Promotor : dr. P. van der Laan
hoogleraar in de wiskunde,
i.h.b. toegepaste statistiek

Co-promotor : dr.ir. L.R. Verdooren
universitair hoofddocent in de
statistische proeftechniek

On statistical selection in plant breeding

C. Johan Dourleijn

On statistical selection in plant breeding

Proefschrift

ter verkrijging van de graad van doctor
in de landbouw- en milieuwetenschappen,
op gezag van de rector magnificus,
dr. H.C. van der Plas,
in het openbaar te verdedigen
op maandag 22 maart 1993
des namiddags te vier uur in de Aula
van de Landbouwuniversiteit te Wageningen.

CIP-gegevens Koninklijke Bibliotheek, Den Haag

Dourleijn, C. Johan

On statistical selection in plant breeding /
C. Johan Dourleijn. - [S.l. : s.n.]. - Ill.
Proefschrift Wageningen. - Met lit. opg. -
Met samenvatting in het Nederlands.
ISBN 90-5485-096-5
Trefw.: plantenveredeling

Ontwerp omslag : Johan Dourleijn
met dank aan Sipko Mous

STELLINGEN

(1)

Het is naïef te denken dat statistische selectiemethoden het 'kwekersoog' kunnen vervangen. Wel kunnen ze dit oog openen voor de onzekerheden waaronder de veredelaar werkt en aldus een waardevolle aanvulling zijn op de huidige praktijk.

(dit proefschrift)

(2)

Bij het vergelijken van een lokaal ras, dat slechts op één locatie beproefd wordt, met andere rassen heeft het vaak zin informatie over de prestaties van laatstgenoemde rassen op andere locaties hierbij te betrekken.

(dit proefschrift)

(3)

In rassenonderzoek is een gecombineerde reeks proeven veel belangrijker dan één individuele proef. Het is daarom niet terecht dat de theorie over het ontwerpen en analyseren van rassenproeven zich grotendeels richt op laatstgenoemd type proef.

(4)

Bij toepassingsgericht onderzoek neemt de onderzoeker te snel de rol van Procrustes op zich. Hij overschat daarbij de flexibiliteit van de randvoorwaarden zoals gesteld door de praktijk.

(5)

Voor statistische computerprogrammatuur, aangeboden op de commerciële markt, zou een keurmerk in het leven moeten worden geroepen. Dit keurmerk zou verleend kunnen worden als de programmatuur technisch correct is en de vooronderstellingen bij en de beperkingen van de gebruikte achterliggende theorie duidelijk vermeld zijn.

(6)

Daar computersimulatie succesvol gebruikt kan worden bij de praktische toepassing van statistiek en in de exploratieve fase in statistisch onderzoek en onderwijs, moet deze techniek deel uitmaken van de denkwereld van iedere statisticus.

(7)

Vakkenevaluatie met behulp van een standaard vragenlijst is éénzijdig en vertekenend. Beter kan het welslagen van een vak bewerkstelligd worden door open overleg tussen docent en studenten gedurende de onderwijsperiode.

(8)

Het houden van referenda om de Bestuurlijke Vernieuwing gestalte te geven leidt niet tot het beoogde doel. Om tot een vernieuwing van de houding van bestuurders te komen, is een communicatieve benadering essentieel.

(9)

Het is wenselijk positieve discriminatie van vrouwen bij sollicitatieprocedures te modelleren en met behulp van simulatiemethoden de kans op correcte selectie te schatten.

(10)

Als een promovendus zijn proefschrift ziet als zijn levenswerk verdient het aanbeveling dat hij ter gelegenheid van zijn pensionering promoveert.

C.J. Dourleijn
On statistical selection in plant breeding
Wageningen, 22 maart 1993

Voor mijn ouders

Aan Miranda

Voorwoord

De titel van dit proefschrift, 'Over statistische selectie in de plantenveredeling', noemt zowel theorie als toepassing. Juist het combineren van deze twee vormde de uitdaging van mijn onderzoek als Assistent In Opleiding. Dit onderzoek werd uitgevoerd van 1 september 1988 tot 1 september 1992 aan de vakgroep Wiskunde, sectie Wiskundige en Toegepaste Statistiek, van de Landbouwuniversiteit te Wageningen. Ik dank mijn promotor prof.dr. Paul van der Laan voor het initiëren van dit AIO-project.

Bij aanvang van mijn onderzoek hielden op de vakgroep Wiskunde dr. Stefan Driessen, dr. Bram van Putten en prof.dr. Paul van der Laan zich met de theorie van statistische rangschikkings- en selectiemethoden bezig. Het is van hen dat ik veel op dit gebied geleerd heb. Vooral met Stefan heb ik veel over het onderwerp gediscussieerd. Ik wil Stefan, Bram en Paul hartelijk danken voor hun begeleiding. Op het gebied van opzet en analyse van rassenproeven heb ik veel geleerd van dr.ir. Rob Verdooren en dr. Aad van Eijnsbergen. Ook hen wil ik hartelijk danken voor de begeleiding en samenwerking.

Natuurlijk past ook een woord van dank aan de overige medewerkers van de vakgroep Wiskunde die, bewust of onbewust, een bijdrage geleverd hebben aan de totstandkoming van dit proefschrift. Ook het afstudeerwerk van ir. Johan Schut en ir. Bert Bos is van invloed geweest op het uiteindelijke resultaat. Ik kijk met plezier terug op de samenwerking met zowel Johan als Bert.

Een proefschrift schrijven over het gebruik van statistische selectie in de plantenveredeling zonder contact met de veredelingspraktijk zou op zijn zachtst gezegd vreemd zijn. Daarom ben ik erg blij dat er goede contacten gegroeid zijn met de onderzoeksafdeling van het veredelingsbedrijf Royal Vanderhave Group te Rilland, Nederland. Ik ben dit bedrijf dankbaar voor de financiële ondersteuning van mijn onderzoeksproject.

Zonder de gesprekken met dr. Marta Lamberts-Morales en ir. Ab van Spijk, beiden werkzaam bij Vanderhave, zou hoofdstuk 2 niet zijn huidige inhoud gekregen hebben. Ik heb hun vriendelijke medewerking bijzonder gewaardeerd.

Johan Dourleijn
Wageningen, 22 maart 1993

Table of contents

CHAPTER 1

Introduction

Theory and practice are often two different worlds. This is also largely true for the theory of statistical selection and the plant breeding practice. However, in this thesis we will try to combine both fields. In **1.1** a brief introduction to the theory of statistical selection is given. Introductory remarks about selection in the plant breeding practice are made in **1.2**. Next, the objective of this doctoral research is described in **1.3**. Finally, the outline of this thesis can be found in **1.4**.

1.1 Statistical selection

In everyday life we frequently come across stochastic variables of which we do not know the expected or true value but only observe their realisations. To make a good guess about the expectation of such variables a statistical model is constructed for the observations and the model parameters are estimated. In the agricultural sciences often several treatments are being compared. These treatments can be different (amounts of) fertilisers, different feeding regimes, different plant varieties, etcetera. To compare the various treatments, the experimenter performs an experiment in which the treatments are applied and observes the outcomes for each treatment. Reality is simplified into a model for the observations that contains parameters corresponding to the treatments. These treatment parameters are estimated, and the treatments are compared through these estimates.

Now assume that the objective of the experimenter is to decide which treatment is the best (with 'best' defined appropriately) and to select this best treatment. This goal cannot be reached with the traditional statistical inference procedures such as testing the assumption that the parameters of all treatments are equal (see e.g. Lehmann, 1986) or multiple comparisons procedures (see e.g. Hochberg & Tamhane, 1987). Therefore, statistical theory was developed to deal with selection problems. The inference corresponding to statistical selection is a statement about the probability of correct selection. A selection is called 'correct' if the best treatment is selected. The founders of this theory were Robert E.

Bechhofer and Shanti S. Gupta. The former started the statistical selection theory with Indifference Zone selection (Bechhofer, 1954), where only one treatment is selected and the probability of correct selection is guaranteed if the distance between the parameter of the best treatment and the parameter of the second-best treatment is at least equal to some prespecified value. Gupta (1965) developed the theory of subset selection, where a random sized subset is selected (with the use of a selection rule) and the inference is made that the probability of correct selection is at least equal to some prespecified value. Thanks to many researchers, the theory of statistical selection has extended enormously, as appears from the hundreds of articles on this topic.

1.2 Selection and plant breeding

The main goal of the experimenters in the plant breeding practice is to develop new varieties that are better than the existing ones, which are currently on the market (see also Kempton & Talbot, 1988). New genotypes are created by traditional crossing or biotechnological methods. These genotypes have to be compared mutually and with control varieties in order to make a selection. 'Selection' in the plant breeding context is often understood as the cycle of choosing a certain percentage of a population of plants and allowing these plants to cross mutually, choosing plants in the next generation and allowing them to cross mutually, and so on. This way a selection programme is obtained that leads to better genotypes. In this thesis however, we mean by 'selection' the process of choosing genotypes because we hope that these particular genotypes can be marketed. The selected genotypes are submitted to the official variety testing authorities and further tested and selected. Consequently, it must be possible to reproduce the selected genotypes.

To compare different varieties, they are grown in various experiments at different sites and sometimes in several years. The results of just one experiment at a certain site in a particular year are not sufficient to base the ultimate selection on. Therefore, series of variety trials are very important in the plant breeding practice.

The statistical selection procedures introduced in **1.1** are rarely used in the plant breeding practice. In text books about plant breeding the type of selection introduced in **1.1** is almost never mentioned. An exception is Mayo (1987). However, the attention he pays to the subject is restricted to a (rather curious) reference to a correction note of Bechhofer, corresponding to one of his articles about Indifference Zone selection. To analyse the experiments statistical (estimation) theory is greatly used, but the decision how many varieties to select is seldomly founded on statistical theory.

1.3 Objective of this study

Since 1954 the theory of statistical selection has grown rapidly. However, the practical application of statistical selection procedures has been an underdeveloped area. This has also been recognised by the statistical selection researchers of the first hour (e.g. Bechhofer, 1985).

