)e^{-h_1 c_1} \leq 1 - P_2^* \quad . \tag{7.2.2}$$

For simplicity we shall consider only those values of (c_1, d_1) and (c_2, d_2) for which sequential boundary intersects at the x-axis, say, at $t = T$. Then

$$\frac{c_1}{d_1} = \frac{c_2}{d_2} = T \quad . \tag{7.2.3}$$

Now putting $h_1 = \frac{\Delta}{2} - d_1$ and $h_2 = \frac{\Delta}{2} - d_2$ in equations (7.2.2) and using a constant λ $(0 > \lambda \geq 1)$ we have two sets of equations

$$(k - 1)e^{d_1(d_1 - \frac{\Delta}{2})T} = \lambda(1 - P_1^*)$$

$$e^{d_2(d_2 - \frac{\Delta}{2})T} + (k-2)e^{d_1(d_1 - \frac{\Delta}{2})T} = 1 - P_2^* \quad .$$

Solving these two equations for d_1 and d_2 for given λ and T we have

$$d_1 = \frac{\Delta}{4} - \frac{1}{2}\sqrt{\frac{\Delta^2}{4} + \frac{4 \log s_1^*}{T}}$$

$$d_2 = \frac{\Delta}{4} - \frac{1}{2}\sqrt{\frac{\Delta^2}{4} + \frac{4 \log s_2^*}{T}} \qquad (7.2.4)$$

where

$$s_1^* = \frac{\lambda(1 - P_1^*)}{k - 1}$$

$$s_2^* = (1 - P_2^*) - \frac{(k-2)}{(k-1)} \lambda (1 - P_1^*) = (1 - P_2^*) - (k - 2)s_1^* .$$

For different values T and λ we obtain different sequential boundaries satisfying the probability requirements. It is clear from (7.2.4) that for obtaining solutions T must satisfy

$$T \geq \text{Max.}\left[- \frac{16 \log s_1^*}{\Delta^2}, - \frac{16 \log s_2^*}{\Delta^2}\right] . \qquad (7.2.5)$$

For variation in λ from 0 to 1 , $- \log s_1^*$ decreases and $- \log s_2^*$ increases. We shall choose that value for which both intersect i.e.

$$\lambda^* = \frac{1 - P_2^*}{1 - P_1^*} .$$

Then λ is chosen as λ^* given by

$$\lambda^* = \text{Min.} \left[1, \frac{1 - P_2^*}{1 - P_1^*}\right] . \qquad (7.2.6)$$

It is not clear what value of T should be optimum but, it is hoped that minimum T satisfying (7.2.5) and (7.2.6) should not be far away from the optimum. Once d_1 and d_2 are obtained with the help of (7.2.4) it is not difficult to obtain c_1 and c_2.

We shall consider a particular case where $P_1^* = P_2^* = P^*$; For this case $s_1^* = s_2^*$. Minimum T is given by

$$T = \frac{16 \log \left[(k-1)/(1-P^*)\right]}{\Delta^2}$$

corresponding to $d = \Delta/4$.

The method discussed above is a generalisation of Paulson's (1962) as he has considered the case of $d = 0$ corresponding to $T = \infty$. Usually this type of selection procedure, giving double protection, requires a large number of observations. For $k = 2$ this procedure is similar to the sequential method of a testing hypothesis with a fixed probability for the first and second kind of errors.

7.2.2 Comparison of the standard variety with the new varieties

In some cases when many new varieties are compared with a standard then there is a tendency to compare the highest yielding new variety with the standard and take a decision on this basis. When this is done then the usual probability levels do not hold and sometimes a decision based on such comparisons can be very misleading. The results of the Lemma given in the Section 7.1.1 can be used to give some protection to the experimenter. In this case the experimenter chooses two quantities $\delta (> 0)$ and P^* such that the probability of selecting a new variety which is worse than the standard by an amount $\delta (\delta > 0)$ should be less than $1 - P^*$. The most unfavourable configuration of means is given by

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_k - \delta$$

where the kth variety is taken as the known standard variety. The sequential scheme for this is as follows.

If

$$\operatorname*{Max}_{1 \leq j \leq k-1} Z_{jm} - Z_{km} \geq c - dm$$

for some m, then select the variety corresponding to the max. Z_{jm}. Discard the variety from further trials if

$$Z_{jm} - Z_{km} \leq dm - c \; ; \quad j \neq k \, .$$

Using the Lemma the values of c and d are given by

$$c = \frac{\log \left[(k - 1)/(1 - P^*) \right]}{\Lambda - d} \; ; \; 0 \leq d < \Lambda \; . \tag{7.2.7}$$

Note that the values of c and d obtained for this purpose are the same as given in (7.1.12).

In a particular case of interest there are only two varieties of which one is standard and the other new. Suppose the second variety is the standard variety with mean θ_0 (unknown). We are interested in comparing it with the new variety in such way that if

$$\theta_1 > \theta_0 + \delta$$

then the probability of preferring the standard is smaller than $1 - P^*$. In the same way if

$$\theta_1 < \theta_0 - \delta$$

then the probability of preferring new variety is smaller than $1 - P^*$. Proceed in the usual way and stop and accept the new variety if

$$Z_{1m} - Z_{2m} \geq c - dm \;\; ;$$

stop and accept the standard variety if

$$Z_{1m} - Z_{2m} \leq dm - c \;\; ;$$

where c and d are obtained from (7.2.7) by putting $k = 2$.

7.3 Results of some Monte Carlo studies

We have done a small number of Monte Carlo (MC) studies to examine the performance of the above mentioned sequential procedures for six and ten varieties. From the data of variety trials conducted on cereals it appears that the value of Δ varies from 1.5 to 2.5 for detecting five to ten percent of differences among the mean yields of varieties. Therefore, for the MC studies we have chosen Δ about 2.0. The observations x_{jt} were generated on a computer as random normal observations with unit standard deviation and mean θ_j. Different configurations of θ_j were considered and are given in Table 7.3.1. All the results are based on the differences within each stage and are not therefore affected by fixed additive stage effect. The Z_{jm} were calculated by the method described in (7.1.5). As the value of P^* increases for fixed k and Δ the maximum number of stages (c/d) increases. In the present study we have restricted our experiments to a small number of stages with $T_0 = 4$, 5 and 6 and $P^* = 0.85$. The results reported here are based on 100 sampling experiments in each case.

In Table 7.3.1, ASN gives the average number of observations (average based on 100 sampling experiments) required for final selection of variety. Average number of stages required for the final selection of variety is also given. Empirical probability gives the fraction of the experiments in which correct variety was chosen, out of 100 experiments.

The minimum value of c/d is obtained by taking $d = \Delta/2$, as discussed in Section 7.1.1. $T_0 = 4$ is the nearest integer corresponding to minimum c/d. Paulson (1964) suggested the use of $d = \Delta/4$ and $T_0 = 5$ corresponds to this value of d approximately. $T_0 = 6$ is larger than the above two choices of T_0.

Now for both studies $(k = 6, 10)$ it appears that there is a considerable discarding of inferior varieties at the early stages resulting in saving of resources as indicated by the ASN. For $d = \Delta/4$ the number of stages to reach the decision seems to be a minimum and the experiment stops much before T_o. As T_o increases the empirical probability of the correct decision increases. As mentioned earlier the empirical probability of the correct decision is higher than is indicated by $P^* = 0.85$.

In the end of Table 7.3.1 we give the results on the performance of this scheme for an alternative configuration, H_A in which θ_i varies uniformly

TABLE 7.3.1

RESULTS FOR SELECTION OF THE HIGHEST YIELDING VARIETY

Number of varieties (k)	Λ	T_o	c	d	ASN	Empirical probability of correct decision	Average number of stages
Configuration $H_o(\theta_i = 0;\ i = 1, 2, \ldots k-1;\ \theta_k = \Lambda)$							
6	2.00	4	2.592	0.648	10.22	$P_k = 0.90$	2.34
		5	2.265	0.453	9.91	$P_k = 0.93$	2.28
		6	2.132	0.355	10.01	$P_k = 0.95$	2.34
10	2.10	4	3.076	0.769	16.66	$P_k = 0.90$	2.63
		5	2.585	0.517	16.66	$P_k = 0.90$	2.60
		6	2.412	0.402	16.57	$P_k = 0.92$	2.63
Configuration $H_A(\theta_i = i.\Lambda/10.0;\ i = 1, 2, \ldots, k)$							
10	2.10	4	3.076	0.769	21.25	$\sum_{i=8}^{10} P_i = 0.78$	3.31
		5	2.585	0.517	20.30	$\sum_{i=8}^{10} P_i = 0.85$	3.49
		6	2.412	0.402	21.40	$\sum_{i=8}^{10} P_i = 0.86$	3.85
Configuration $H_A(\theta_1 = \theta_2 = \ldots = 0;\ \theta_9 = 0.9\Lambda;\ \theta_{10} = \Lambda)$							
10	2.10	4	3.076	0.769	17.44	$P_9 = 0.43; P_{10} = 0.56$	3.00
		5	2.585	0.517	16.95	$P_9 = 0.42; P_{10} = 0.57$	3.08
		6	2.412	0.402	17.09	$P_9 = 0.41; P_{10} = 0.58$	3.25

between 0 and Δ. As expected the ASN and the average number of stages increase. The minimum ASN is obtained by taking d near $\Delta/4$ ($T_o = 5$), but the minimum number of stages by taking $d = \Delta/2$ ($T_o = 4$). As far as empirical probability is concerned, it will be more appropriate to pool the probabilities of the highest two or three populations, because the differences among them are so small that even if they are wrongly classified it will not result in much economic loss. We have pooled the empirical probabilities of selecting the highest three populations and they are 0.78, 0.85 and 0.86 corresponding to $T_o = 4$, 5 and 6, respectively. This shows that even under a very different configuration of means these sequential schemes will select the superior populations with a high probability. Another configuration, in which one variety is very close to the highest yielding variety and the rest are at the other end, gives similar results as indicated in Table 7.3.1. There is a considerable saving in the number of observations and in the number of stages required to reach the final conclusion. From these studies it appears that $d = \Delta/4$ is an adequate choice.

In Table 7.3.2 we have taken one variety (k^{th}) as the standard and compared with the maximum of those remaining as discussed in Section 7.2.2. The results follow very similar lines as that of Table 7.3.1. We have also considered an alternative configuration H_A where the mean varies uniformly between o and Δ. The standard variety has mean Δ. The pooled probabilities for the highest four populations have been obtained as 0.89, 0.85 and 0.86 for $T_o = 4$, 5 and 6, respectively. There is a considerable

TABLE 7.3.2

RESULTS FOR SELECTING THE HIGHEST YIELDING VARIETY IN RELATION TO THE STANDARD VARIETY

Number of varieties (k)	Δ	T_o	c	d	ASN	Empirical probability of correct decision	Average number of stages
Configuration $H_o(\theta_i = o;\ i = 1,2,\dots,k-1;\ \theta_k = \Delta)$							
6	2.00	4	2.592	0.648	10.49	$P_k = 0.94$	2.34
		5	2.265	0.453	9.99	$P_k = 0.91$	2.21
		6	2.132	0.355	10.19	$P_k = 0.93$	2.30
10	2.10	4	3.706	0.769	18.05	$P_k = 0.90$	2.66
		5	2.585	0.517	17.18	$P_k = 0.90$	2.60
		6	2.412	0.402	17.04	$P_k = 0.93$	2.60
Configuration $H_A(\theta_i = i.\Delta/10.00;\ i = 1,2,\dots,k)$; k^{th} variety is the standard variety							
10	2.10	4	3.706	0.769	21.37	$\sum_{i=7}^{10} P_i = 0.89$	2.88
		5	2.585	0.517	21.23	$\sum_{i=7}^{10} P_i = 0.85$	3.02
		6	2.412	0.402	20.81	$\sum_{i=7}^{10} P_i = 0.86$	3.09

saving in the number of stages and in the number of observations required even under H_A .