The objective of this study is to explore the possibilities to use statistical selection in the plant breeding practice. The plant breeding practice seems a suitable application field because selection is the essence of plant breeding. However, problems have to be expected because the experimental designs used are often not the relatively simple ones such as the completely randomised designs or randomised complete block designs. Also, the selection has to be based on 'mean performance' results from a series of trials.

Before the selection rules can be executed, parameters corresponding to the varieties have to be estimated as good as possible. The specific designs and the series of experiments used in the plant breeding practice can hamper the estimation of the variety parameters. Before applying selection rules these difficulties have to be dealt with.

Besides translating existing statistical selection procedures into variety testing terminology, new selection procedures have to be developed if necessary. It must be made possible to execute the selection rules corresponding to the various selection procedures for every experimental design used in the breeding practice, including designs associated with series of trials. This means that it must be possible to calculate the so-called selection constants of the selection rules for every design.

The practical application of statistical selection procedures in plant breeding might ask the development of computer software. With large numbers of varieties to work with, the use of the computer to calculate the selection constants and to execute the selection rules is a must.

Briefly said, the objective of this study is to bridge the gap between statistical selection theory and the plant breeding practice.

1.4 Outline of this thesis

Since this thesis deals with the application of statistical selection procedures, the application field has to be studied first. Therefore, **chapter 2** is concerned with the course of the sugar beet breeding practice. The author spend some time at the research unit of the Royal Vanderhave Group at Rilland, The Netherlands, to study this branch of plant breeding. In chapter 2 the breeding programme of sugar beets, the types of experiments performed during the breeding process and the current selection methods are described. This chapter reveals several difficulties that have to be overcome before statistical selection procedures can be used successfully in the breeding practice.

In our view statistical selection comes in two parts. The first part is to estimate the parameters on which the selection has to be based as good as possible. The second part is the performance of statistical selection procedures. **Chapter 3** deals with the estimation of contrasts between variety values, using linear models for the observations of the varieties. These contrasts are sufficient for selection purposes. Much of chapter 3 was written in close co-operation with dr. A.C. van Eijnsbergen.

First, the estimation procedure for a single trial is described, using the fixed additive model or the mixed additive model. However, in the plant breeding practice series of experiments are often more important than a single trial.

In the second part of chapter 3 the best linear unbiased estimators (BLUEs) corresponding to a series of experiments are obtained by combining the BLUEs from the separate trials. How this must be done depends also on the model chosen. It appears that the series of experiments can be designed in such a way that the separate estimates can be combined relatively easy. An article based on this part of chapter 3 is submitted to the *Journal of Statistical Planning and Inference*, and is entitled : Combining estimators from a series of variety trials, by C.J. Dourleijn & A.C. van Eijnsbergen. This article appeared as prepublication as Technical Note 92-08 of the Department of Mathematics, Agricultural University Wageningen.

The third part of chapter 3 deals with another type of design typical for the plant breeding practice : the concatenated trial. This trial can be divided into a number of subtrials that are only connected with each other through a small number of control varieties. Here the BLUEs corresponding to the concatenated trial can be obtained by combining local BLUEs corresponding to the subtrials. The theory of this part is also elaborated in an article entitled : Combining estimators from variety trials that are connected by control varieties only, by C.J. Dourleijn. This article is submitted to *Communications in Statistics - Theory and Methods*, but is already available as prepublication as Technical Note 92-09 of the Department of Mathematics, Agricultural University Wageningen.

In **chapter 4** the second part of statistical selection is studied. First, the most important available selection procedures are described. The distinction is made between Model I selection and Model II selection. In this thesis we mainly pay attention to subset selection procedures, which correspond to Model I, because they are associated with the analysis of a trial. The Indifference Zone selection procedures are more interesting from the design point of view.

In the second part of chapter **4** new subset selection rules for randomised experiments are given. Some of these rules are very convenient to use in the breeding practice. My research partner in this study and co-author of *4.2.1-4.2.4* was dr. S.G.A.J. Driessen. An article covering *4.2.1-4.2.4* is accepted for publication in the *Biometrical Journal* (to appear about April 1993). The title of the article is : Subset selection procedures for randomized designs, by C.J. Dourleijn & S.G.A.J. Driessen. It is also available as Technical Note 91-02 of the Department of Mathematics, Agricultural University Wageningen.

The third part of this chapter deals with the use of computer simulation to calculate the selection constants (which are necessary to execute a selection rule), the probability of correct selection and the expected subset size. Especially for the typical plant breeding trials we have to use simulation methods to calculate the important statistics. This part of chapter **4** is based on Technical Note 91-03 of the Department of Mathematics, Agricultural University Wageningen, entitled : The use of simulation in statistical selection, by C.J. Dourleijn.

With the new selection rules and the possibility to approximate the selection constants included in these rules, subset selection can be applied to plant breeding selection problems. This is elaborated in the fourth part of chapter **4**. The results are also written in an article, entitled : Subset selection in the plant breeding practice, by C.J. Dourleijn. This article will be submitted to *Euphytica*. Some results can also be found in Dourleijn, C.J. (1991) : Subset selection in the plant breeding practice; *Proceedings of the 8th meeting of the Eucarpia section Biometrics in Plant Breeding*, held July 1-6 1991 at Brno, Czechoslovakia, 287-296.

In the fifth and last part of this chapter we propose some modifications of subset selection procedures. With these modifications we hope that the subset selection procedures gain in value for the practice.

Chapter 5 deals with the execution of the subset selection rules. In the plant breeding practice, where often tens to hundreds of varieties are tested, computer software to actually make the selection is indispensible. Such software, written by the author, is described in the first part of chapter **5**.

In the second part of this chapter we consider a case study. The data were generously supplied by the research unit of the Royal Vanderhave Group at Rilland, The Netherlands. Most of the aspects described in chapters **2** to **4** are incorporated in this case study.

Finally, **chapter 6** gives a discussion and conclusions.

REFERENCES

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. (1985). Selection and ranking procedures - some personal reminiscences, and thoughts about its past, present, and future. *Am. J. of Math. and Man. Sc.*, vol. 5 nos. 3 & 4, 201-234.

Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* 7, 225-245.

Hochberg, Y. & A.C. Tamhane (1987). *Multiple comparison procedures*. Wiley & Sons, New York.

Kempton, R.A. & M. Talbot (1988). The development of new crop varieties. *J. R. Statist. Soc.* A, 151 Part 2, 327-341.

Lehmann, E.L. (1986). *Testing statistical hypotheses*. 2nd. edition. Wiley & Sons, New York.

Mayo, O. (1987). *The theory of plant breeding*. 2nd. edition. Clarendon Press, Oxford.

CHAPTER 2

Sugar beet breeding

The application field that runs through this thesis like a continuous thread is the plant breeding practice. As we will see, design and analysis of experiments and selection problems are very important there. In this chapter we will concentrate on the breeding of sugar beets (*Beta vulgaris* L.), although the theory in the following chapters is not restricted to a particular crop. In **2.1** we will briefly describe the breeding programme of sugar beet varieties at the research centre of the Royal Vanderhave Group in Rilland, The Netherlands. Of special statistical interest are the types of experiments and designs used. These are described in **2.2**. Finally, in **2.3**, we will make an inventory of the selection problems typical of sugar beet breeding.

2.1 The breeding programme

When we consult the Dutch descriptive list of agricultural crop varieties of 1991, we notice that most of the modern sugar beet varieties carry three sets of chromosomes : they are triploid, denoted by 3N. These varieties are hybrids, often produced by a cross between a diploid (2N) mother and a tetraploid (4N) father. To produce this cross on a routine basis, the mother plants have to be male sterile. This can be achieved by using cytoplasmic male sterility, which was discovered for sugar beet by Owen (1945). A plant is male sterile if it has sterile cytoplasm and two pairs of recessive fertility restoring genes in the nucleus. Plants with two pairs of recessive fertility restoring genes but with normal cytoplasm (and therefore male-fertile) are called O(wen)-types. If a male-sterile plant is crossed with an O-type, the offspring (which has the maternal cytoplasm) is male sterile. The female parent of a triploid sugar beet variety is a single cross (SC) hybrid, the result of a cross between a male-sterile inbred line and an inbred O-type. Although sugar beet, which is a cross-pollinator, has a self-incompatibility mechanism to prevent self-pollination, it is not completely self-sterile. This makes it possible to produce inbred lines by repeated selfing (Ellerton & Arnold, 1982).

The varieties in the Dutch descriptive list are all monogerm. This means that the fruits (often called 'seeds' for the sake of simplicity) contain only one embryo, so one seed gives one seedling. Monogermy is a character of the female parent rather than a character of the offspring : if the female parent produces single flowers instead of clusters of flowers the seeds are monogerm. Therefore the plant breeder selects female parents that are, besides male sterile, also monogerm. Because the females have to be *mo*nogerm and *m*ale *s*terile, they are qualified as being 'moms'. The moms female parents and the male parents (or pollinators) of the commercial varieties are developed in separate breeding procedures.

The development of moms SC hybrids starts with the selection of individual plants from a population of diploid plants. The selected plants are monogerm and have recessive fertility restoring genes but normal cytoplasm. Repeated selfing of these O-types results in inbred O-types. With each O-type an isogenic moms inbred line is developed by repeated back crossing. Ideally, the only difference between such a moms inbred and its associated O-type is the male sterility of the former. An inbred O-type and its associated moms inbred line are called a 'couple'. The production of new moms inbred lines and their associated inbred O-types can be seen as a separate selection programme (Sneep & Hendriksen, 1979). We will only describe the selection procedure that follows. In the first year the new moms inbreds are crossed with four standard, unrelated inbred O-types. A cross of a moms inbred line with an unrelated inbred O-type gives a moms SC hybrid. The produced seed is sown the second year on experimental fields with conditions favourable for bolting. A bolter is a plant that enters the generative phase already in the first year after sowing. The moms SC hybrids that show too many bolters are discarded. The hybrids are also screened for male sterility and monogermy. Seed is being reserved for crossing of the selected moms SC hybrids with four different standard 4N pollinators in the third year. The 3N experimental hybrids that are produced this way are grown on the experimental fields in the fourth year. The composition of the experimental hybrids is as follows (mother × father) :

moms inbred line (2N) × unrelated inbred O-type (2N)
↓
moms SC hybrid (2N) × pollinator (4N)
↓
experimental hybrid (3N)

The experimental 3N hybrids are produced to estimate the general combining ability of the new moms inbred lines. The breeder searches for moms inbreds that will be good parents in the future. However, a moms inbred line determines only 1/6 of the genetic content of an experimental hybrid. The question may arise whether the genetical variation is not too small to make a meaningful selection. In practice about 20% of the new moms inbred lines are selected. In the fifth year the selected moms inbreds are crossed with new, unrelated inbred O-types according to a (partial) diallel scheme. This way the seed of various new moms SC hybrids is produced. These new moms SC hybrids are checked in the sixth year on monogermy, male sterility and bolting resistance. After that, in year seven, the remaining moms SC hybrids are crossed with ten different, standard 4N pollinators. The 3N experimental hybrids which are thereby produced, appear in year eight on the experimental fields. Besides estimation of the general combining ability of the new moms inbred lines and the new O-types, the estimation of the general combining ability of the moms SC hybrids is of paramount importance in this selection phase. The moms SC hybrids that are expected to be good female parents of a triploid hybrid variety are selected. However, a moms SC hybrid determines only 1/3 of the genetic content of the 3N experimental hybrids. Again the question may arise whether efforts bear proportion to the selection results. The breeding procedure after the forming of couples is recapitulated in Table 2.1.

Table 2.1. Breeding procedure of new moms single cross (SC) hybrids. Abbreviations : moms = monogerm + male sterile, TO = 2N inbred O-types, Poll. = 4N pollinators.

Year	
...	Formation of couples.
1	New 2N moms inbreds × 4 standard, unrelated TO → 2N moms SC hybrids.
2	Testing moms SC hybrids for monogermy, male sterility and bolting resistance.
3	Remaining moms SC hybrids × 4 standard Poll. → 3N experimental hybrids.
4	Testing experimental hybrids.
5	Selected 2N moms inbreds × new, unrelated TO's → 2N moms SC hybrids.
6	Testing moms SC hybrids for monogermy, male sterility and bolting resistance.
7	Remaining moms SC hybrids × 10 standard Poll. → 3N experimental hybrids.
8	Testing experimental hybrids.

The selection of pollinators starts with the formation of a population of tetraploid plants. The chromosome number of diploid plants can be doubled by a colchicine treatment of germinating seeds (Sneep & Hendriksen, 1979). From a (F2) population of tetraploids *i*ndividual *p*lants (ip's) are selected. Consider this as preparation for the following breeding procedure. In the first year of this breeding procedure a standard moms SC hybrid is crossed with the various ip's. At the same time the ip's are maintained in vitro. The produced experimental hybrids are grown on the experimental fields in the second year. The genetic variation among these experimental hybrids is due to 2/3 of the genetic content. If we compare this with the experimental hybrids in the moms hybrids breeding procedure, we can expect the selection of superior pollinators to be more successful. A number of ip's is selected and multiplied in vitro, thereby creating *ip c*lones (ipc's). The selected experimental hybrids are reproduced in the third year by crossing a standard moms SC hybrid with the selected ipc's. In year four the resulting experimental hybrids are grown on the experimental fields and a selection is made. The selected ipc's are further multiplied in vitro. However, the ipc's cannot be used as male parents of commercial sugar beet hybrids, because the in vitro propagation would be too labour intensive. In practice the F2 of a (poly)cross between ipc's is used as male parent. In the fifth year the ipc's are crossed mutually, resulting in *poly*cross *F1*'s (polyF1's). A standard moms SC hybrid is crossed with these polyF1's in the sixth year. The resulting experimental hybrids are tested on the experimental fields in year seven. After selection, the ipc's that correspond with the selected polyF1's are further multiplied in vitro. With this plant material the polyF1's can be reproduced in year eight, followed in year nine by the reproduction of experimental hybrids. These hybrids can be tested a year later. In Table 2.2 the breeding procedure of pollinators is summarized.

In both breeding procedures there are several years where experimental hybrids are tested and a selection is made. In the breeding procedure of new moms SC hybrids this is the case in years 4 and 8. The main objective of this procedure is to select good female parents of future 3N hybrid varieties. In the breeding procedure of new pollinators the selections are made in years 2, 4, 7 and 10. There, promising male parents of future varieties are selected. In the final years the individual breeding procedures become opaque, because the moms hybrid procedure and the pollinator procedure are combined in these years. Potential varieties are produced by crossing good moms SC hybrids with the F2's of good crosses between ipc's. These potential varieties are tested several years at various sites. Finally, the apparently best potential varieties are selected.

Table 2.2. Breeding procedure of new pollinators. ip = 4N individual plant, ipc = ip clone, moms = monogerm + male sterile, polyF1 = 4N polycross F1.

Year	
...	Formation of 4N population; selection of ip's.
1	Standard 2N moms SC hybrid × ip's → 3N experimental hybrids; in vitro maintenance of ip's.
2	Testing experimental hybrids; multiplication of ip's to ipc's.
3	Standard 2N moms SC hybrid × selected ipc's → 3N experimental hybrids.
4	Testing experimental hybrids; multiplication of ipc's.
5	New ipc's × new ipc's → polyF1's.
6	Standard moms SC hybrid × polyF1's → experimental hybrids.
7	Testing experimental hybrids; multiplication of ipc's.
8	Reproduction of polyF1's.
9	Reproduction of experimental hybrids.
10	Testing experimental hybrids.

In general, we can consider the situation where a number of experimental hybrids is tested with the use of field experiments. We can distinguish two goals: (1) The selection of good parents of future varieties; (2) The selection of potential varieties themselves. In both situations the produced genotypes, which can be reproduced, are of primary interest to the breeder and not the (theoretical) population of all possible genotypes. Although the goal in the pollinator breeding procedure is (1), it is effectively the same as (2), because only one standard moms SC hybrid is used to produce the experimental 3N hybrids. We will mainly pay attention to trials, set up for goal (2). Although the experimental hybrids are not varieties (yet), this situation is in general denoted by 'variety testing'. We will adopt this notation for reasons of generality and simplicity. Thus instead of 'experimental hybrid' we will write 'variety'. Although different breeders may use different breeding procedures, they always will perform variety trials.

2.2 Types of experiments performed

Numerous experiments are performed every year. Some of them are designed to test moms SC hybrids for monogermy and male sterility (see Table 2.1). These experiments are performed in the open air and in greenhouses. Other experiments are specifically set up to test bolting resistance of the varieties. The bolter trials are located on sites with relatively low temperatures in spring. Also early sowing stimulates the occurrence of bolting. Although the above experiments are very important, we will concentrate on the most performed type of trials, in general denoted by 'variety trials'.

Depending on the selection phase, tens to hundreds of varieties are included in those performance trials. For example, in the second year of the pollinator breeding procedure about 600 varieties are tested and in the fourth year approximately 150. The varieties are grown at a number of different sites. Varieties in an advanced selection phase are grown at more sites than varieties in earlier phases. For example, the number of sites increases from 4 in the second year of the pollinator breeding procedure to 10 in the fourth year. The final selection of commercial hybrids is based on the results of more than 20 sites. For varieties intended for the North-West European market these sites include mainly locations in The Netherlands, Germany, France and the United Kingdom. First consider the field experiments at one site.

To take account of differences within the trial environment almost always an experiment with an incomplete block design is used to test the varieties. The number of varieties is much too large to use complete blocks. When the number of varieties is very large it is often not feasible to include all varieties in one large experiment with a suitable incomplete block design. There are a number of reasons why this it not feasible. First, up to the very last moment varieties can be added to the cohort which has to be tested. Thus the experimental design to be used can only be selected shortly before sowing, which is very inconvenient. Second, it is desirable that varieties which share a common genetical background are not grown too far apart. Grouping of the plant material makes the experimental field more surveyable for the plant breeder. Third, for organizing the various experimental field activities it is very convenient if standard designs can be used every year. Use of standard designs also reduces the number of mistakes made.

Because of the aforesaid reasons it is often easier for the plant breeder to organize several small trials at one site instead of one large experiment. The varieties are then divided into disjunct sets and each set is laid out in an experiment with an incomplete block design. The varieties in one set are often more or less genetically related to each other. However, all varieties have to be compared with one another and a selection has to be made. Therefore, control varieties are used to connect the trials (see also section 3.3). The varieties can then be compared via these control varieties. The incomplete block designs used are nearly always unbalanced, because there is only a limited amount of seed available. A realistic example is the following : If there are 154 varieties, 7 disjunct sets of 22 varieties are made. Together with 3 control varieties, 22 varieties can be included in a 5×5 lattice design. Often a 5×5 lattice design with 3 or 4 replications is used. Within each replication a 5×5 lattice design has 5 incomplete blocks containing 5 varieties each. The lattice design can be found in textbooks on experimental design, e.g. Cochran & Cox (1957).

Each set of varieties is grown at a number of sites. However, due to the large number of varieties, the limited area of experimental fields and the limited amount of seed, not all sets are grown at the same sites. Therefore, if we would make a variety × site table, it would be incomplete. As an example, part of the variety × site table corresponding to the second year of the pollinator breeding procedure at the research centre of the Royal Vanderhave Group (see Table 2.2) in 1986 will be given. In that year and selection phase 26 disjunct sets, each containing 22 new experimental hybrids and 3 control varieties, were distributed over 11 sites in North-West Europe. In Table 2.3 the incidence of 12 sets at the various (coded) sites is given. Notice that the scheme in Table 2.3 has a lot of empty cells. This is typical for the experimental situation in this selection phase.