Thus by a proper choice of Δ and P^* these schemes may help in choosing the higher yielding varieties, even when the configuration differs considerably from the one assumed for obtaining the analytical results. These schemes may lead to a considerable saving of resources.

7.4 Discussion of the results

Closed sequential plans usually lead to a considerable saving of resources when there are no differences among population means. The problem considered here is not a hypothesis-testing problem, but for $k = 2$ the method discussed in Section 7.2.1 is similar. The closed sequential plans have an advantage over the open ones, in that one can give an upper bound to the maximum number of stages. This upper bound on the number of stages (T_o) will depend on the values of P^*, Δ and k . The actual number of stages required to reach a decision is a random variable and can be considerably smaller than T_o as observed in MC studies.

In the variety trials when interaction component (variety-year) of σ^2 i.e. σ_{vy} is relatively smaller than σ_{vs} , σ can be reduced considerably by increasing n . In cases when σ_{vy} is negligibly small then there is not much advantage in conducting trials over years but in some cases this may lead to a considerable saving of resources or better utilisation of resources. In the data of variety trials which we have analysed, the variety - year interaction is substantial and it is desirable to spread the trials over years. It is necessary to keep σ as small as possible because for the same δ and P^* the value of T_o can be made small by decreasing σ. By increasing n the value of σ can be made small. However, the rate of decrease in σ decreases as n increases, particularly when σ_{vy} is relatively large. No optimum value of n can be obtained unless some cost function approach is assumed and we shall not consider this any further. For each case the rate of decrease in σ should be obserbed by varying n and r, and when no substantial change is found by increasing n and r any further, those values should be chosen. In the above discussions we have assumed that the value of σ is known but it is rarely known in practice. When σ is unknown then the analytical solution for this problem becomes very difficult. Hence we shall assume that a good estimate of σ is available from past experience.

The other two parameters required are δ and P^*. As we have seen P^* is an upper bound of the probability, therefore it should be chosen lower than what is usually used in testing a hypothesis. The choice P^* near 0.90 should be adequate for all practical purposes. The choice of δ depends much on the experimenter. In most cases five to seven percent of the mean should be adequate. The value of T_o and the boundaries are very sensitive to Δ and they increase very rapidly as Δ decreases.

From the MC studies it appears that when we want to select for the highest yielding variety then there is a considerable saving in the actual number of observations and stages required, even under H_A. In most cases the choice $d = \Delta/4$ should be adequate. In cases when more than one variety, say, q varieties are highest yielding (and may have almost similar means among themselves) and the rest are inferior then the inferior varieties will be first to be eliminated. Grouping of q varieties near the maximum will be apparent from the cumulative totals as they will be very close to each other even after several stages of experimentation. In such cases it may be advisable to select all the q highest yielding varieties, since be pursuing such a scheme to the final stage may only increase the number of stages, and there may not be much to gain by discriminating among such varieties.

When comparing with the standard variety two types of probability protection may be necessary, as described in Section 7.2.1. This usually increases T_o and makes the continuation region much wider. This also increases the expected number of stages and expected number of observations required. This is similar to the case of testing a hypothesis, when probabilities of the first and the second kind of errors are fixed. We have not done any MC studies for this case but we have considered an example and this will be discussed in the next section. When some of the varieties (at least one) is markedly superior than the standard (or the standard is markedly superior than the others) then it is expected that the number of stages required for reaching decision will be considerably smaller than T_o, and the experimental program will terminate much before T_o.

In practical situations the discarded varieties in the trials will be replaced by new varieties at every stage and it will also be desirable to compare the varieties which are at different stages. Any analytical solution for this problem is not easy to obtain because the effect of stages will also be involved. Even if this is eliminated the problem of different variances for different comparisons may be difficult to tackle. However, it is possible to apply the methods discussed here using information on the varieties based on the common number of stages.

7.5 Application of the results

In Table 7.5.1 we have reproduced the means of six varieties (mean over fifteen or sixteen sites within each year) from Table 3.3.3.

TABLE 7.5.1

WHEAT VARIETY MEAN YIELDS (GRAINS t/ha)

Variety Names	x_{jt}			Standardized cumulative totals Z_{jm}		
	1969	1970	1971	1969	1970	1971
Cappelle	5.33	4.95	5.01	28.50	54.97*	(81.76)
Bouquet	5.39	5.06	5.43	28.82	55.88*	(84.92)
Cama	5.62	5.33	5.12	30.05	58.55	85.93*
Maris Beacon	5.88	5.49	5.87	31.44	60.80	92.19
Maris Settler	5.65	5.22	4.82	30.21	58.12	83.89*
Tommy	5.43	5.12	5.23	29.04	56.42*	(84.39)

We shall use the estimated value of $\sigma = 0.187$* (from Table 3.3.6). Suppose we want to search for the highest yielding variety such that the probability of discriminating the maximum from the next, which may differ by about 5% of the general mean ($\delta = 0.25$), is greater than 0.90 ($P^* = 0.90$). This specification immediately gives

---

* where σ is defined in (7.1.1) and is estimated by $\sqrt{\frac{.536}{15.333}}$

$$\Delta = \frac{\delta}{\sigma} = 1.337 .$$

Now using the sequential scheme discussed in Section 7.1.1 and taking $d = \Delta/4$ we obtain

$$c = \frac{4 \log [5/(1-P^*)]}{3\Delta} = 3.901$$

$$d = \Delta/4 = 0.334 .$$

For this choice of c and d, T_0 is given by 12.

The standardised cumulative totals for each variety are also given in Table 7.5.1. The boundaries of continuation region $(c - dm)$ are goven below:

m	1	2	3	4
c-dm	3.567	3.233	2.899	2.565

As soon as the difference between any variety and the highest yielding variety exceeds c-dm at the m^{th} stage the lower yielding variety should be discarded from future trials. The varieties marked with asterisks should be discarded from future trials. The trial ends at the third stage $(m = 3)$ with the selection of Maris Beacon.

As discussed in Section 7.2.1, when new varieties are compared with the standard variety then two types of probability protection are desired. This usually increases the sample size and makes these boundaries much wider. In this example we shall consider the case $k = 2$, taking Cama as the standard variety and Maris Beacon as the new variety. We shall take $P_1^* = P_2^* = P^* = 0.90$

and $\Delta = 1.337$. Taking $d = \Delta/4$ (this choice of d minimises T_o),

$$c = \frac{4 \log \left[1/1(1 - P^*)\right]}{\Delta} = 6.889$$

$$d = \Delta/4 = 0.334$$

m	1	2	3	4
c-dm	6.553	6.221	5.887	5.553
$R_{jm} = (Z_{4m} - Z_{3m} - \frac{m\Delta}{2})$	0.721	0.913	4.255	-

These boundaries can be represented graphically and are shown in Graph 7.5.1. Further observations are needed as long as R_{4m} remains within the triangle which represents the continuation region. As soon as R_{4m} crosses the triangle the appropriate decision should be taken. For this particular case some more observations are needed for reaching any definite conclusion.

In some cases only one type of protection is needed when comparing with the standard variety, i.e. the probability that the selected highest yielding variety (from the new varieties) is worse than the standard variety by more than Δ, does not exceed $1 - P^*$. For this, a similar continuation region is obtained as given in selecting the highest yielding variety. Taking the same values of P^*, Δ, c and d we obtain $c-dm$ for different values of m. Now taking Cama as the standard variety we can calculate the cumulative differences between each variety and the standard variety and they are given in Table 7.5.3. These differences have been plotted in Graph 7.5.2. From this graph it is evident that Cappelle should be rejected after the second year.

TABLE 7.5.3

CUMULATIVE DIFFERENCES OF VARIETIES FROM CAMA (STANDARD)

Variety Name	1969	1970	1971
1. Cappelle	-1.55	-3.58	-4.17
2. Bouquet	-1.23	-2.67	-1.01
3. Cama	-	-	-
4. Maris Beacon	1.39	2.25	6.26
5. Maris Settler	0.16	-0.43	-2.04
6. Tommy	-1.01	-2.13	-1.54

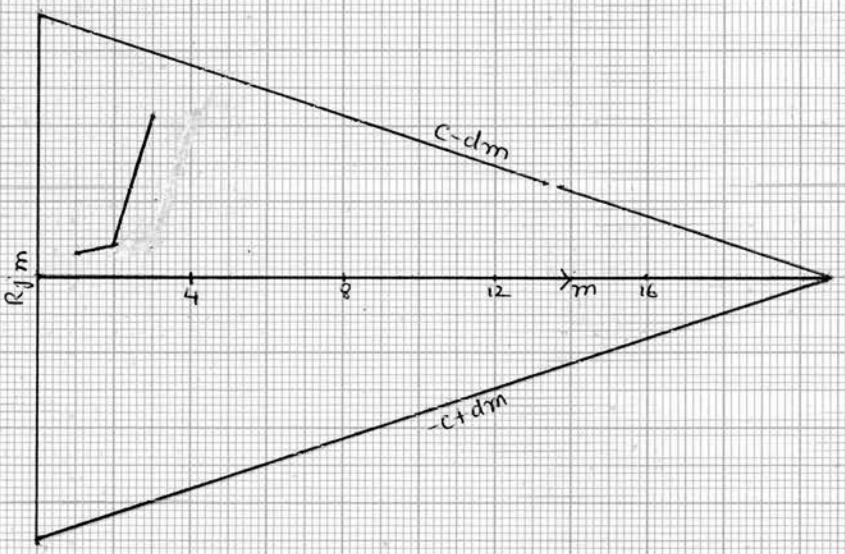
The trial in the third year results in the selection of Maris Beacon and no more trials are necessary. This procedure is similar to that considered at the beginning of this section. Go on conducting trials and discarding the varieties on the basis of their difference with the highest yielding variety. As soon as the standard variety is discarded no more trials are necessary and the highest yielding variety is selected. If the aim is to select the highest yielding variety then further trials may be necessary till only one variety is left.

GRAPH 7.5.1

SEQUENTIAL BOUNDARY FOR COMPARING STANDARD WITH A NEW VARIETY

STANDARD = CAMA

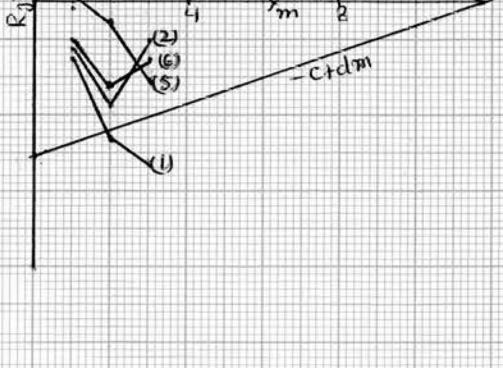
NEW VARIETY = MARIS BEACON



GRAPH 7.5.2

SEQUENTIAL BOUNDARY FOR COMPARING STANDARD WITH MANY NEW VARIETIES

STANDARD = CAMA



REFERENCES

Anderson, T.W.(1958). *Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Anderson, T.W. and Rubin, H.(1956). Statistical inference in factor analysis. *Proc. Third Berkeley Symp. Math. Statist. Probab. 5, 111-150*.

Baker, R.J.(1969). Genotype-environment interactions in yield of wheat. *Can. J. Plant Science* 49, 743-751.