Table 2.3. Incidence scheme of 12 of the 26 sets of varieties tested in the second year of the pollinator breeding procedure of the Royal Vanderhave Group in 1986, at 11 sites. Each set of varieties contains 22 new experimental hybrids and 3 control varieties. The actual site names are coded (S1, ..., S11) to maintain trade secrecy.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
Set no.											
1	•		•			•				•	
2	•	•				•				•	
3	•	•				•				•	
4	•					•				•	
5	•	•				•				•	
6	•	•					•			•	
7						•				•	
8								•			•
9	•		•	•			•				
10	•		•	•		•					
11	•		•	•		•					
12	•	•					•			•	

2.3 Selection

There are a number of characters on which the selection decisions are based. The root yield is often corrected for missing plants (gaps) in the plots, and is therefore denoted by *c*orrected *r*oot *y*ield (CRY). The correction method has been determined empirically. The CRY is usually given in ton/ha. The *s*ugar *c*ontent (SC) in % is determined with a polarimeter, as well as the α-amino N content (N%). Also the potassium content (K%) and the sodium content (Na%) are determined. A sugar beet is composed of approximately 75% water and 25% dry matter. The 25% dry matter can be divided into about 20% soluble matter and 5% insoluble matter. The soluble part can in its turn be divided into approximately 16% sucrose, 1.8% N-containing organic matter, 1.4% N-free organic matter and 0.8% minerals (Johnson *et al.*, 1969). The *t*are (T) , in kilogram per ton clean beets, is calculated as the difference between gross root yield (in ton/ha) and net root yield (in ton/ha), divided by the net root yield and multiplied by 1000. There are a number of characters that are calculated from the above mentioned characters. The *c*orrected *s*ugar *y*ield (CSY) is calculated as CRY×SC/100. The *w*hite *s*ugar *c*ontent (WSC) is calculated as WSC = SC - {0.343 (Na% + K%) + 0.094 (N%) + 0.29}, an empirical formula by N.J. van Geijn. The *w*hite *s*ugar *y*ield (WSY) is calculated as CRY×WSC/100. Other characters are *s*ugar *l*oss (SL), defined as SC-WSC, the *e*xtraction *i*ndex (EI), calculated as 100×WSC/SC and the sum of the potassium- and sodium content (KNa%), calculated as K% + Na%.

In special bolter trials the number of *bol*ting plants (BOL) is counted for each variety. This number reflects the bolting resistance of a variety. Varieties with a large number of bolters have little resistance to bolting. In breeding for disease resistance attention is paid to rhizomania resistance, yellows resistance and resistance to beet cyst-nematodes (*Heterodera schachtii*). Qualitative characters like beet shape are of minor importance and therefore we will restrict ourselves to the quantitative characters.

After the analyses of the trials at the various sites, the results of the different sites are combined into a so-called mean performance for each variety. The selection decisions are mainly based on the mean performance of the varieties, and secondly the question arises whether the performance of the varieties is relatively stable over the various sites. This is especially the case if the number of varieties is very large, and inspection of all varieties at the individual site level would be too time consuming. If the number of varieties is relatively small, which is the case in advanced selection phases, the individual site results are also important. The main interest of a sugar beet breeder lies in varieties that are stable over a wide range of sites. However, due to variety × site interactions the region in which a variety is more or less superior is limited. Therefore, the plant breeder develops different varieties for different regions.

Yield has always been a very important character for selection, but in the last decade the internal quality of the beets has become increasingly important. The internal quality is made up by the sugar content and the extraction index. The internal quality is high if both the SC and the EI are high. A large EI is achieved if the juice purity is high. This is the case if the percentage α-amino N, K and Na is small. A large sugar yield can be the result of a large root yield or a high sugar content. There is a negative correlation between CRY and SC. The selection is based primarily on CRY, CSY, WSY, SC, WSC and BOL. The decisions are made with three control varieties as reference. Plant breeders like to express the values of new varieties relatively w.r.t. the average value of the control varieties. The various characters can partially compensate each other, so the selected varieties do not necessarily have to be better than the control varieties for all characters. However, lack of bolter resistance cannot be compensated by any other character; all varieties with BOL greater than a certain threshold value are discarded. The selection decisions are made in a rather subjective way. The number of varieties to be selected is restricted by the capacity of the experimental fields. When the total number of selected varieties is too large, the selection limits are adjusted. Also from the logistics point of view it is inconvenient to have much fluctuation over the years in the number of retained varieties. Whereas statistics plays an important role in the estimation of variety parameters, it does not (at present) in making selection decisions. Much time and money is spend to estimate the variety parameters properly, followed by selection based on a rule of thumb and the subjective breeder's eye. Although selection of the really best variety would of course be welcomed by the plant breeder, he is satisfied if in the end a few varieties are selected that are better than the control varieties.

The ultimate aim of commercial sugar beet breeding is to produce varieties that are preferred by the farmer to varieties of rival plant breeding companies. The choice of the farmer mainly depends on the *fin*ancial yield (FIN) of the varieties. The financial yield is the amount of money a farmer receives from the sugar factory for the yield of one hectare of the variety. The financial yield of a variety is determined by the characters CRY, SC and juice purity, and of course by the type of formula used. The appearance of the variety in the field is of minor importance to the farmer, except frequent occurrence of bolters. Because the financial yield primarily determines the choice of the farmer, this linear combination of various characters could be used as a selection index. Let V be the threshold value of tare in kilogram per ton beets beneath which no penalty for tare is given by sugar factories. Let the penalty per ton tare above V be denoted by M. The price of one ton beets is denoted by P. The sugar factories give a bonus or penalty S, say, per ton beets per % sugar content higher or lower, respectively, than 16%. Then the financial yield (in Dfl.) is calculated in The Netherlands as :

$$\mathrm{FIN} = \mathrm{CRY}\left[\mathrm{P} + \mathrm{S}\{(\mathrm{SC}-16) + 0.08(\mathrm{EI}-85)\} - \frac{\mathrm{M}(\mathrm{T}-\mathrm{V})^{+}}{1000}\right],$$

with $(\mathrm{T}-\mathrm{V})^{+} = \max(\mathrm{T}-\mathrm{V}, 0)$. Consider the following small example. A farmer has 10 ha sugar beets with a net yield of CRY=50 ton/ha. Let SC = 17%, EI = 80%, T = 100 kg/ton beets, P = Dfl. 105, S = Dfl. 9, M = Dfl. 20, V = 75 kg/ton beets. Then FIN = 50 [105 + 9 {1 + 0.08(-5)} - 0.5] = Dfl. 5495 /ha. We have to bear in mind that the FIN formula differs for different countries and/or sugar factories. At this moment, little use is made of the financial yield for selection purposes. The benefit of using a selection index is that the various characters are combined into one new character. The amount of information is reduced and the selection will be easier. However, the problem is to find a satisfactory selection index. Characters like WSY and WSC are also indices, but do not represent all the aspects a breeder wants to base the selection on.

REFERENCES

Cochran, W.G. & G.M. Cox (1957). *Experimental designs.* 2nd edition. Wiley & Sons, New York.

Ellerton, S. & M.H. Arnold (1982). Sugar beet breeding. In : *Fifty years of sugar beet research.* International institute for sugar beet research, Brussels, 19-33.

Johnson, R.T. *et al.* (Eds.) (1969). *Advances in sugar beet production, principles and practices.* The Iowa State University Press, Ames, Iowa.

Owen, F.V. (1945). Cytoplasmically inherited male sterility in sugar beets. *Journal of Agricultural Research* 71, 423-440.

Sneep, J. & A.J.T. Hendriksen (Eds.) (1979). *Plant breeding perspectives.* Centennial publication of Koninklijk Kweekbedrijf en Zaadhandel D.J. van der Have. Centre for Agricultural Publishing and Documentation, Wageningen.

CHAPTER 3

The estimation of contrasts between variety values

The ultimate goal of the plant breeder is to select new varieties that are better than the varieties currently available. In order to get a good assessment of the agricultural value of a new variety, this variety is included in experiments performed at various sites and sometimes in several years. The observations are described by a model, and specific linear combinations of the parameters of this model are used for the evaluation of the new varieties. It is therefore of the utmost importance to the plant breeder that these linear combinations of parameters are estimated as good as possible. However, different models can be used, often resulting in different estimates and, even more important, a different ranking of these estimates. In this chapter we will discuss (generalised) least squares estimation of contrasts between variety values for various models. First, in **3.1**, estimation at a single site in a particular year is described. Next, the estimates from the various sites and years can be combined. This is described in **3.2**. Finally, in **3.3**, we will study estimation in case trials at one site are subdivided into subtrials, a situation frequently occurring in the sugar beet breeding practice.

3.1 Estimation at the individual sites

We will now focus on the estimation of model parameters for a single site and year. Suppose t varieties are grown in an experiment. In all the subsequent models the variety contributions are taken fixed, because we are specifically interested in the varieties actually used and not in an underlying population of varieties. Assume that the experiment performed has a randomised (in)complete block design with b blocks, and that there are n_{ij} observations of variety i $(i=1,\dots,t)$ in block j $(j=1,\dots,b)$. The total number of observations will be denoted by n. It is common practice to use an additive model for the observations of a variety trial with a block design at a certain site and year. This additive model can also be the result of a logarithm transformation of a multiplicative model. In an additive model an observation of variety i in block j is modelled as the sum of a parameter related to variety i, a term related to block j and a quantity depending on the particular experimental unit on which the observation was made. The latter quantities are called errors and are assumed to be uncorrelated random variables with zero expectation and common variance. Using an additive model it is assumed that there is no interaction between varieties and blocks. Which type of additive model has to be used depends on the way the blocks were chosen. If the blocks represent a random sample from a large population of blocks, the block terms can be considered random and a mixed additive model can be used. This model is described in *3.1.2*. If the blocks were purposively chosen to reduce intra-block variation, the block terms can be considered fixed and a fixed additive model is more appropriate. In *3.1.1* this model is treated.

We assume that the design of the experiment is connected. Loosely stated, a design is connected if it is possible to follow a path through all non-empty cells in the variety × block table, with the restriction that this path only links cells corresponding to the same variety or the same block. Because much of the theory in the sequel of this chapter is valid for connected designs only, we additionally will explain connectedness by means of three small examples. Consider the following three variety × block tables, with V1 = variety 1, B1 = block 1, and so on. The varieties included in a block are indicated by a dot (•).

(I)	B1	B2	B3	B4
V1	•		•	
V2		•		•
V3	•		•	
V4		•		•

(II)	B1	B2	B3	B4
V1	•		•	•
V2		•		•
V3	•		•	
V4		•		•

(III)	B1	B2	B3	B4
V1	•		•	
V2		•	•	
V3	•			•
V4		•		•

Design (I) is disconnected; it has to be apprehended as two separate designs : varieties 1 and 3 in blocks 1 and 3, and varieties 2 and 4 in blocks 2 and 4. If we assume a fixed additive model, then in design (I) only the differences between varieties 1 and 3 and varieties 2 and 4 can be estimated free from block contributions. For the other variety differences this is not possible. By adding variety 1 in block 4, resulting in Design (II), the design becomes connected and all differences can be estimated free from block contributions. Also connected is Design (III), which has the same number of observations as Design (I) but a different allocation of the varieties to the blocks.

3.1.1 The fixed additive model

After the experiment has been performed, characters as mentioned in chapter **2** are observed. For each character least squares estimates of linear combinations of model parameters, reflecting the values of the varieties for that character, have to be calculated. The character observed at the k^{th} plot with variety i in block j is denoted by Y_{ijk}. The model we will use for the observations is :

$$Y_{ijk} = \lambda + \tau_i + \beta_j + E_{ijk} , \qquad i = 1, \ldots, t , \; j = 1, \ldots, b , \; k = 1, \ldots, n_{ij} , \tag{3.1}$$

with λ a general level parameter, τ_i the parameter corresponding to variety i, β_j the parameter corresponding to block j and E_{ijk} the plot error. The E_{ijk} are assumed to be uncorrelated random variables with zero expectation and common variance σ^2.

Before using this model, it should be checked whether the model assumptions are not seriously violated. Sometimes special experimental techniques, such as the use of guard rows or discard rows, are necessary to achieve uncorrelatedness of the errors. In addition to that randomisation is beneficial to reduce the correlation among neighbouring plots. The randomisation procedure should effect that the probability of assignment to variety i ($i = 1, \ldots, t$) is the same for each plot. This probability is equal to $\Sigma_j\, n_{ij}/n$ for variety i. In order to get a randomisation procedure that satisfies this probability requirement, the blocks are first randomised. For trials with blocks of equal size (so-called proper designs), this randomisation can also be accomplished by assigning a set of varieties (varieties which should remain together in a block) randomly to a block (Mead, 1988). The second stage of the randomisation procedure consists of randomly assigning the plots to the varieties included in that block, according to the design (see also **4.2**). Whatever the complexity of the design, in the end we have n plots over which the varieties are randomly distributed.

For an experiment with blocks it is difficult to test the assumption of a common variance. Visual inspection of estimated errors (called residuals) plotted against estimated expectation values of Y_{ijk} can reveal heterogeneity of variances. If this scatter plot shows a certain pattern, this is a warning that the variances are not equal. For completely randomised designs, which have no blocks, the homogeneity assumption can, for Normally distributed errors, be tested by Bartlett's test

(Bartlett, 1937). However, this test relies heavily on the Normality assumption. If the Normality assumption is violated, the equal variances hypothesis may be rejected even if the variances are indeed equal. Although we will use the Normality assumption in chapter **4**, it is not needed for estimation. A test for homogeneity of variances that is much more robust against deviations from a Normal distribution of observations, is the test of Levene (Levene, 1960). This test can also be used for experiments with blocks. The test of Levene is based on an analysis of variance of the absolute values of the residuals.

The value of variety i ($i = 1, \ldots, t$) can be defined as a weighted average of the expectation of Y_{ijk} over all j and k. So, with the fixed additive model a variety value is defined as a weighted average of $E[Y_{ijk}] = \lambda + \tau_i + \beta_j$ over all j. If equal weights are chosen, the value of variety i is defined in terms of the parameters as

$$\lambda + \tau_i + \overline{\beta}. \quad , \text{ with } \overline{\beta}. = \frac{1}{b} \sum_{j=1}^{b} \beta_j \ .$$

Because a variety value is an average of cell expectations in the variety × block scheme, it is estimable. Because of the overparameterisation in the model the variety parameter τ_i itself is not estimable. However, in order to compare the varieties it is sufficient to estimate differences between variety parameters. These differences are equal to those between the variety values. Hence, we are especially interested in the estimation of contrasts between variety parameters. These contrasts are denoted in the sequel by $\mathbf{p}'\tau$, with τ the column vector of variety parameters and $\mathbf{p}$ a column vector with $\mathbf{p}'\mathbf{1}_t = 0$. Using model (3.1), all contrasts between variety parameters are estimable if the design is connected (Dey, 1986).

Remark

Consider the following reparameterisation of model (3.1) : $E[Y_{ijk}] = \lambda^{\#} + \tau_i^{\#} + \beta_j^{\#}$, where $\lambda^{\#} = \lambda + \overline{\tau}. + \overline{\beta}.$ (with $\overline{\tau}. = 1/t \Sigma\, \tau_i$ and $\overline{\beta}. = 1/b \Sigma\, \beta_j$), $\tau_i^{\#} = \tau_i - \overline{\tau}.$ and $\beta_j^{\#} = \beta_j - \overline{\beta}.$. Consequently, $\overline{\tau}^{\#}_{.} = 0$ and $\overline{\beta}^{\#}_{.} = 0$. Now $\tau_i^{\#}$ is called the variety effect or variety deviation, and $\beta_j^{\#}$ is called the block effect. With the above reparameterisation the value of variety i is equal to $\lambda^{\#} + \tau_i^{\#}$. The above reparameterisation can be generalised by replacing $\overline{\tau}.$ by the weighted average $\Sigma\, w_i \tau_i$, with $\Sigma\, w_i = 1$, and $\overline{\beta}.$ by the weighted average $\Sigma\, v_j \beta_j$, with $\Sigma\, v_j = 1$. For instance, the Kuiper-Corsten iterative method of finding a solution of the normal equations (see also *3.3.1*) gives estimates which correspond to a reparameterisation with $w_i = \Sigma_j\, n_{ij}/n$ and $v_j = \Sigma_i\, n_{ij}/n$.

Let $\mathbf{X}$ be the design matrix of the variety parameters and let $\mathbf{Z}$ be the design matrix of the block parameters. These two matrices show for each observation to which variety and block, respectively, it belongs. The incidence matrix $\mathbf{N}$ shows how often variety i ($i = 1, \ldots, t$) occurs in block j ($j = 1, \ldots, b$) and can be calculated as $\mathbf{X}'\mathbf{Z}$. The rank of $\mathbf{X}$ is equal to t and the rank of $\mathbf{Z}$ is equal to b. In the Euclidean space $\mathcal{R}^n$ the expectation subspace is spanned by the columns of $\mathbf{X}$ and $\mathbf{Z}$. If a design is connected, then the intersection of the subspace spanned by the columns of $\mathbf{X}$ and the subspace spanned by the columns of $\mathbf{Z}$ is the subspace spanned by the unit vector $\mathbf{1}_n$ (Corsten, 1976). The vector of block parameters is denoted by β. Then, given model (3.1), the expectation of the observations at a certain site can be written in matrix notation as

$$
\begin{aligned}
E[\mathbf{Y}] &= \mathbf{1}_n\lambda + \mathbf{X}\tau + \mathbf{Z}\beta \\
&= \mathbf{R}_\mathbf{Z}\mathbf{X}\tau + \mathbf{P}_\mathbf{Z}\mathbf{X}\tau + \mathbf{Z}\beta + \mathbf{1}_n\lambda, \quad \text{with } \mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \text{ and } \mathbf{R}_\mathbf{Z} = \mathbf{I}_n - \mathbf{P}_\mathbf{Z}, \\
&= \mathbf{R}_\mathbf{Z}\mathbf{X}\tau + \mathbf{Z}\beta^*, \quad \text{with } \beta^* = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\tau + \beta + \mathbf{1}_b\lambda,
\end{aligned}
\tag{3.2}
$$

because $\mathbf{1}_n = \mathbf{Z}\mathbf{1}_b$. The matrix $\mathbf{P}_\mathbf{Z}$ denotes the orthogonal projection on the subspace spanned by the columns of $\mathbf{Z}$. Because of the orthogonalisation in (3.2) the normal equations have the following simple form :

$$
\begin{pmatrix} \mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}'\mathbf{Z} \end{pmatrix}
\begin{pmatrix} \tilde{\tau} \\ \beta^* \end{pmatrix}
=
\begin{pmatrix} \mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{pmatrix}.
$$

(The sub-matrix $\mathbf{0}$ denotes a null submatrix of the appropriate size.)
Hence a solution of the least squares estimates of β^*, denoted by $\tilde{\beta}^*$, is $(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. Further a solution of the least squares estimates of τ, denoted by $\tilde{\tau}$, can be calculated as

$$
\tilde{\tau} = (\mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{Y}, \tag{3.3}
$$

where $\mathbf{A}^-$ denotes a pseudo-inverse of $\mathbf{A}$, which has the property that $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$. For later use we choose a pseudo-inverse that satisfies the additional conditions $\mathbf{A}^- = (\mathbf{A}^-)'$ and $\mathbf{A}^-\mathbf{A} = (\mathbf{A}^-\mathbf{A})'$. A different choice of $\mathbf{A}^-$ gives a different solution of $\tilde{\tau}$, but contrasts between the estimates which correspond to contrasts between the variety parameters are unique (Dey, 1986). We now define the pseudo-variance/covariance matrix (also called pseudo-dispersion matrix) of $\tilde{\tau}$, say $\dot{\mathbf{D}}[\tilde{\tau}]$, as the matrix which can be interpreted as the variance/covariance matrix of $\tilde{\tau}$ if it is used for the calculation of variances and covariances of estimators corresponding to estimable functions of variety parameters, e.g. contrasts $\mathbf{p}'\tilde{\tau}$. Here $\dot{\mathbf{D}}[\tilde{\tau}] = (\mathbf{X}'\mathbf{R}_Z\mathbf{X})^-\sigma^2$, and the variance of $\mathbf{p}'\tilde{\tau}$ is equal to $\text{var}(\mathbf{p}'\tilde{\tau}) = \mathbf{p}'(\mathbf{X}'\mathbf{R}_Z\mathbf{X})^-\mathbf{p}\sigma^2$.