Bartlett, M.S.(1951). The goodness of fit of a single hypothetical discriminant function in the case of several groups. *Ann of Eug.* 16, 199-214.

Bechhofer, R.E.(1954). A single-sample multiple-decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.* 25, 16-39.

Bechhofer, R.E.(1958). A sequential multiple decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics* 14, 408-429.

Bechhofer, R.E. and Blumenthal, S.(1962). A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance II: Monte Carlo sampling results and new computing formulae. *Biometrics* 18, 52-67.

Bechhofer, R.E., Kiefer, J. and Sobel, M.(1968). *Sequential Identification and Ranking Procedures*. The University of Chicago Press, Chicago and London.

Box, G.E.P.(1949). A general distribution theory for a class of likelihood criteria. *Biometrika* 36, 317-346.

Box, G.E.P.(1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statist.* 25, 290-302

Box, G.E.P.(1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effect of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Statist*. 25, 484-498.

Box, G.E.P. and Cox, D.R.(1964). An analysis of transformations. *J. Roy. Statist. Soc. B*. 26, 211-252.

Caliňski, T.(1966). On the distribution of the F-type statistics in the analysis of group of experiments. *J. Roy. Statist. Soc. B*. 28, 526-542.

Cochran, W.G.(1937). Problem arising in the analysis of a series of similar experiments. *J. Roy. Statist. Soc. Supp*. 4, 102-118.

Cochran, W.G.(1954). The combination of estimates from different experiments. *Biometrics* 10, 101-129.

Cochran, W.G. and Cox, G,M.(1957). *Experimental Designs*. 2nd ed., Wiley, New York.

Curnow, R.N.(1961). Optimal programmes for varietal selection. *J. Roy. Statist. Soc. B*, 23, 282-318.

Dunnett, C.W.(1960). On selecting the largest of k normal population means. *J. Roy. Statist. Soc. B*, 22, 1-40.

Ehrenberg, A.S.C.(1950). The unbiased estimation of heterogeneous error variances. *Biometrika* 37, 347-357.

Elandt, R.C.(1963). Optimal and sufficient allocation of multiple varietal experiments. *Biometrics* 19, 615-628.

Finlay, K.W. and Wilkinson, G.N.(1963). The analysis of adaptation in a plant breeding program. *Aust. J. Agr. Res.* 14, 742-754.

Finney, D.J.(1958a). Statistical problems in plant selection. *Bull. Int. Statist. Inst*. 36, 242-268

Finney, D.J.(1958b). Plant selection for yield improvement. *Euphytica* 7, 83-106.

Finney, D.J.(1960). A simple example of the external economy of varietal selection. *Bull. Int. Statist. Inst*. 37, 91-106.

Fisher, R.A. and Mackenzie, W.A.(1923). Studies in crop variation. II. The manurial response of different potato varieties. *J. Agr. Sci*. 13, 311-320.

Fisher, R.A.(1924). The influence of rainfall on the yield of wheat at Rothamsted, *Phil. Trans. Roy. Soc. B*, 213, 89-142.

Fisher, R.A.(1949). *The Design of Experiments*. 5th ed. Oliver and Boyd, Edinburgh.

Freeman, G.H. and Perkins, J.M.(1971). Environmental and genotype-environmental components of variability. VIII. Relations between genotype grown in different environments and measure of these environments. *Heredity* 26, 15-23

Gollob, H.F.(1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33, 73-115.

Gomes, F.P. and Guimarães, R.F.(1958). Joint analysis of experiments in complete randomised blocks with some common treatments. *Biometrics* 14, 521-526.

Graybill, F.A.(1954). Variance heterogeneity in a randomised block design. *Biometrics* 10, 516-520.

Grubbs, F.E.(1948). On estimating precision of measuring instruments and product variability. *J. Amer. Statist. Ass*. 43, 243-264.

Grubbs, F.E.(1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics* 15, 53-66.

Grundy, P.M., Healy, M.J.R. and Rees, D.H.(1956). Economic choice of the amount of experimentation. *J. Roy. Statist. Soc. B*, 18, 32-55.

Han, C.P.(1969). Testing the homogeneity of variances in a two-way classification. *Biometrics* 25, 153-158.

Hardwick, R.C. and Wood, J.T.(1972). Regression methods for studying genotype-environment interactions. *Heredity* 28, 209-223.

Hill, G.W. and Davies, A.W.(1968). Generalised asymptotic expansions of Cornish-Fisher type. *Ann. Math. Statist*. 39, 1264-1273.

Jaech, J.L.(1971). Further tests of significance of Grubbs's estimators. *Biometrics* 27, 1097-1101.

Jinks, J.L. and Perkins, J.M.(1970). Environmental and genotype-environmental components of variability. VII. Simultaneous prediction across environments and generations. *Heredity*. 25, 474-480.

Johnson, N.L.(1962). Some notes on the investigation of heterogeneity in interactions. *Revista Trabajos De Estadistica* 13, 183-199.

Kshirsagar, A.M.(1961). The goodness of fit of a single (non-isotropic) hypothetical principal component. *Biometrika* 48, 397-407.

Lawley, D.N. and Maxwell, A.E.(1971). *Factor Analysis as a Statistical Method*. 2nd ed., Butterworths, London.

Maloney, C.J. and Rastogi, S.C.(1970). Significance test for Grubbs's estimators. *Biometrics* 26, 671-676.

Mandel, J.(1961). Non-additivity in two-way analysis of variance. *J. Ame. Statist. Ass*. 56, 878-888.

Mandel, J.(1969). The partitioning of interaction in analysis of variance. *J. Res. National Bureau of Standards-B. Math. Sci.* 73B, No. 4, 309-328.

Mandel, J.(1971). A new analysis of variance model for non-additive data. *Technometrics* 13, 1-18.

Mauchly, J.W.(1940). Significance test for sphericity of a normal n-variate distribution. *Ann. Math. Statist.* 11, 204-209.

McIntyre, G.A.(1972). Homogeneity of variance in a two-way classification. Letter. *Biometrics* 28, 246.

Morgan, W.A.(1939). A test for the significance of the difference between the two variances in a sample from a normal bi-variate population. *Biometrika* 31, 13-19.

Neyman, J., Iwaszkiewicz, K. and St. Kolodziejczyk.(1935). Statistical problems in agricultural experimentation. *J. Roy. Statist. Soc. Supp.* 2, 107-180.

Paulson, E.(1962). A sequential procedure for comparing several experimental categories with a standard or control. *Ann. Math. Statist.* 33, 438-443.

Paulson, E.(1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.* 35, 174-180.

Pavate, M.V.(1961). Combined analysis of balanced incomplete block designs with some common treatments. *Biometrics* 17, 111-119.

Perkins, J.M. and Jinks, J.L.(1968). Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. *Heredity* 23, 339-356.

Pitman, E.J.G.(1939). A note on normal correlation. *Biometrika* 31, 9-12.

Raiffa, H. and Schlaifer, R.(1961). *Applied Statistical Decision Theory*. Harvard University. Boston.

Rao, C.R.(1970). Estimation of heteroscedastic variances in linear models. *J. Ame. Statist. Ass.* 65, 161-172.

Rao, C.R.(1972). Estimating variances and covariances components in linear models. *J. Ame. Statist. Ass.* 67, 112-115.

Russell, T.S. and Bradley, R.A.(1958). One-way variance in a two-way classification. Biometrika 45, 111-129.

Scheffé, H.(1956). A mixed model for the analysis of variance. Ann. Math. Statist. 27, 23-36.

Shukla, G.K.(1972a). Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29, 237-245.

Shukla, G.K.(1972b). An invariant test for the homogeneity of variances in a two-way classification. Biometrics 28, 1063-1072.

Shukla, G.K.(1973). Some exact tests of hypotheses about Grubbs's estimators. Biometrics 29, 373-377.

Sprague, G.F. and Federer, W.T.(1951). A comparison of variance components in corn yield trials: ii. Error, year × variety, location × variety, and variety components. Agronomy Journal 43, 535-541.

Tai, G.C.C.(1971). Genotype stability analysis and its application to potato regional trials. Crop Science 11, 184-190.

Tukey, J.W.(1949). One degree of freedom for non-additivity. Biometrics 5, 232-242.

Tukey, J.W.(1962). The future of data analysis. Ann. Math. Statist. 33, 1-67.

Tyagi, B.N., Kathuria, O.P. and Rao, P.R.(1970). The analysis of group of experiments involving several factors. J. Ind. Soc. Agr. Statist. 22, 27-42.

Whittle, P.(1953). A principal component and least squares method of factor analysis. Skand. Akt. 35, 223-239.

Wilks, S.S.(1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. Ann. Math. Statist. 17, 257-281.

Williams, E.J.(1952). The interpretation of interaction in factorial experiments. Biometrika 39, 65-81.

Yates, F. and Cochran, W.G.(1938). The analysis of groups of experiments. J. Agr. Sci. 28, 556-580.

Yates, F.(1952). Principles governing the amount of experimentation in developmental work. Nature 170, 138.

Some Statistical Aspects of Partitioning Genotype-Environmental Components of Variability

G. K. SHUKLA

A.R.C. Unit of Statistics, 21 Buccleuch Place, University of Edinburgh, Edinburgh, EH8 9LN

Reprinted from HEREDITY, Volume 29, Part 2, pp. 237-245, October 1972

SOME STATISTICAL ASPECTS OF PARTITIONING GENOTYPE-ENVIRONMENTAL COMPONENTS OF VARIABILITY

G. K. SHUKLA
A.R.C. Unit of Statistics, 21 Buccleuch Place, University of Edinburgh, Edinburgh EH8 9LN

Received 17.xii.71

1. INTRODUCTION

RECENTLY, Freeman and Perkins (1971) examined some of the existing methods of partitioning genotype-environmental component of variability and their statistical validity. They have considered the usual practice of calculating the regression of genotype means on the environmental means calculated by taking the average of all genotypes in that environment, first used by Yates and Cochran (1938) and later used by Finlay and Wilkinson (1963) and since then used by several other authors of which references have been cited by Freeman and Perkins (1971). They have shown the statistical invalidity of using such regressions and their sums of squares for testing homogeneity. They have further suggested that some genotypes should be taken in each environment (not included in calculating the mean of each genotype) as a measure of environment and regression of genotype means should be calculated on the independent measure of environment thus making the procedure statistically more valid.

In the first part of this paper we have shown how by looking at the model in a different way one could draw statistically valid conclusions of certain hypotheses from the same analysis of Yates and Cochran (1938). In the second part of this paper we have suggested a method of estimating a component of genotype-environmental interaction corresponding to each genotype, thus giving a better measure of genotype stability. This paper is concerned with the presentation of practical methods rather than with statistical theory (to which references are given); it does not itself contain much that is new, but a more general statistical treatment is being prepared and will be published elsewhere.

2. REGRESSION OF GENOTYPE MEANS ON ENVIRONMENTAL MEANS

We have used mostly the same model and notation as used by Perkins and Jinks (1968) for t genotypes, s environments and r replications of each genotype within each environment and the model could be represented as,

$$y_{ijk} = \mu + d_i + \epsilon_j + g_{ij} + e_{ijk}, \tag{1}$$

where μ is the grand mean, $d_i (i = 1, \ldots, t)$ the additive genetic contribution of the ith genotype, $\epsilon_j (j = 1, \ldots, s)$ the additive environmental contribution of the jth environment, g_{ij} the genotype-environment interaction of the ith genotype in the jth environment and $e_{ijk} (k = 1, \ldots, r)$ is the residual variation contributed by the kth replicate of the ith genotype in the jth environment. We shall assume that

$$E(\bar{e}_{ij}) = 0; \;\; V(\bar{e}_{ij}) = \sigma_0^2, \tag{2}$$

for all i and j where $\bar{e}_{ij}$ is the mean of e_{ijk} over r replications and σ_0^2 is the within environment error variance for the mean of r replications. We shall further assume that environment effects ϵ_j's are random effects with population mean zero and variance σ_ε^2. A random sample of s environments has been selected from an infinite population of environments. We shall estimate σ_0^2 as usual by $\hat{\sigma}_0^2$, where

$$\hat{\sigma}_0^2 = \sum_i \sum_j \sum_k (y_{ijk}-\bar{y}_{ij})^2/str(r-1), \tag{3}$$

with $st(r-1)$ degrees of freedom. Hereafter we shall work with $\bar{y}_{ij}$ (the mean of r replications of the ith genotype at the jth environment). In the present paper we shall confine ourselves to the genotype-environment $(G \times E)$ part of the analysis of variance.

Working with the means, the model in (1) can be written

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j + g_{ij} + \bar{e}_{ij}. \tag{4}$$

Putting

$$g_{ij} = b_i'\epsilon_j + \eta_{ij} \quad \text{and} \quad \alpha_{ij} = \eta_{ij} + \bar{e}_{ij}$$

we obtain from (4)

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j + b_i'\epsilon_j + \alpha_{ij}. \tag{5}$$

Putting $b_i = b_i' - \bar{b}'$ where $\bar{b}' = \sum_i b_i'/t$ we obtain from (5)

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j(1+\bar{b}') + b_i\epsilon_j + \alpha_{ij}. \tag{6}$$

The model in (5) is reparameterised in (6) in such a way that $\sum_i b_i = 0$. When all b_i's are 0 (or b_i' are equal) then the model in (6) becomes as in (7),

$$\bar{y}_{ij} = \mu + d_i + \epsilon_j(1+b') + \alpha_{ij}. \tag{7}$$

Thus the problem of testing the equality of all b_i''s becomes the problem of testing model (7) against model (6). This is equivalent to testing the presence of the non-additivity term $b_i\epsilon_j$ in (6) when ϵ_j are taken as fixed effects. The test for presence of non-additivity of this type was given by Mandel (1961), which is a generalisation of Tukey (1949). Same results hold good even if ϵ_j are taken as random effects. He has estimated b_i by $\hat{b}_i$ and calculated the sum of squares due to non-additivity (S) as follows:

$$\hat{b}_i = \frac{\sum_j (\bar{y}_{ij}-\bar{y}_{.j})(\bar{y}_{.j}-\bar{y}_{..})}{\sum_j (\bar{y}_{.j}-\bar{y}_{..})^2},$$
$$S = \Sigma_i \hat{b}_i^2 \Sigma_j (\bar{y}_{.j}-\bar{y}_{..})^2, \tag{8}$$

and " Balance " = Interaction Sum of Square (G × E) − S.

The sum of squares in (8) is similar to that calculated by Yates and Cochran (1938) and the same as that due to heterogeneity in regression in Freeman and Perkins (1971) table 2, with z_j in Freeman and Perkins' notations replaced by $\bar{y}_{.j}-\bar{y}_{..}$. The estimated regression coefficients, $\hat{b}_i$, here are similar to those given by Yates and Cochran (1938) except that we have regressed deviation of genotype means from the environmental means rather than genotype means. The sum of squares S indeed is a ratio of quartic terms and quadratic terms in the y's.

In the absence of interaction ($g_{ij} = 0$ for all i and j) it can be shown that S/σ_0^2 and "Balance"$/\sigma_0^2$ are both independently distributed as χ^2 on $(t-1)$ and $(t-1)(s-2)$ degrees of freedom, respectively. However, in the presence of interaction S/σ_0^2 and "Balance"$/\sigma_0^2$ are not distributed as χ^2 even if all $b_i = 0$, though they are independently distributed of each other. In the presence of interaction the appropriate test statistic for all $b_i = 0$ will be F' as given in (9). ?

$$F' = \frac{S/(t-1)}{\text{“Balance”}/(t-1)(s-2)} \tag{9}$$

F' will be distributed as F on $(t-1)$ and $(t-1)(s-2)$ degrees of freedom. The same test was proposed by Perkins and Jinks (1968) and this statistic gives correct probability level. Equality of any two b_i's could be tested by doing the similar analysis for any two genotypes of interest. The above argument can be generalised for other arrangements of genotypes within and between environments and also in the presence of non-orthogonality in the data (Milliken and Graybill, 1970).

To use the $\hat{b}_i$'s in the usual sense of regression would not be valid, but they give rough guidance about the relation of genotype means to environmental means. The $\hat{b}_i$'s are biased estimators of the b_i's. In general this bias will be small when σ_ε^2 is large but could be corrected in the way suggested by Tai (1971). The partition of sum of squares into two components is also only approximate, but this may be quite satisfactory for practical purposes. More efficient estimates of b_i's and ϵ_j's, and their sum of squares could be obtained by fitting the model in (6) by a non-linear least squares method as suggested by Elston (1961) and Tai (1971) and approximate tests could be obtained. Under these circumstances, independent measure of environment based on more genotype means may not be worthwhile.

3. COMPONENTS OF INTERACTION SUM OF SQUARES

The characterisation of genotypes on the basis of regression coefficients may not be very effective when only a small fraction of the interaction sum of squares ($G \times E$) can be attributed to heterogeneity among the regressions. It might be then of great interest to partition $G \times E$ into t components, one corresponding to each genotype, as mentioned by Baker (1969). Put

$$g_{ij} + \bar{e}_{ij} = \nu_{ij}.$$

Let us further assume that

$$E(\nu_{ij}) = 0;\ V(\nu_{ij}) = \sigma_i^2;\ E(\nu_{ij}, \nu_{i'j'}) = 0 \text{ for } i \neq i' \text{ or } j \neq j';$$
$$V(g_{ij}) = \sigma_i'^2;\ E(g_{ij}\bar{e}_{ij}) = 0;\ i = 1, \ldots, t. \tag{10}$$

Then,

$$\sigma_i^2 = \sigma_i'^2 + \sigma_0^2.$$

In the above expression σ_i^2 could be taken as the sum of two components, *viz.* within environmental variance (σ_0^2) and between environmental variance ($\sigma_i'^2$) of the ith genotype (after correcting for additive common effect of environment ϵ_j), and we shall name it the " stability variance " of the ith genotype. We shall call a genotype stable if its stability variance (σ_i^2) is equal to within environmental variance (σ_0^2) which means that $\sigma_i'^2 = 0$. Relatively large values of σ_i^2 will indicate more instability of genotype.

Estimation of σ_i^2 is analogous to the problem of estimating heterogeneous variances in a two-way classification when they change in one way considered by Ehrenberg (1950) and later by Russell and Bradley (1958). Rao (1970) has further generalised the above prcoedure for any classification and also considered some optimum properties of the above estimators. Without going into detail, we give the unbiased estimators of σ_i^2, denoted by $\hat{\sigma}_i^2$, as

$$\begin{aligned}\hat{\sigma}_i^2 &= \frac{1}{(s-1)(t-1)(t-2)}\,[t(t-1)\sum_j(\bar{y}_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2 \\ &\qquad\qquad - \sum_i\sum_j(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2] \\ &= \frac{1}{(s-1)(t-1)(t-2)}\,[t(t-1)\sum_j(u_{ij}-\bar{u}_{i.})^2-\sum_i\sum_j(u_{ij}-\bar{u}_{i.})^2] \qquad (11)\end{aligned}$$

where

$$u_{ij} = \bar{y}_{ij}-\bar{y}_{.j} \text{ and } \bar{u}_{i.} = \sum_j u_{ij}/s.$$

They are obtained as linear combinations of squares of residuals

$$(\bar{y}_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..}),$$

therefore, they are independent of μ, d_i and variance of ϵ_j. It is not difficult to verify that they are unbiased estimators of σ_i^2. Under the assumption of symmetrical distribution of σ_i^2's, Rao (1970) has proved that on average they have minimum variance among all possible quadratic unbiased estimators (MINQUE) of σ_i^2. It is not difficult to see that their mean is the same as the mean sum of squares ($G\times E$). Therefore, by multiplying each $\hat{\sigma}_i^2$ by $(t-1)(s-1)/t$ we shall obtain t components of $G\times E$, one corresponding to each genotype. These components are not statistically independent; as they are differences of two sums of squares, they can be negative, but negative estimates of variances are not uncommon in variance components problems.

The essential difference between the present approach and Baker's (1969) approach is this that his method estimates $((t-2)\sigma_i^2+\bar{\sigma}^2)/t$ while the above method estimates σ_i^2 where

$$\bar{\sigma}^2 = \frac{1}{t}\sum_{i=1}^{t}\sigma_i^2.$$

The same is true for deviation from regression component.

The variance of $\hat{\sigma}_i^2$ is not only a function of σ_i^2 but of variances of other genotypes $\sigma_j^2 (j \neq i)$ taken in trial. Such estimators are only available when $t \geqq 3$. The problem of testing homogeneity of σ_i^2's has been considered by Russell and Bradley (1958), Johnson (1962), Han (1969) and Shukla (1971). The method proposed by Shukla could be easily applied for testing the homogeneity of all the variances or any pair of them.

It might be of some interest to test whether certain genotypes are stable or not. Johnson (1962) suggested a test criterion based on the ratio $\hat{\sigma}_i^2/\hat{\sigma}_0^2$.

It is difficult to derive the exact distribution of $\hat{\sigma}_i^2$, but when t is large the variance of $\hat{\sigma}_i^2$ is approximated as

$$V(\hat{\sigma}_i^2) \cong 2\sigma_i^4/(s-1). \tag{12}$$

The above expression helps us in obtaining an approximate distribution. When t is large, $(s-1)\hat{\sigma}_i^2/\sigma_i^2$ will be approximately distributed as χ^2 on $(s-1)$ degrees of freedom; thus, under the hypothesis that $\sigma_i'^2 = 0$, F^* will have an approximate F distribution on $(s-1)$ and $st(r-1)$ degrees of freedom where

$$F^* = \hat{\sigma}_i^2/\hat{\sigma}_0^2. \tag{13}$$

When $\hat{\sigma}_i^2$ is negative or less than $\hat{\sigma}_0^2$ then $\sigma_i'^2$ may be taken equal to zero as usual.

4. FURTHER EXTENSION OF MODEL

For further progress in the interpretation of instability, we shall reconsider the model in (5). To keep the treatment general, we replace ϵ_j by z_j in the non-additive term $b_i'\epsilon_j$ and rewrite the model as

$$y_{ij} = \mu + d_i + \epsilon_j' + b_i z_j + \alpha_{ij}; \tag{14}$$

where z_j is a measure of some characteristic of jth environment; by taking deviation from the mean we can make $\sum_j z_j = 0$, and $\epsilon_j' = \epsilon_j + \bar{b}' z_j$.