Solution (3.3) is identical with the solution of the well known so-called reduced normal equations for treatment parameters, often written as $\mathbf{C}\tilde{\tau} = \mathbf{Q}$. The matrix $\mathbf{C}$ is calculated as $\mathbf{C} = \mathbf{R} - \mathbf{N}\mathbf{K}^{-1}\mathbf{N}' = \mathbf{X}'\mathbf{R}_Z\mathbf{X}$ and the vector $\mathbf{Q}$ as $\mathbf{Q} = \mathbf{T} - \mathbf{N}\mathbf{K}^{-1}\mathbf{B} = \mathbf{X}'\mathbf{R}_Z\mathbf{Y}$, where $\mathbf{R} = \mathbf{X}'\mathbf{X}$ denotes the diagonal replication matrix, $\mathbf{K} = \mathbf{Z}'\mathbf{Z}$ is the diagonal block size matrix, $\mathbf{N} = \mathbf{X}'\mathbf{Z}$ is the incidence matrix, $\mathbf{T} = \mathbf{X}'\mathbf{Y}$ is the vector with treatment totals and $\mathbf{B} = \mathbf{Z}'\mathbf{Y}$ is the vector with block totals. A solution of the reduced normal equations for treatment parameters can now be written as $\tilde{\tau} = \mathbf{C}^-\mathbf{Q}$, hence the contrast $\mathbf{p}'\tau$ can be estimated as $\mathbf{p}'\tilde{\tau} = \mathbf{p}'\mathbf{C}^-\mathbf{Q}$. This is the best linear unbiased estimator (BLUE), or Gauss-Markov estimator, of $\mathbf{p}'\tau$ (John, 1971). In reduced normal equation notation, the pseudo-variance/covariance matrix of $\tilde{\tau}$ is equal to $\dot{\mathbf{D}}[\tilde{\tau}] = \mathbf{C}^-\sigma^2$.

The error variance can be estimated by the Sum of Squares for Error (SS_E) divided by the degrees of freedom for error (df_e). The SS_E can be calculated as the sum of squares of the residuals. The residuals are the estimated errors ($\tilde{E}_{ijk}$) and the vector of residuals is denoted by $\tilde{\mathbf{E}}$. It is calculated by subtracting the estimated expectation of $\mathbf{Y}$ from $\mathbf{Y}$ itself :

$$\begin{aligned}
\tilde{\mathbf{E}} &= \mathbf{Y} - \mathbf{R}_Z\mathbf{X}\tilde{\tau} - \mathbf{Z}\beta^* \\
&= \mathbf{Y} - \mathbf{R}_Z\mathbf{X}\tilde{\tau} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \\
&= \mathbf{R}_Z(\mathbf{Y} - \mathbf{X}\tilde{\tau})\,.
\end{aligned}$$

The value of $\tilde{\mathbf{E}}$ does not depend on the specific choice of the pseudo-inverse of $\mathbf{X}'\mathbf{R}_Z\mathbf{X}$, as does the value of $\tilde{\tau}$. Notice that it is not necessary to estimate the block parameters in order to calculate the residuals. Now $SS_E = \tilde{\mathbf{E}}'\tilde{\mathbf{E}}$. For a connected design, the degrees of freedom for error can be calculated as $df_e = n - t - b + 1$, which is the dimension of the observation space minus the dimension of the expectation subspace. The latter dimension is equal to $t + b - 1$, because the intersection of the subspace spanned by the columns of $\mathbf{X}$ and the subspace spanned by the columns of $\mathbf{Z}$ has dimension 1, if the design is connected. The estimate of the error variance, denoted by s^2, is equal to $\tilde{\mathbf{E}}'\tilde{\mathbf{E}}/(n - t - b + 1)$.

3.1.2 The mixed additive model

If we use a mixed additive model with fixed variety contributions we assume that the blocks represent a random sample from all possible blocks in the experimental field. We then can write the model for the observations as

$$Y_{ijk} = \lambda + \tau_i + B_j + E_{ijk}\,, \qquad i = 1,\ldots,t\,,\ j = 1,\ldots,b\,,\ k = 1,\ldots,n_{ij}\,. \tag{3.4}$$

Besides the error terms, now also the block terms are uncorrelated random variables. The latter have zero expectation and common variance σ_B^2. Furthermore we assume that $\text{cov}(B_j, E_{ijk}) = 0$ for all j and i,j,k. The expectation of Y_{ijk} is equal to $\lambda+\tau_i$, and therefore the value of variety i is defined as $\lambda+\tau_i$. The vector of observations **Y** contains random variables with the variance/covariance matrix :

$$\begin{aligned}\mathbf{D}[\mathbf{Y}] &= \mathbf{ZZ}'\sigma_B^2 + \mathbf{I}_n\sigma^2 \\ &= \mathbf{V}\sigma^2\,, \quad \text{with } \mathbf{V} = \mathbf{ZZ}'\frac{\sigma_B^2}{\sigma^2} + \mathbf{I}_n\,.\end{aligned}$$

In matrix notation, the expectation of **Y** can be written as

$$\begin{aligned}E[\mathbf{Y}] &= \mathbf{1}_n\lambda + \mathbf{X}\tau \\ &= \boldsymbol{R}_1\mathbf{X}\tau + \boldsymbol{P}_1\mathbf{X}\tau + \mathbf{1}_n\lambda\,, \quad \text{with } \boldsymbol{P}_1 = \mathbf{1}_n(\mathbf{1}'_n\mathbf{V}^{-1}\mathbf{1}_n)^{-1}\mathbf{1}'_n\mathbf{V}^{-1} \text{ and } \boldsymbol{R}_1 = \mathbf{I}_n - \boldsymbol{P}_1\,, \\ &= \boldsymbol{R}_1\mathbf{X}\tau + \mathbf{1}_n\lambda^*\,, \quad \text{with } \lambda^* = (\mathbf{1}'_n\mathbf{V}^{-1}\mathbf{1}_n)^{-1}\mathbf{1}'_n\mathbf{V}^{-1}\mathbf{X}\tau + \lambda\,.\end{aligned} \tag{3.5}$$

$\boldsymbol{P}_1$ denotes a projection on the subspace spanned by the unit vector. The normal equations are now equal to :

$$\begin{pmatrix}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_n\mathbf{V}^{-1}\mathbf{1}_n\end{pmatrix}\begin{pmatrix}\tilde{\tau} \\ \tilde{\lambda}^*\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{Y} \\ \mathbf{1}'_n\mathbf{V}^{-1}\mathbf{Y}\end{pmatrix}.$$

A (generalised least squares) solution of the normal equations for $\tilde{\tau}$ can be calculated as

$$\tilde{\tau} = (\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{Y}\,, \tag{3.6}$$

with pseudo-variance/covariance matrix $\dot{\mathbf{D}}[\tilde{\tau}] = (\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{X})^{-}\sigma^2$. The matrix **V** is a $n\times n$ matrix. The dimension of this matrix often causes difficulties in the calculation of its inverse. However, the inverse of **V** can be written as $\mathbf{V}^{-1} = \mathbf{I}_n - \mathbf{Z}((\sigma^2/\sigma_B^2)\mathbf{I}_b + \mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. Now the matrix that has to be inverted is only of dimension $b \times b$.

If we subtract the estimated expectation of **Y** from **Y** itself we obtain the vector of residuals, denoted here by $\tilde{\mathbf{F}}$:

$\tilde{\mathbf{F}} = \boldsymbol{R}_1(\mathbf{Y} - \mathbf{X}\tilde{\tau})$.

With the variance ratio σ_B^2/σ^2 assumed known, hence **V** known, the degrees of freedom for error in model (3.4) are equal to $df_f = n - t$. Then σ^2 can be estimated by $\tilde{\mathbf{F}}'\tilde{\mathbf{F}}/df_f$.

Remark

With model (3.4) the vector with variety value estimators has a very convenient form : $\tilde{\tau} + \mathbf{1}_t\tilde{\lambda} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$. Therefore in practice these estimators are often used. Of course contrasts between variety values are identical with contrasts between variety parameters. The approach above this remark has been chosen for later use.

The estimation procedure assumes that we *a priori* know the ratio σ_B^2/σ^2 in the matrix **V**. The plant breeding practice usually has a long history of similar experiments performed in the past years, hence the breeder often has a fairly good idea about the variance ratio. For estimation of variance components from prior experiments the reader is referred to Verdooren (1988). It would be wise to estimate a lower bound and an upper bound for the variance ratio σ_B^2/σ^2, and to calculate the best estimates of contrasts between variety parameters for several values of the ratio within the range between these bounds. If the ranking of the estimates changes for different ratios, we have to be very cautious in making decisions. If there is no long history of similar experiments to get a good guess of σ_B^2/σ^2, then the plant breeder might decide to estimate the variance ratio not from prior experiments but from the current experiment. If the estimates of the variance components are obtained by the ANOVA procedure and the estimate of σ_B^2/σ^2 is used in **V**, the estimators of contrasts between variety parameters are equal to the combined intra- and inter-block estimators introduced by Yates (1940). The use of inter-block estimates is called recovery of inter-block information. Using other estimators for the variance components and inserting these estimates in σ_B^2/σ^2 in **V** gives other methods of recovery of inter-block information (see Verdooren, 1989). However, it is questionable whether procedures which use an estimator for

σ_B^2/σ^2 lead towards a combined estimator that is better than the ordinary least squares estimator described in *3.1.1*, for the very reason that an estimator for the variance ratio is used.

If the ratio σ_B^2/σ^2 approaches infinity, then $\mathbf{V}^{-1} = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \mathbf{R}_\mathbf{Z}$. Then $(\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_1\mathbf{Y} = (\mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{Y}$ because $\mathbf{V}^{-1}\boldsymbol{R}_1 = \mathbf{R}_\mathbf{Z}\boldsymbol{R}_1 = \mathbf{R}_\mathbf{Z}(\mathbf{I}_n - \boldsymbol{P}_1) = \mathbf{R}_\mathbf{Z}$. Then the estimators of τ and the pseudo-variance/covariance matrix of these estimators are the same for the fixed and the mixed model. In practice the ratio σ_B^2/σ^2 is often large, resulting in estimates and a pseudo-variance/covariance matrix with the mixed model which are approximately equal to those obtained with the fixed model. Added to that the extra work and the uncertainty corresponding to the estimation of variance components, the choice is often made to use the fixed model. The large ratio of σ_B^2/σ^2 found in practice is likely to be caused by the fact that the blocks are often not selected at random, but are specifically chosen in such a way that the variation among the blocks is as large as possible, and the variation within a block as small as possible (Mead, 1988). There are situations for which the estimators of the variety parameters are identical for both the fixed and the mixed model, irrespective of the ratio σ_B^2/σ^2. It has been proven (e.g. Baksalary & Kala, 1983) that this is the case if the subspace spanned by the columns of **VX** is included in the subspace spanned by the columns of **X**. This is e.g. the case if a complete block design is used.