We shall further assume that

$$V(\alpha_{ij}) = s_i^2; \quad (i = 1, \ldots, t),$$

and then discuss the method of estimation of s_i^2. The usual estimator of b_i, by the method of unweighted least squares, can be obtained as

$$\hat{b}_i = \sum_j \frac{(u_{ij} - \bar{u}_{i.})z_j}{\sum_j z_j^2}. \tag{15}$$

Using methods as in Section 3, unbiased estimators of s_i^2 for extended model in (14) could be obtained as $\hat{s}_i^2$:

$$\hat{s}_i^2 = \frac{t}{(t-2)(s-2)}\left[S_i - \sum_i \frac{S_i}{t(t-1)}\right] \tag{16}$$

where

$$S_i = \sum_{j=1}^{s} (u_{ij} - \bar{u}_{i.} - \hat{b}_i z_j)^2.$$

It is apparent that the model in (14) is just the extension of the model in (7) to take into account a covariate z_j. The estimators obtained in (16) are quadratic (in y's) estimators of s_i^2 and have the properties of MINQUE estimators. When t is large, the variance of $\hat{s}_i^2$ can be approximated by

$$V(\hat{s}_i^2) = \frac{2s_i^4}{(s-2)} \tag{17}$$

and their distribution can be approximated by χ^2 on $(s-2)$ degrees of freedom. An approximate significance test against $\hat{\sigma}_0^2$ is possible as in (13). If some of the genotypes become stable after taking the covariate into account, it may be inferred that the instability was introduced by the linear effect of the covariate and such information may be useful. The above approach could also be extended to more than one covariate.

5. RELATIONSHIP BETWEEN REGRESSION APPROACH AND THE " STABILITY VARIANCE " APPROACH

The definition of stability is similar to Baker (1969) and Eberhart and Russell (1966). A significant departure of the regression of a genotype from zero will be indicated by a relatively high " stability variance ", but a regression coefficient of zero need not mean that the particular genotype is stable. A zero regression will be obtained if there is no linear relationship between genotype mean and environmental mean, yet the " stability variance " (σ_i^2) may be greater than σ_0^2.

Once some of the genotypes are found unstable, it may be of interest to examine further the reasons for instability. The approach of Section 4 may be followed if observations are available on covariates which are likely to affect the genotypes differentially. We can examine the effect on stability variance of linear regression on environmental means by the method in the previous section. To examine any effect of differential fertility we have used $z_j = \bar{y}_{.j} - \bar{y}_{..}$ as used by many other authors mentioned in Sections 1 and 2. It must be noted here that the estimators of s_i^2 obtained by putting $z_j = \bar{y}_{.j} - \bar{y}_{..}$ in (16) will not be quadratic estimators of y's and therefore the optimum properties described in Section 4 may not hold. Again as mentioned earlier the effect of departure from optimality may be small when σ_ε^2 is large. The effect of such differential regression on the stability of genotypes could be tested as above under the assumption that z_j's are constant. The estimation of individual s_i^2 is analogous to what Perkins and Jinks (1968) have suggested by the mean sum of squares $\sum_j \delta_{ij}^2/(s-2)$ and Baker (1969) by deviation from regression sum of squares but the above approach has an advantage as they are unbiased estimates of s_i^2 (free from any other nuisance parameters) and the mean of $\hat{s}_i^2$ is the same as the mean sum of squares of departure from regressions (" Balance ") and this could be taken as equivalent to dividing the " Balance " into components corresponding to each genotype.

Recently Tai (1971) has worked with the above problem. The difference between our method and his method is this, that he has considered the model under certain side conditions on the interaction and we have not imposed any such conditions on them. It would not be very justifiable to impose any condition on interaction while estimating the individual component. The definition of stability is also different. According to his definition of stability one should have $b_i' = -1$ and $s_i^2 = \sigma_0^2$. Our definition of stability coincides with his definition of average stability ($\alpha_i = 0$; $\lambda_i = 1$ in Tai, 1971, notations). By our definition of stability we only mean that the performance of a genotype is sum of additive genotypic effect, additive environmental effect and a random error without any interaction between genotype and environment.

Prediction of the expected performance can also be made with reasonable accuracy for a given environment, if either the interaction is not present or most of it can be accounted by linear regression term (Jinks and Perkins, 1970).

6. Numerical example

For illustration purposes we have considered the data analysed by Yates and Cochran (1938). We shall only consider the part of the table dealing with $G \times E$.

Table 1

Variety × Place totals over the years

	Places						
Varieties	1	2	3	4	5	6	Total
Manchusia	161·7	247·0	185·4	218·7	165·3	154·6	1132·7
Svansota	187·7	257·5	182·4	183·3	138·9	143·8	1093·6
Velvet	200·1	262·9	194·9	220·2	165·8	146·3	1190·2
Tribi	196·9	339·2	271·2	266·3	151·2	193·6	1418·4
Peatland	182·5	253·8	219·2	200·5	184·4	190·1	1230·5
Total	928·9	1360·4	1053·1	1089·0	805·6	828·4	6065·4

Table 2 gives the values of u_{ij} obtained from table 1.

Table 2

u_{ij}*'s and regression coefficients*

	Places						
Varieties	1	2	3	4	5	6	$\hat{b}_i$
Manchusia	−24·08	−25·08	−25·22	0·90	4·18	−11·08	−0·156
Svansota	1·92	−14·58	−28·22	−34·50	−22·22	−21·88	−0·014
Velvet	14·32	−9·18	−15·72	2·40	4·68	−19·38	−0·054
Tribi	11·12	67·12	60·58	48·50	−9·92	27·92	0·609
Peatland	−3·28	−18·28	8·58	−17·30	23·28	24·42	−0·385

Table 3

$u_{ij}-\hat{b}_i z_j$

	Places					
Varieties	1	2	3	4	5	6
Manchusia	−26·14	−14·18	−23·90	3·34	−2·23	−16·77
Svansota	1·69	−13·60	−28·10	−34·28	−22·79	−22·39
Velvet	13·43	−5·41	−15·26	3·24	2·46	−21·35
Tribi	21·11	24·55	55·44	38·99	15·08	50·17
Peatland	−9·59	8·63	11·83	−11·29	7·47	10·37

To obtain the component of variances on the same unit as the sum of squares (units of single plot) in the Analysis of Variance (table 4) we have divided them by 6.

Comparison of $\hat{\sigma}_i^2$'s with $\hat{\sigma}_0^2$ shows that Tribi and Peatland are unstable. Further regression analysis shows that Tribi remains unstable, though, its variability has reduced considerably. Peatland becomes stable after taking covariate into consideration. Similar conclusions were drawn by Yates and

Cochran (1938) but the above type of analysis in general may be advantageous.

TABLE 4

Analysis of variance table and " stability variances " (units of single plot)

Source	D.F.	S.S.	M.S.	F
Places	5	7072·92	—	—
Varieties	4	1770·28	—	—
$V \times P$	20	1477·84	73·89	—
$\hat{\sigma}_1^2$	—	—	25·88	1·11
$\hat{\sigma}_2^2$	—	—	19·60	0·84
$\hat{\sigma}_3^2$	—	—	22·73	0·98
$\hat{\sigma}_4^2$	—	—	225·53	9·69**
$\hat{\sigma}_5^2$	—	—	75·68	3·25**
Source	D.F.	S.S.	M.S.	F
Heterogeneity	4	773·16	193·29	—
Balance	16	704·69	44·04	—
$\hat{s}_1^2$	—	—	34·10	1·46
$\hat{s}_2^2$	—	—	40·48	1·74
$\hat{s}_3^2$	—	—	42·78	1·84
$\hat{s}_4^2$	—	—	79·70	3·42*
$\hat{s}_5^2$	—	—	23·27	1·00
$\hat{\sigma}_0^2$	216	—	23·28	—

7. SUMMARY

1. The usual regression approach of explaining genotype-environment interaction has been considered by using a non-additive model and the statistical validity of the analysis has been discussed.

2. Alternative approach of dividing genotype-environmental interaction into components, one corresponding to each genotype has been proposed and the optimum properties have been discussed.

3. The alternative approach has been extended to take into account a covariate.

4. The relationship of new approach to the regression approach has been discussed.

5. A numerical example has been given as an illustration.

Acknowledgment.—I am very grateful to Professor D. J. Finney and Mr H. D. Patterson for their valuable suggestions and help. I am grateful to referees for their valuable suggestions.

8. REFERENCES

BAKER, R. J. 1969. Genotype-environment interaction in yield of wheat. *Can. J. Plant Sci.*, *49*, 743-751.

EBERHART, S. A., AND RUSSELL, W. L. 1966. Stability parameters for comparing varieties. *Crop Sci.*, *6*, 36-40.

EHRENBERG, A. S. C. 1950. The unbiased estimation of heterogeneous error variances. *Biometrika*, *37*, 347-357.

ELSTON, R. C. 1961. On additivity in the analysis of variance. *Biometrics*, *17*, 209-219.

FINLAY, K. W., AND WILKINSON, G. N. 1963. The analysis of adaption in a plant breeding programme. *Aust. J. Agric. Res.*, *14*, 742-754.

FREEMAN, G. H., AND PERKINS, JEAN M. 1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measure of these environments. *Heredity*, *26*, 15-23.

JINKS, J. L., AND PERKINS, JEAN, M. 1970. Environmental and genotype-environmental components of variability. VII. Simultaneous prediction across environments and generations. *Heredity*, *25*, 475-480.

HAN, C. P. 1969. Testing the homogeneity of variances in a two-way classification. *Biometrics*, *25*, 153-158.

JOHNSON, N. L. 1962. Some notes on the investigation of heterogeneity in interaction. *Trabajos de Estadistics*, *XIII*, 183-199.

MANDEL, J. 1961. Non-additivity in two-way analysis of variance. *J. Am. Stat. Assoc.*, *56*, 878-888.

MILLIKEN, G. A., AND GRAYBILL, F. A. 1970. Extension of general linear hypothesis. *J. Am. Stat. Assoc.*, *65*, 797-807.

PERKINS, JEAN M., AND JINKS, J. L. 1968. Environmental and genotype environmental components of variability. III. Multiple lines and crosses. *Heredity*, *23*, 239-256.

RAO, C. R. 1970. Estimation of heteroscedastic variances in linear models. *J. Am. Stat. Assoc.*, *65*, 161-172.

RUSSELL, T. S., AND BRADLEY, R. A. 1958. One-way variances in a two-way classification. *Biometrika*, *45*, 111-129.

SHUKLA, G. K. 1971. An invariant test for the homogeneity of variances in a two-way classification. To appear in *Biometrics*.

TAI, G. C. C. 1971. Genotype stability analysis and its application to potato regional trials. *Crop Science*, *11*, 184-190.

TUKEY, J. W. 1949. One degree of freedom of non-additivity. *Biometrics*, *5*, 232-242.

YATES, F., AND COCHRAN, W. G. 1938. The analysis of group experiments. *J. Agric. Sci.*, *28*, 556-580.

PRINTED IN GREAT BRITAIN BY T. & A. CONSTABLE LTD., EDINBURGH

Reprinted from
BIOMETRICS *28*, 1063–1072
December 1972

AN INVARIANT TEST FOR THE HOMOGENEITY OF VARIANCES IN A TWO-WAY CLASSIFICATION

G. K. SHUKLA

ARC Unit of Statistics, University of Edinburgh, 21 Buccleuch Place, Edinburgh EH8 9LN, Scotland

SUMMARY

A test is proposed of the equality of variances in different arrays (rows or columns) of a two-way balanced crossed classification without interaction. This test differs from another test recently proposed by Han [1969] in that it is invariant to changes in the nuisance parameters representing the effect of arrays. The paper also generalizes the likelihood ratio test introduced by Russell and Bradley [1958] for classifications with only three columns.

1. INTRODUCTION

In this paper a new test is proposed for the equality of variances in different arrays of a balanced two-way classification. We suppose the problem arises as follows. The observations y_{ij} are arranged in n rows and t columns ($i = 1, 2, \cdots, n; j = 1, 2, \cdots, t$). An appropriate model relating the y_{ij} to unknown parameters μ, α_i, β_j is considered to be

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \tag{1}$$

where e_{ij} is a random error with zero expectation and

$$V(e_{ij}) = \sigma_j^2 \, ; E(e_{ij}e_{i'j'}) = 0 \quad \text{for} \quad i \neq i' \quad \text{or} \quad j \neq j'.$$

In many applications $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_t^2$ is known to be fulfilled but in some applications

$$H_A : \sigma_j^2 \neq \sigma_{j'}^2$$

for at least one pair of j, j' ($j \neq j'$) may be more appropriate. We wish to test H_0 against H_A.

In this paper we shall proceed with the test of equality of column variances, but the same method can be used to test the equality of row variances (assuming that they are constant along the columns) with an appropriate change in the notation. Once these variances are shown to be different we may wish to test whether two particular column variances differ i.e. to test the hypothesis $H_0^* : \sigma_j^2 = \sigma_{j'}^2$.

An application of this is in the combination of results of a series of variety trials. The n rows represent n separate trials, the t columns t varieties and the y_{ij} are means in individual trials.

Morgan [1939] and Pitman [1939] considered the problem of testing the equality of variances in a bivariate normal population. Their results have been applied by Maloney and Rastogi [1970] to the comparison of variances in a two-way classification with only two columns. The results of Morgan [1939] and Pitman [1939] have also been applied by Curnow [1957] to test the equality of variances of sub-plot treatments in a split-plot design when there are only two sub-plot treatments. Russell and Bradley [1958] developed a likelihood ratio test of H_0 for the case $t = 3$. Johnson [1962] described a test of H_0^* for any $t > 2$.

Han [1968] proposed several tests of equality of variances in a one-way classification. One of these tests is a generalization of Morgan [1939] and Pitman [1939]. Later Han [1969] extended one of his tests to two-way classifications. His test criterion is the multiple correlation, R say, of $\bar{y}_{i.}$ on $y_{ij} - \bar{y}_{i.}$ $(j = 2, 3, \cdots, t)$, where $\bar{y}_{i.}$ is the marginal mean in row i. Under the null hypothesis H_0 the multiple correlation is zero. Han [1969] therefore suggests that a significant value of R indicates departure from H_0. A drawback of this test is that the sample multiple correlation depends on the value of the α_i. This can be shown by adding a constant μ_i to all entries in row i. The variance-covariance matrix of $y_{ij} - \bar{y}_{i.}$ is unaffected but the value of the multiple correlation coefficient is changed. When row means are very variable the correlation tends to be small and hence the power of the test is small. McIntyre [1972] has very effectively demonstrated this point by means of a numerical example. In the present paper a test has been proposed which is invariant for changes in α_i and β_j. We consider only the case when all observations are independent ($\rho = 0$ in Han's [1969] notation). The number of columns must be greater than 2 and n should be moderately large. In section 2 we have described the test statistic and section 3 deals with the distribution of the test statistic. Section 4 gives the test for comparing any pair of variances and section 5 considers the likelihood ratio test. A worked example is in section 6.

2. METHOD

It is convenient to write the model (1) in the form

$$Y_i = 1\mu + 1\alpha_i + \beta + e_i , \tag{2}$$

where

$$Y_i' = (y_{i1}, y_{i2}, \cdots, y_{it}), \quad \beta' = (\beta_1, \beta_2, \cdots, \beta_t),$$

$e_i' = (e_{i1}, e_{i2}, \cdots, e_{it})$ and 1 is a vector of ones. Now define an $(m - 1) \times m$ matrix L_m such that

$$L_m 1 = 0, \quad L_m L_m' = I. \tag{3}$$

Now a vector Z_i of $m - 1$ orthogonal contrasts among the observations in row i can be obtained as follows.

$$Z_i = L_t Y_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{i(t-1)} \end{bmatrix}; \qquad (i = 1, \cdots, n). \tag{4}$$

It is easy to show that

$$E(Z_i) = L_t \beta; \quad V(Z_i) = L_t \Sigma L_t' = V, \text{ say,} \tag{5}$$

where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_t^2 \end{bmatrix}. \tag{6}$$

Under the null hypothesis $H_0 : \sigma_j^2 = \sigma^2$, say, for all j we have

$$V(Z_i) = \sigma^2 I \tag{7}$$

We test H_0 by (i) estimating V by S defined below and (ii) applying the test of sphericity described by Mauchly [1940]. The matrix S has elements $s_{kk'}$ given by

$$s_{kk'} = \frac{1}{n-1} \sum_{i=1}^{n} (z_{ik} - \bar{z}_{.k})(z_{ik'} - \bar{z}_{.k'}); \qquad (k, k' = 1, 2, \cdots, t-1) \tag{8}$$

The test statistic is λ defined by

$$\lambda = \frac{|S|^{n/2}}{\left\{\frac{\operatorname{tr}(S)}{p}\right\}^{np/2}}, \tag{9}$$

where $p = t - 1$.

Under the null hypothesis of equality of variances the value of λ should be 1 apart from sampling errors.

It is not difficult to show that the statistic λ is invariant for any chosen set of orthogonal contrasts. For, consider any other set of orthogonal contrasts $M_t Y_i$ where M_t is given by

$$M_t = D L_t ; \tag{10}$$

and D is an orthogonal matrix. Hence

$$DD' = I \quad \text{and} \quad M_t M_t' = I.$$

Now let λ_1 be the value of λ when M_t is substituted for L_t in equation (9). We obtain

$$\lambda_1^{2/n} = \frac{|DSD'|}{\left(\frac{\operatorname{tr}(DSD')}{p}\right)^p} = \frac{|S|}{\left(\frac{\operatorname{tr}(S)}{p}\right)^p} = \lambda^{2/n}. \tag{11}$$

Hence λ is independent of the choice of orthogonal contrasts. Also λ is obviously not affected by any changes in μ, α_i's and β_j's.

3. DISTRIBUTION OF λ

Moments of the statistic λ under the null hypothesis are given by Anderson [1958] pp. 262. For simplicity of notation we take $N = n - 1$ and $W = \lambda^{2/n}$. The hth moment of W is given by

$$E(W^h) = \frac{p^{hp}\Gamma\left(\frac{pN}{2}\right)}{\Gamma\left(\frac{pN}{2} + ph\right)} \prod_{i=1}^{p} \frac{\Gamma(\frac{1}{2}(N + 1 - i) + h)}{\Gamma(\frac{1}{2}(N + 1 - i))}. \tag{12}$$

Box [1949] has discussed a general method for approximating the distribution of a likelihood ratio criterion by the first few terms in an infinite series of χ^2's. Anderson [1958] has applied this method to the sphericity test criterion. Using the first two terms he obtained the following theorem *Theorem*: Under the null hypothesis of equality of all variances $-2\rho \log_e \lambda$ is approximately distributed as χ^2 with $[t(t - 1)]/2 - 1$ degrees of freedom (D.F.), where ρ is given by

$$\rho = 1 - \frac{2p^2 + p + 2}{6p(n - 1)}.$$

This result is adequate for large n. A more accurate result (correct up to order n^{-2}) can be obtained by including three terms of the infinite series. We then have

$$\Pr\{-2\rho \log_e \lambda \le \eta\} = \Pr\{\chi_f^2 \le \eta\} + \omega_2\{\Pr(\chi_{f+4}^2 \le \eta) - \Pr(\chi_f^2 \le \eta)\} + O(n^{-3}), \tag{13}$$

where

$$\omega_2 = \frac{(p + 2)(p - 1)(p - 2)(2p^3 + 6p^2 + 3p + 2)}{288p^2N^2\rho^2}, \quad \text{and}$$

$$f = \frac{t(t - 1)}{2} - 1.$$

Better approximations can also be obtained by using the results of Hill and Davis [1968]. Mathai and Rathie [1971] also deal with the exact distribution of the likelihood ratio in a related problem and their results can be extended to give the exact distribution of λ considered here.

4. TEST OF THE EQUALITY OF TWO COLUMN VARIANCES

Suppose that H_0 has been shown to be unacceptable. We may now wish to compare a particular pair of column variances, for example σ_1^2 and σ_2^2. The null hypothesis is as follows:

$$H_0^*: \sigma_1^2 = \sigma_2^2.$$

To test the hypothesis we first define the contrasts

$$z_{i1} = (y_{i1} - y_{i2})/\sqrt{2},$$
$$z_{i2} = (y_{i1} + y_{i2} - 2\bar{y}_{i(t-2)})/\sqrt{6},$$

where

$$\bar{y}_{i(t-2)} = \sum_{j=3}^{t} y_{ij}/(t-2).$$

The expected values of z_{i1} and z_{i2} are given by

$$E(z_{i1}) = (\beta_1 - \beta_2)/\sqrt{2},$$
$$E(z_{i2}) = (\beta_1 + \beta_2 - 2\bar{\beta}_{t-2})/\sqrt{6},$$

and the variances and covariances are given by

$$V(z_{i1}) = (\sigma_1^2 + \sigma_2^2)/2$$
$$V(z_{i2}) = \left(\sigma_1^2 + \sigma_2^2 + 4\,\frac{\bar{\sigma}_{t-2}^2}{t-2}\right)\Big/ 6,$$
$$\operatorname{Cov}(z_{i1}, z_{i2}) = (\sigma_1^2 - \sigma_2^2)/\sqrt{12}$$

where

$$\bar{\sigma}_{t-2}^2 = \sum_{j=3}^{t} \sigma_j^2/(t-2) \quad \text{and} \quad \bar{\beta}_{t-2} = \sum_{j=3}^{t} \beta_j/(t-2).$$

Under the null hypothesis H_0^* correlation between z_{i1} and z_{i2} is zero. Hence we can test H_0^* by comparing

$$\frac{r^*}{\sqrt{(1 - r^{*2})}}\,(n-2)^{1/2} \tag{14}$$

with the t distribution ($n - 2$ D.F.), where r^* is the sample correlation coefficient between z_{i1} and z_{i2}.

Note, however, that $V(z_{i2})$ and hence the power of the test depends on $\bar{\sigma}_{t-2}^2$, i.e. on the variances in columns other than the two under consideration. Some improvement in the test may be possible if we replace $\bar{y}_{i(t-2)}$ by another mean with a smaller variance. For example, if we know a priori that $\sigma_{j'}^2$ is large we might replace $\bar{y}_{i(t-2)}$ by the mean for columns 3, 4, $\cdots$, t excluding column j'. The power of the test is, however, independent of μ, α_i, β_j. In cases when α_i's are very variable the above test may provide a more powerful test than the one suggested by Maloney and Rastogi [1970].

Johnson [1962] suggested a test of H_0^* based on the ratio

$$F = \frac{\sum_{i=1}^{n} (y_{i1} - \bar{y}_{i.} - \bar{y}_{.1} + \bar{y}_{..})^2}{\sum_{i=1}^{n} (y_{i2} - \bar{y}_{i.} - \bar{y}_{.2} + \bar{y}_{..})^2}.$$

He pointed out that both numerator and denominator of this ratio depend on the variances in column 3, 4, $\cdots$, t and the power may be small when variances in remaining columns ($t = 3, \cdots , t$) are much larger than the variances of the first two columns under test. The test proposed here is exact, simple to construct, and may provide a more powerful test in some cases.

5. LIKELIHOOD RATIO TEST

Russell and Bradley [1958] derived the likelihood ratio (LR) test of equality of variances for the particular case of $t = 3$. When $t > 3$ explicit expressions for the maximum likelihood (ML) estimators of variances are no longer available but the equations of estimation can be solved by a process of successive approximation. However, for $t > 3$ it may not be easy to obtain the iterative solution, as the estimate of the matrix V may not always be positive definite.