In so-called resolvable designs the blocks can be grouped into r replications or 'super blocks', a group of blocks forming one complete replication of the varieties. In replication j $(j = 1, \ldots, r)$ there are b_j blocks, with $\Sigma\, b_j = b$. The position of the replications in the experimental field is chosen in such a way that the variation between replications is as large as possible. Therefore the replications have to be considered fixed. The blocks within the replications can then be considered random. Examples of resolvable designs are lattice designs and alpha designs. A mixed model for the observations from an experiment with a resolvable block design can be written as

$$Y_{ijk} = \lambda + \tau_i + \rho_j + B_{k(j)} + E_{ijk}\,, \qquad i = 1, \ldots, t\,,\ j = 1, \ldots, r\,,\ k = 1, \ldots, b_j\,. \qquad (3.7)$$

The replication terms (ρ_j) are fixed and the contributions of the blocks within the replications $(B_{k(j)})$ are uncorrelated random variables with zero expectation and common variance σ_B^2. Furthermore it is assumed that $\mathrm{cov}(B_{k(j)}, E_{ijk}) = 0$. We now introduce **M**, the design matrix of the replication parameters with rank r. The variance/covariance matrix of the observations remains equal to $\mathbf{V}\sigma^2$, the variance/covariance matrix for the mixed model without fixed replication parameters. Using the reparameterisation technique described above for a mixed model, similar expressions for $\tilde{\tau}$ and $\dot{\mathbf{D}}[\tilde{\tau}]$ can be found as for the situation without fixed replication parameters. These expressions can be obtained from the latter by replacing $\mathbf{1}_n$ by **M** and λ by ρ, with ρ the column vector of replication parameters. We then find $\tilde{\tau} = (\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_\mathbf{M}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_\mathbf{M}\mathbf{Y}$ and $\dot{\mathbf{D}}[\tilde{\tau}] = (\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{R}_\mathbf{M}\mathbf{X})^{-}\sigma^2$, with $\boldsymbol{R}_\mathbf{M} = \mathbf{I}_n - \mathbf{M}(\mathbf{M}'\mathbf{V}^{-1}\mathbf{M})^{-1}\mathbf{M}'\mathbf{V}^{-1}$. Because of the frequent occurrence of resolvable designs in the plant breeding practice, model (3.7) can often be used in that field. Also if a super block contains more than one complete replication, hence if some varieties appear more than once in a super block, the above is valid. In model (3.7) Y_{ijk} and E_{ijk} are then replaced by Y_{ijkl} and E_{ijkl}, respectively, with $l = 1, \ldots, n_{ij}$ and n_{ij} the number of observations of variety i in super block j.

3.2 Combining estimators from a series of experiments

The selection of superior new varieties for just one site and one year is already a difficult task. The plant breeder has to study numerous characters, most of them masked by micro-environmental noise. All these characters have to be combined into an index, written in black and white or determined intuitively by the breeder looking at the performance of the varieties in the field and combining his impressions; the last index being often referred to as the 'breeder's eye'. The task of selecting the best new variety becomes even more difficult when the results of various sites or years or both have to be combined. The calculation of a 'mean performance' of a variety is very important for the breeder, because the results of just one trial are not sufficient to base the ultimate selection on. Besides the question which index to choose, now other questions become urgent. The plant breeder has to decide which sites to combine, together forming a region of interest for which a separate variety has to be developed. Another important question is whether the model for the observations has to contain fixed or random site contributions, and whether variety × site interaction terms have to be included in this model. Because the various models result in different estimates and ranking of these estimates, the model has to be chosen very carefully. More about model choice is written in *3.2.1*. For the time being assume that there is only one year, so only the results of various sites are combined. The calculation of combined estimates is elaborated in *3.2.2*, *3.2.3*, *3.2.4* and *3.2.5*. In each section a different model is used. In *3.2.6* the traditional analysis of a series of experiments is described and compared with the proposed procedures.

3.2.1 Choosing the type of model

Suppose a plant breeder wants to select a variety for use in the region which is characterised by the chosen sites. Assume there are m sites, and that the trials at these sites have been analysed separately, either by using a fixed additive model or a mixed additive model. If the breeder does not have much experience with the crop, it is sensible to first compare the m error variances through their estimates. If the error variances cannot be considered equal for all sites, combining the results from the various sites is hazardous. The interpretation of the set of trials should then be based on the individual analyses of the data from the separate experiments (Mead, 1988). The hypothesis that the m error variances are equal can be tested with the test of Bartlett (1937), if the errors are Normally distributed, or Levene (1960). If the error variances can be considered equal for all sites, the observations from all the trials together can be described by a single linear model of the usual form. An additive model includes terms for varieties, blocks within sites, sites and errors. In addition to that an interaction model contains terms for the variety × site interactions. An important question is whether the site terms and the interaction terms should be considered random or fixed. This question is inevitably connected with the purpose of the series of experiments and the way the sampling of the sites has been done.

Assume that the region of interest for which a breeder wants to develop a variety is more or less known. Then a number of sites within this region has to be chosen. One option is to specifically select sites that represent the different environmental conditions of the region. If, for instance, there are three soil types in the region, we may on purpose select a site with soil type 1, a site with soil type 2 and a site with soil type 3. If we choose the sites in such a way, and assume that the other sources of variation which are present at the site level (e.g. farmer's skill, rainfall, sunshine) are neglectable with respect to soil type, we have to see the site terms as being fixed. The results of the analysis apply to this group of particular sites only. If the goal of the breeder is to select varieties which are the best, averaged over the sites actually used, a model with fixed site terms is appropriate. The variety

× site interactions are also fixed terms in this situation. Mead (1988) strongly advocates the method of selecting sites in a non-random way. He is of the opinion that a breeder should try to characterise the major differences within the region and select sites that span these characteristics.

The other option is to choose the sites fully at random, hence creating a random sample from the population of sites. The situation of truly random sampling of the sites will never occur in the plant breeding practice. However, usually the sample is considered effectively random. The inferences made after the analysis of the data apply to the whole population of sites, which represents the region. This type of inference is of great importance to the plant breeder, because he wants to develop varieties that are superior in the region, and not specifically at the sites actually used. It is much more difficult to make inferences about a variety value for the population of sites, compared with the situation of fixed sites. If an interaction model is used, the variance of the estimator of the difference between two variety values increases because the interaction component of variance is added. Because the variance increases, it is more difficult to select the best variety. This is not surprising, because the question to which we want an answer is more difficult.

Consider again the above mentioned example of a region with three soil types. Often, not only one site but a number of sites with the same soil type are chosen. The differences between sites with different soil types are large because of the purposeful choice of the sites. Hence we have to classify the groups of sites with the same soil type as fixed. The soil types can be introduced as a new fixed factor. The sites within each group with a particular soil type can be seen as a random sample of sites. Thus a third option is to extend the basic model with random site terms by introducing fixed group terms. Because situations similar to the soil type example frequently occur in the plant breeding practice, a model with both fixed and random site terms is of paramount importance there.

The comparison of the individual analyses of the separate trials often reveals whether interaction is present. Then there are two possibilities. First, the plant breeder can decide that the region has to be redefined. By subdividing the region into smaller regions with sites that are more alike, the additive model within each subregion may be used. But the development of separate varieties for each subregion is often not desirable because of economical drawbacks. The second possibility is to use an interaction model.

We will use fixed interaction terms if the site terms have been chosen fixed, and random interaction terms if the sites are random. In the mixed interaction models we will use, random interaction terms have a common variance. Although common interaction variances may not be very realistic, models which allow unequal interaction variances are not at all convenient to work with in practice (Patterson & Silvey, 1980).

3.2.2 The fixed additive model

It is often not feasible to analyse the trials at the various sites as one large experiment. In practice, (contrasts between) variety values are estimated first for each individual trial. Next the estimates from the various sites are combined. This has to be done in such a way that the ultimate estimates are identical with the outcomes of the best linear unbiased estimators (BLUEs), had these been calculated using a model for the joint observations of the various sites, thus analysing the large experiment as a whole. Again we assume that the design is connected. First suppose that all varieties are present at all sites. Assume that the t varieties are grown at site k ($k = 1, \ldots, m$) in an experiment with an (in)complete block design with b_k blocks, and that variety i ($i = 1, \ldots, t$) has $n_{ij(k)}$ observations in block j at site k. Let n denote the total number of observations. The additive model for the joint observations can be written as

$$Y_{ijkl} = \mu + \tau_i + \beta_{j(k)} + \lambda_k + E_{ijkl} \,, \tag{3.8}$$

$$i = 1, \ldots, t \;,\; j = 1, \ldots, b_k \;,\; k = 1, \ldots, m \;,\; l = 1, \ldots, n_{ij(k)} \,,$$

where μ is the general level parameter, τ_i the variety parameter for variety i, $\beta_{j(k)}$ the block parameter for the j^{th} block at site k, λ_k the site parameter for site k and E_{ijkl} the plot error. The errors are assumed to be uncorrelated random variables with expectation zero and common variance σ^2. We define a variety value in terms of the parameters as

$$\mu + \tau_i + \sum_{k=1}^{m} w_k(\overline{\beta}_{\cdot(k)} + \lambda_k) \,, \quad \text{with } \overline{\beta}_{\cdot(k)} = \frac{1}{b_k} \sum_{j=1}^{b_k} \beta_{j(k)} \,,$$

where it is assumed that all blocks at a certain site have equal weight $1/b_k$ and that site k has weight w_k, with $\Sigma\, w_k = 1$. The weights have to be chosen by the plant breeder. If all sites are of the same importance, it is logical to use equal weights

$w_k = 1/m$. With the fixed model, a variety value can be seen as a weighted average of the separate variety values from the sites *actually* used. A contrast between variety values is equal to that contrast between variety parameters and does not depend on the weights chosen. The parameters μ, $\beta_{j(k)}$, λ_k can be replaced by a single parameter. In fact, $\mu + \beta_{j(k)} + \lambda_k$ is equivalent to $\lambda + \beta_j$ of the fixed model at the k^{th} site level.