In this section we outline the procedure in the general case. Following Russell and Bradley [1958] we first eliminate the nuisance parameters μ, α_i , β_i by choosing suitable contrasts between observations in rows and columns. We then obtain estimates of σ_i^2 by maximizing the joint likelihood of the selected contrasts.

A suitable set of orthogonal row $\times$ column contrasts x_{km} , $k = 1, \cdots , n - 1$; $m = 1, \cdots , t - 1$ is given by the elements of an $(n - 1) \times (t - 1)$ matrix X, where

$$X = L_n Y L_t' = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_{n-1}' \end{bmatrix}.$$

It can be shown that

$$E(X) = 0; \quad V(X_k) = V; \quad \operatorname{cov}(X_k , X_{k'}) = 0 \quad \text{for} \quad (k \neq k'),$$

where V is defined by equation (5).

The joint distribution of x_{km} is given by

$$f(x) = (2\pi)^{-(n-1)(t-1)/2} \, |V^{-1}|^{(n-1)/2} \exp - \frac{1}{2} \sum_{k=1}^{n-1} X_k' V^{-1} X_k. \qquad (15)$$

Let $\hat{\sigma}_i^2$ be the estimate of σ_i^2 maximizing $\log f(x)$. Also let $L(\Omega)$, $\hat{V}$ be the expression for $\log f(x)$ and V obtained by substituting $\hat{\sigma}_i^2$ for σ_i^2 .

Under the null hypothesis H_0 the ML estimator of the common variance σ^2 is simply $\hat{\sigma}^2$, the residual mean square in the conventional analysis of variance for an orthogonal two-way classification. Let $L(\omega)$ be the log likelihood under the null hypothesis and let λ' be the likelihood ratio such that

$$\log_e \lambda' = L(\omega) - L(\Omega).$$

Then

$$\log_e \lambda' = \frac{(n-1)(t-1)}{2}\left[-\log \hat{\sigma}^2 - \frac{1}{(t-1)} \log |\hat{V}^{-1}| + \frac{1}{(t-1)(n-1)} \sum_{k=1}^{n-1} X_k' \hat{V}^{-1} X_k - 1\right]. \qquad (16)$$

The quantity $-2 \log_e \lambda'$ is distributed asymptotically as χ^2 with $(t-1)$ D.F. under the null hypothesis.

It is interesting to observe that in the particular case of $t = 3$ the statistics λ of equation (9) and λ' of equation (16) are the same for large n, i.e. the sphericity test and the LR test are asymptotically equivalent for $t = 3$. This can be shown by taking

$$L_3 = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{pmatrix}$$

in equation (4).

Russell and Bradley (1958) show that

$$-2 \log_e \lambda' = -(n-1)[\log_e (\hat{\sigma}_1^2 \hat{\sigma}_2^2 + \hat{\sigma}_1^2 \hat{\sigma}_3^2 + \hat{\sigma}_2^2 \hat{\sigma}_3^2) - 2 \log_e \hat{\sigma}^2 - \log_e 3] \qquad (17)$$

where $\hat{\sigma}_j^2$ is given by

$$\hat{\sigma}_j^2 = \frac{1}{2(n-1)}\left[6 \sum_{i=1}^{n} (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 - \sum_{i=1}^{n} \sum_{j=1}^{3} (y_{ij} - \bar{y}_{i.} - y_{.j} + \bar{y}_{..})^2\right]. \qquad (18)$$

The $\hat{\sigma}_j^2$ can also be expressed in terms of the elements of S defined by equation (8). The expressions are as follows

$$\hat{\sigma}_1^2 = s_{11} + \sqrt{3}\, s_{12}\,,$$

$$\hat{\sigma}_2^2 = s_{11} - \sqrt{3}\, s_{12}\,,$$

$$\hat{\sigma}_3^2 = (3s_{22} - s_{11})/2.$$

Substituting in equation (17) we have

$$-2 \log_e \lambda' = -(n-1)[\log_e (s_{11}s_{22} - s_{12}^2) - 2 \log_e \hat{\sigma}^2]. \qquad (19)$$

This is the same as $-2 \log_e \lambda$ obtained from equation (9),

$$-2 \log_e \lambda = -n[\log_e (s_{11}s_{22} - s_{12}^2) - 2 \log_e \hat{\sigma}^2], \qquad (20)$$

apart from the multiplication factor of n in place of $n - 1$. The discrepancy is unimportant for large n. However, for $t = 3$ the results in section 3 give better approximation to χ^2 for moderate values of n.

It is not clear how the test proposed above and the likelihood ratio test compare in general. Looking at the degrees of freedom for the two tests it appears that the LR test may always give a more powerful test for $t > 3$ but in general the value of $-2 \log_e \lambda$ will be larger than $-2 \log_e \lambda'$. Moreover, the test proposed in section 2 is much easier to obtain and gives a better approximation to χ^2 even for moderate values of n. It must be noted here that the sphericity test, as described by Mauchly [1940], was obtained by the likelihood ratio criterion to test the equality of all variances, and equality of covariances to zero, in a multivariate normal distribution. In our case, under the alternative hypothesis H_A, the transformed variates z_{ij}, with reduced dimensions, will not have the sphericity property, under the assumptions that the errors e_{ij} are independent, and so the above test can be used in this situation.

6. NUMERICAL EXAMPLE

In this section we illustrate the methods of the present analyses of the variety trial data set out in Table 1. These data were presented by Graybill [1954] and have also been used by Han [1969].

The problem is to determine whether or not the four varieties are associated with different error variances.

First we describe the calculations leading to the sphericity test. We take the matrix L_4 as

$$L_4 = \frac{1}{2}\begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{bmatrix}.$$

As we have shown the statistic λ is independent of the matrix chosen. We could use other matrices of the form L_t, e.g. a matrix of orthogonal polynomial coefficients. The sample variance-covariance matrix S is as follows.

$$S = \begin{bmatrix} 41.090 & 9.552 & 54.261 \\ 9.552 & 15.212 & 25.752 \\ 54.261 & 25.752 & 112.067 \end{bmatrix}.$$

The following quantities arise in the calculations:

$$|S| = 14770.289$$

$$\hat{\sigma}^2 = \text{tr}\,(S)/3 = 56.303$$

$$-\cancel{2}n \log_e \lambda^{2/n} = 32.39$$

$$\rho = 0.893$$

$$-\cancel{2}n\rho \log_e \lambda^{2/n} = 28.92$$

$$\omega_2 = 0.0039.$$

TABLE 1
YIELDS OF WHEAT (CWT/ACRE)

	Variety			
Location	1	2	3	4
1	43.60	24.05	19.47	19.41
2	40.40	21.76	16.61	23.84
3	18.08	14.19	16.69	16.08
4	19.57	18.61	17.78	18.29
5	45.20	29.33	20.19	30.08
6	25.87	25.60	23.31	27.04
7	55.20	38.77	21.15	39.95
8	55.32	34.19	18.56	25.12
9	19.79	21.65	23.31	22.45
10	46.24	31.52	22.48	29.28
11	14.88	15.68	19.79	22.56
12	7.52	4.69	20.53	22.08
13	41.17	32.59	29.25	43.95

As the value of ω_2 is small in this case the approximation up to n^{-1} will be adequate. The value of $-\cancel{2}n\rho \log_e \lambda^{2/n}$ is highly significant compared with the tabulated χ^2 with 5 D.F. for the 5% probability level.

We now consider the LR test.

The following ML estimates were obtained on a computer using an iteration technique:

$$\hat{\sigma}_1^2 = 70.843; \qquad \hat{\sigma}_2^2 = -16.373; \qquad \hat{\sigma}_3^2 = 102.282; \qquad \hat{\sigma}_4^2 = 77.016.$$

Hence using equation (20)

$$-2 \log_e \lambda' = -2[L(\omega) - L(\Omega)] = 14.048.$$

Under the null hypothesis this should have asymptotically a χ^2 distribution with 3 D.F. Again the null hypothesis must be rejected.

ACKNOWLEDGMENT

I wish to thank Mr. H. D. Patterson for going through the manuscript and pointing out certain irregularities. I am also thankful to Dr. D. N. Lawley for making various suggestions and to the referees for their helpful comments, and for bringing to my notice some interesting references.

UN TEST INVARIANT POUR L'HOMOGENEITE DES VARIANCES DANS UNE CLASSIFICATION A DEUX VOIES

RESUME

On propose un test d'égalité des variances dans différents tableaux (lignes ou colonne) d'une classification croisée, équilibrée à deux voies sans interaction. Ce test diffère d'un autre test récemment proposé par Han [1969] en ce qu'il est invariant pour des changements des paramètres gênants représentant l'effet des tableaux. L'article généralise aussi le test du rapport de vraisemblance introduit par Russell et Bradley [1958] pour des classification avec seulement trois colonnes.

REFERENCES

Anderson, T. W. [1958]. *An Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Box. G. E. P. [1949]. A general distribution theory for a class of likelihood criteria. *Biometrika 36*, 317–46.

Curnow, R. N. [1957]. Heterogeneous error variances in Split-plot Experiments. *Biometrika 44*, 378–83.

Graybill, F. [1954]. Variance heterogeneity in a randomised block design. *Biometrics 10*, 516–20.

Han, C. P. [1968]. Testing the homogeneity of a set of correlated variance. *Biometrika 55*, 317–26.

Han, C. P. [1969]. Testing the homogeneity of variances in a two-way classification. *Biometrics 25*, 153–8.

Hill, G. W. and Davis, A. W. [1968]. Generalized asymptotic expansions of Cornish-Fisher type. *Ann. Math. Statist. 39*, 1264–73.

Johnson, N. L. [1962]. Some notes on the investigation of heterogeneity in interactions. *Revista Trabajos De Estadistica*, XIII, 183–99.

McIntyre, G. A. [1972]. Homogeneity of variance in a two-way classification. Letter. *Biometrics 28*, 246.

Maloney, C. J. and Rastogi, S. C. [1970]. Significance test for Grubb's estimators. *Biometrics 26*, 671–6.

Mathai, A. M. and Rathie, P. N. [1971]. The problem of testing independence. *38th Session ISI*. 254–58.

Mauchly, J. W. [1940]. Significance test for sphericity of a normal n-variate distribution. *Ann. Math. Statist. 11*, 204–9.

Morgan, W. A. [1939]. A test of the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–9.

Pitman, E. J. G. [1939]. A note on normal correlation. *Biometrika 31*, 9–12.

Russell, T. S. and Bradley, R. A. [1958]. One-way variance in a two-way classification. *Biometrika 45*, 111–29.

Received March 1971, *Revised May* 1972

Key Words: Homogeneity of variances; Combination of results; Sphericity test; Comparison of pairs of variances.

Reprinted from
BIOMETRICS *29*, 373–377
June 1973

SOME EXACT TESTS OF HYPOTHESES ABOUT GRUBBS'S ESTIMATORS

G. K. SHUKLA

*ARC Unit of Statistics, University of Edinburgh, 21 Buccleuch Place,
Edinburgh EH8 9LN, Scotland*

SUMMARY

When two instruments or techniques are used to measure the same item, the measurement precisions may be estimated using a method proposed by Grubbs [1948]. The present paper generalizes certain tests given by earlier authors. Using this general result, some exact tests of certain hypotheses are obtained for which only approximate results were available. The results can also be used to construct exact confidence intervals for the relative precision of two instruments.

1. INTRODUCTION

Recently, Maloney and Rastogi [1970] and Jaech [1971] considered some tests of hypotheses about Grubbs's [1948] estimators for comparing the precision of two instruments. They used the methods of Morgan [1939] and Pitman [1939]. Jaech [1971] has suggested some approximate tests based on the likelihood ratio criterion; one of which tests a hypothesis about the relative precision of the two instruments. The results obtained here generalize those of Pitman [1939]. Curnow [1957] has also considered a similar problem in a different context. In section 2 we restate the notation and define the parameters consistent with that of Maloney and Rastogi [1970]. Section 3 gives a general result and some particular cases of interest are considered. Section 4 shows that the test statistic proposed in section 3 is the likelihood ratio (LR) statistic, thus giving the exact distribution of the LR criterion. In section 5, these results have been used to derive exact confidence limits for the relative precision.

2. FORMULAE AND NOTATION

Measurements from two instruments on the ith item (drawn randomly from an infinite population of items) are denoted by x_i and y_i ($i = 1, 2, \cdots, n$),

$$x_i = \tau_i + \xi_i\,; \quad y_i = \tau_i + \eta_i\,; \quad i = 1, 2, \cdots, n; \tag{1}$$

where τ_i is the correct unknown value of the ith item and ξ_i and η_i are the measurement errors and are assumed to be independent. Let us assume that

$$V(\tau_i) = \sigma^2; \quad V(\xi_i) = \sigma_1^2; \quad V(\eta_i) = \sigma_2^2\,; \tag{2}$$

then,

$$V(x_i) = \sigma^2 + \sigma_1^2 ; \quad V(y_i) = \sigma^2 + \sigma_2^2 ; \quad \text{Cov}\,(x_i , y_i) = \sigma^2. \tag{3}$$

Denote the sample estimates of the variances and covariance of x and y by

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n - 1); \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n - 1)$$
$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/(n - 1). \tag{4}$$

It is easy to see that

$$E(S_{xx}) = \sigma^2 + \sigma_1^2 ; \quad E(S_{yy}) = \sigma^2 + \sigma_2^2 ; \quad E(S_{xy}) = \sigma^2. \tag{5}$$

Grubbs [1948] estimated σ_1^2 , σ_2^2 , and σ^2, respectively by

$$\hat{\sigma}_1^2 = S_{xx} - S_{xy} ; \quad \hat{\sigma}_2^2 = S_{yy} - S_{xy} ; \quad \hat{\sigma}^2 = S_{xy} . \tag{6}$$

He also showed that they are maximum likelihood (ML) estimators for normally distributed variates.

3. TEST OF HYPOTHESIS

The present work is concerned only with the null hypothesis

$$H_0 : \sigma_1^2 = k\sigma_2^2 \tag{7}$$

where k is a specified constant ($k \geq 0$). This will be tested against

$$H_A : \sigma_1^2 \neq k\sigma_2^2 . \tag{8}$$

Jaech [1971] suggested the likelihood ratio test for testing the above hypothesis, and obtained a result equivalent to that given here. Consider an alternative simpler approach. Write

$$u_i = x_i + ky_i ; \quad v_i = x_i - y_i ; \quad (i = 1, 2, \cdots , n). \tag{9}$$

The variances and covariance of u and v are given by

$$V(u) = \sigma_1^2 + k^2\sigma_2^2 + \sigma^2(1 + k)^2; \quad V(v) = \sigma_1^2 + \sigma_2^2 ; \quad \text{Cov}\,(u, v) = \sigma_1^2 - k\sigma_2^2 . \tag{10}$$

Denote the population correlation between u and v by ρ; then

$$\rho = \frac{\sigma_1^2 - k\sigma_2^2}{[\{\sigma_1^2 + \sigma_2^2\}\{\sigma_1^2 + k^2\sigma_2^2 + \sigma^2(1 + k)^2\}]^{1/2}}. \tag{11}$$

A necessary and sufficient condition for ρ to be zero is that $\sigma_1^2 = k\sigma_2^2$, which suggests that the test of H_0 is equivalent to a test of u and v having zero correlation. Denote the sample correlation coefficient between u and v by r; this can be expressed in terms of S_{xx} , S_{yy} , and S_{xy} as follows. Write the sample variances and covariance of u and v as

$$S_{uu} = S_{xx} + k^2S_{yy} + 2kS_{xy} ; \quad S_{vv} = S_{xx} + S_{yy} - 2S_{xy} ;$$
$$S_{uv} = S_{xx} - kS_{yy} + (k - 1)S_{xy} . \tag{12}$$

Then, by definition,

$$r = \frac{S_{uv}}{(S_{uu}S_{vv})^{1/2}} = \frac{S_{xx} - kS_{yy} + (k-1)S_{xy}}{[(S_{xx} + S_{yy} - 2S_{xy})(S_{xx} + k^2S_{yy} + 2kS_{xy})]^{1/2}}. \tag{13}$$

Thus under the null hypothesis H_0,

$$t_0 = r\sqrt{\frac{n-2}{1-r^2}} \tag{14}$$

will have a Student's t distribution with $(n - 2)$ degrees of freedom (D.F.).

The particular case of $k = 1$ was considered by Maloney and Rastogi [1970] in detail, and the test they obtained is the same as above.

The particular case of $k = 0$, testing the hypothesis that $\sigma_1^2 = 0$, was considered by Maloney and Rastogi [1970] and by Jaech [1971], and the result in this section gives the exact distribution of the test statistic they considered.

4. LIKELIHOOD RATIO TEST

It can be easily seen that the test statistic used in section 3 is the same as that derived by the likelihood ratio given by Jaech [1971]. For simplicity, work with the transformed variates u_i and v_i $(i = 1, 2, \cdots, n)$ instead of x_i and y_i ; u_i and v_i have a bivariate normal distribution with variances and covariance given in (10). Let us denote the maximum value of the likelihood under the alternative hypothesis H_A by $L(\Omega)$.

$$\begin{aligned}\log L(\Omega) &= -\frac{n}{2} - \frac{n}{2}\log(S_{uu}S_{vv} - S_{uv}^2) - n\log(2\pi)\\ &= -\frac{n}{2} - \frac{n}{2}\log(S_{xx}S_{yy} - S_{xy}^2) - n\log(1+k) - n\log(2\pi).\end{aligned} \tag{15}$$

Under H_0, u and v are independently distributed. Denote the maximum value of the likelihood under H_0 by $L(\omega)$, and put $S_{uv}^2 = 0$ since u and v are independent; then (15) becomes

$$\begin{aligned}\log L(\omega) &= -\frac{n}{2} - \frac{n}{2}\log(S_{uu}S_{vv}) - n\log(2\pi)\\ &= -\frac{n}{2} - \frac{n}{2}\log[(S_{xx} + S_{yy} - 2S_{xy})(S_{xx} + k^2S_{yy} + 2kS_{xy})]\\ &\quad - n\log(2\pi).\end{aligned} \tag{16}$$

Thus,

$$\begin{aligned}\log\lambda &= \log L(\omega) - \log L(\Omega)\\ &= \frac{n}{2}\log\left[\frac{S_{xx}S_{yy} - S_{xy}^2}{(S_{xx} + S_{yy} - 2S_{xy})(S_{xx} + k^2S_{yy} + 2kS_{xy})}\right] + n\log(1+k).\end{aligned}$$

But from (13)

$$1 - r^2 = \frac{(1 + k)^2(S_{xx}S_{yy} - S_{xy}^2)}{(S_{xx} + S_{yy} - 2S_{xy})(S_{xx} + k^2S_{yy} + 2kS_{xy})},$$

thus

$$\log \lambda = \frac{n}{2} \log (1 - r^2). \tag{17}$$

The same value of log λ is obtained by Jaech's [1971] approach, using $\hat{\sigma}^2$ and $\hat{\sigma}_2$ obtained in (12) and (13) of his paper and putting them in his (5), (6), and (7).

The disadvantage of the above test is that the power depends upon σ^2 (the larger σ^2, the smaller the power). This has been discussed in more detail by Maloney and Rastogi [1970]. If the data from more than two instruments are available then a test can be obtained whose power is independent of σ^2 (Shukla [1972]). When $\sigma^2 = 0$ (one item being measured repeatedly) then the obvious test for $\sigma_1^2 = k\sigma_2^2$ is to consider the ratio of S_{vv} and S_{uu} and test by an F-test, where $u_i = \sqrt{k}y_i$ and $v_i = x_i$; $(i = 1, 2, \cdots, n)$.

5. CONFIDENCE INTERVALS FOR RELATIVE PRECISION

The results of section 3 can be used to obtain confidence intervals for k. Let us denote the Student's t value at $\alpha/2$ probability level with $(n - 2)$ D.F. by $t_{\alpha/2}$. Then $(1 - \alpha)\%$ confidence limits for k can be obtained from (14) by substituting the value of $t_{\alpha/2}$ in place of t_0 and substituting the value of ratio $r^2/(1 - r^2)$ in terms of S_{xx}, S_{yy}, S_{xy} and k.

The quadratic equation in k reduces to

$$k^2(a^2 - P) - 2k(ab + P) + (b^2 - P) = 0 \tag{18}$$

where

$$a = S_{yy} - S_{xy} ; \qquad b = S_{xx} - S_{xy} ; \qquad P = \frac{t_{\alpha/2}^2(S_{xx}S_{yy} - S_{xy}^2)}{(n - 2)}.$$

Solution of (18) gives K_U and K_L as upper and lower limits

$$K_L = \frac{b - \sqrt{P}}{a + \sqrt{P}} ; \qquad K_U = \frac{b + \sqrt{P}}{a - \sqrt{P}} ; \tag{19}$$

such that

$$P(K_L \leq k \leq K_U) = 1 - \alpha.$$

In example 1, considered by Maloney and Rastogi [1970], $n = 181$, $S_{xx} = 0.6069$, $S_{yy} = 0.5533$, $S_{xy} = 0.3054$, and $t_{\alpha/2} = 1.96$. The 95% confidence interval for k calculated by the above method is given by

$$K_L = 0.7168; \quad K_U = 2.1258.$$

ACKNOWLEDGMENTS

I am thankful to Professor D. J. Finney for helpful suggestions.

QUELQUES TESTS D'HYPOTHESES EXACTS CONCERNANT LES ESTIMATEURS DE GRUBBS

RESUME

Quand deux instruments ou deux techniques sont utilisées pour mesurer la même quantité, on peut estimer les précisions des mesures à l'aide d'une méthode proposée par Grubbs. Cet article généralise certains tests donnés par de précédents auteurs. En utilisant ce résultat général, on obtient quelques tests de certaines hypothèses exactes alors que seules étaient disponibles des approximations. On peut aussi utiliser ces résultats pour construire des intervalles de confiance exacts pour la précision relative des deux instruments.

REFERENCES

Curnow, R. N. [1957]. Heterogeneous error variances in split-plot experiments. *Biometrika 44*, 378–83.

Grubbs, F. E. [1948]. On estimating precision of measuring instruments and product variability. *J. Amer. Statist. Ass. 43*, 243–64, (corrections, pp. 564).

Jaech, J. L. [1971]. Further tests of significance for Grubbs's estimators. *Biometrics 27*, 1097–101.

Maloney, G. F. and Rastogi, S. C. [1970]. Significance test for Grubbs's estimators. *Biometrics 26*, 671–6.

Morgan, W. A. [1939]. A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–9.

Pitman, E. J. G. [1939]. A note on normal correlation. *Biometrika 31*, 9–12.

Shukla, G. K. [1972]. An invariant test for the homogeneity of variances in a two-way classification. *Biometrics 28*, 1063–72.

Received April 1972, *Revised October* 1972

Key Words: Grubbs's estimators; Exact tests of hypotheses; Confidence interval for precision.

Reprinted from
BIOMETRICS Copyright © 1973
THE BIOMETRIC SOCIETY, Vol. 29, No. 2, June 1973