We will now change over to matrix notation. If we look back at the separate analyses at the individual sites, we notice that the vector of variety parameters τ is equal for each site and is not associated with a particular site. The vectors of block parameters are different for all sites, and will be denoted by β_k ($k = 1, \ldots, m$). Because all the varieties are present at a certain site, all $\mathbf{X}_k$ have the same number of columns. Using the results of **3.1**, the expectation of Y_{ijkl} in model (3.8) can be replaced by

$$E\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{\mathbf{Z}_1}\mathbf{X}_1 \\ \mathbf{R}_{\mathbf{Z}_2}\mathbf{X}_2 \\ \vdots \\ \mathbf{R}_{\mathbf{Z}_m}\mathbf{X}_m \end{pmatrix}\tau + \begin{pmatrix} \mathbf{Z}_1 & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{Z}_2 & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{Z}_m \end{pmatrix}\begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_m^* \end{pmatrix}, \tag{3.9}$$

with $\beta_k^* = (\mathbf{Z}'_k\mathbf{Z}_k)^{-1}\mathbf{Z}'_k\mathbf{X}_k\tau + \beta_k + \mathbf{1}_{b_k}(\lambda_k + \mu)$.

(The symbol **-0-** means that all the empty positions in the matrix are filled with zeros)

The normal equations then become :

$$\begin{pmatrix} \sum_{k=1}^{m} \mathbf{X}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{X}_k & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{Z}'_1\mathbf{Z}_1 & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{Z}'_m\mathbf{Z}_m \end{pmatrix}\begin{pmatrix} \hat{\tau} \\ \beta_1^* \\ \vdots \\ \beta_m^* \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{m} \mathbf{X}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{Y}_k \\ \mathbf{Z}'_1\mathbf{Y}_1 \\ \vdots \\ \mathbf{Z}'_m\mathbf{Y}_m \end{pmatrix},$$

or in reduced normal equation notation :

$$\begin{pmatrix} \sum_{k=1}^{m} \mathbf{C}_k & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{K}_1 & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{K}_m \end{pmatrix}\begin{pmatrix} \hat{\tau} \\ \beta_1^* \\ \vdots \\ \beta_m^* \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{m} \mathbf{Q}_k \\ \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_m \end{pmatrix}.$$

In the sequel $\Sigma\, \mathbf{C}_k$ will be denoted by $\mathbf{C}.$ for simplicity. A solution of the normal equations for $\hat{\tau}$ can now be calculated as

$$\hat{\tau} = \mathbf{C}_{.}^{-} \sum_{k=1}^{m} \mathbf{Q}_k = \sum_{k=1}^{m} \mathbf{C}_{.}^{-}\mathbf{C}_k\mathbf{C}_k^{-}\mathbf{Q}_k \qquad \text{, because } \mathbf{C}_k\mathbf{C}_k^{-}\mathbf{Q}_k = \mathbf{Q}_k \text{ ,}$$

$$= \sum_{k=1}^{m} \mathbf{W}_k\tilde{\tau}_k \qquad \text{, with } \mathbf{W}_k = \mathbf{C}_{.}^{-}\mathbf{C}_k \text{ .} \tag{3.10}$$

The pseudo-variance/covariance matrix of $\hat{\tau}$ can easily be calculated as $\dot{\mathbf{D}}[\hat{\tau}] = \mathbf{C}_{.}^{-}\sigma^2$.

From (3.10) we see that the BLUE of $\mathbf{p}'\tau$ is equal to $\mathbf{p}'\hat{\tau} = \mathbf{p}'\Sigma\, \mathbf{W}_k\tilde{\tau}_k$. Notice that in general this is not a univariately weighted average of the BLUEs $\mathbf{p}'\tilde{\tau}_k$ at the individual sites, but that a multivariately weighted average is required. By 'univariately weighted' it is meant that $\mathbf{p}'\hat{\tau}$ can be calculated as a weighted average of the estimators of $\mathbf{p}'\tau$ at the individual sites. By 'multivariately weighted' it is meant that for the calculation of $\mathbf{p}'\hat{\tau}$ also estimators of other contrasts than $\mathbf{p}'\tau$ (at the individual sites) are used for the weighted average. It is well known that univariate weights, say w_k, should be chosen inversely proportional to the variances of the $\mathbf{p}'\tilde{\tau}_k$, with $\Sigma\, w_k = 1$. The multivariate weights in the weight matrices $\mathbf{W}_k$ satisfy similar conditions :

$$\mathbf{W}_k = \mathbf{C}_{.}^{-}\mathbf{C}_k \text{ ,}$$

and

$$\sum_{k=1}^{m} \mathbf{W}_k = \mathbf{C}_{.}^{-}\mathbf{C}_{.} = \mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}'_t \text{ , so } \mathbf{p}'\left(\sum_{k=1}^{m} \mathbf{W}_k\right)\mathbf{p} = \mathbf{p}'\mathbf{I}_t\mathbf{p} \text{ .}$$

The second condition is developed as follows : Because the individual experiments at the various sites have connected designs, the only dependence between the rows of $\mathbf{C}_k$ is reflected by $\mathbf{C}_k\mathbf{1}_t = \mathbf{0}$, hence $\mathbf{C}_{.}^{-}\mathbf{C}_{.}\mathbf{1}_t = \mathbf{0}$. Matrix $\mathbf{C}_{.}^{-}\mathbf{C}_{.}$, which is chosen symmetric (see the characteristics of the pseudo-inverses chosen in *3.1.1*), is a projection matrix since $(\mathbf{C}_{.}^{-}\mathbf{C}_{.})(\mathbf{C}_{.}^{-}\mathbf{C}_{.}) = \mathbf{C}_{.}^{-}\mathbf{C}_{.}$, and all idempotent symmetric matrices are projection matrices. Because $\mathbf{C}_{.}^{-}\mathbf{C}_{.}$ is orthogonal to $\mathbf{1}_t$, it must be equal to $\mathbf{I}_t - (1/t)\mathbf{1}_t\mathbf{1}'_t$.

The multivariate weights reduce to the univariate weights w_k if and only if

$$\mathbf{p}'\mathbf{W}_k = w_k\mathbf{p}' \text{ , so } \mathbf{W}'_k\mathbf{p} = w_k\mathbf{p} \text{ ,}$$

hence if $\mathbf{p}$ is a common eigenvector of all $\mathbf{W}'_k$ matrices with corresponding eigenvalue w_k. Notice that the above condition can be satisfied for contrast $\mathbf{p}'\boldsymbol{\tau}$ with eigenvalues $w_{\mathbf{p}k}$ and for a different contrast $\mathbf{q}'\boldsymbol{\tau}$ with eigenvalues $w_{\mathbf{q}k}$. However, if the best estimates for linear combinations of $\mathbf{p}'\boldsymbol{\tau}$ and $\mathbf{q}'\boldsymbol{\tau}$ have to be equal to the linear combinations of $\Sigma\, w_{\mathbf{p}k}\mathbf{p}'\tilde{\boldsymbol{\tau}}_k$ and $\Sigma\, w_{\mathbf{q}k}\mathbf{q}'\tilde{\boldsymbol{\tau}}_k$, then $w_{\mathbf{p}k}$ has to be equal to $w_{\mathbf{q}k}$ for each k ($k = 1, \ldots, m$). This is for instance the case if the $\mathbf{C}_k$ matrices are identical. The common eigenvalue of the $\mathbf{W}'_k$ must then be $1/m$ in order to sum up to unity.

If the $\mathbf{C}_k$ matrices are proportional to each other, they can be written as $\mathbf{C}_k = d_k\mathbf{C}_.$, with $\Sigma\, d_k = 1$. Then

$$\mathbf{W}_k = \mathbf{C}_.^-\mathbf{C}_k = d_k\mathbf{C}_.^-\mathbf{C}_. = d_k\left(\mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}'_t\right),$$

hence $\mathbf{p}'\hat{\boldsymbol{\tau}} = \mathbf{p}'\Sigma\,\mathbf{W}_k\tilde{\boldsymbol{\tau}}_k = \Sigma\, d_k\mathbf{p}'\tilde{\boldsymbol{\tau}}_k$, since $\mathbf{p}'(\mathbf{I}_t - (1/t)\mathbf{1}_t\mathbf{1}'_t) = \mathbf{p}'$. So $d_k = w_k$, the univariate weight and eigenvalue of $\mathbf{W}'_k = \mathbf{W}_k$ with eigenvector $\mathbf{p}$. Because d_k is a value independent of $\mathbf{p}$, the $\mathbf{W}_k$ must have an eigenvalue w_k with multiplicity $t-1$ and one zero eigenvalue. This is the case if all the experiments have variance-balanced designs. In a variance-balanced design the least squares estimators of all pairwise contrasts between two variety parameters have the same variance. If all the experiments have variance-balanced designs, then all the non-zero eigenvalues of $\mathbf{C}_k$ are equal, say θ_k, and $\mathbf{C}_k$ can be written as $\mathbf{C}_k = \theta_k(\mathbf{I}_t - (1/t)\mathbf{1}_t\mathbf{1}'_t)$ (Dey, 1986). In that case $\mathbf{W}_k = (\theta_k/\Sigma\,\theta_k)\,(\mathbf{I}_t - (1/t)\mathbf{1}_t\mathbf{1}'_t)$ and $w_k = \theta_k/\Sigma\,\theta_k$. Examples of variance-balanced designs are an equireplicated completely randomised design (E-CRD), a randomised complete block design (RCBD) and a balanced incomplete block design (BIBD). The $\mathbf{C}$ matrices of these designs are proportional to each other since they can be written as

$$\text{E-CRD}: \mathbf{C} = r\left(\mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}'_t\right),$$

$$\text{RCBD}: \mathbf{C} = b\left(\mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}'_t\right),$$

$$\text{BIBD}: \mathbf{C} = \frac{b(k-1)}{(t-1)}\left(\mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}'_t\right) = \frac{\lambda t}{k}\left(\mathbf{I}_t - \frac{1}{t}\mathbf{1}_t\mathbf{1}'_t\right),$$

with r the number of replications, t the number of varieties, b the number of blocks, k the block size and λ the parameter that indicates how often each pair of treatments appears in the same block (and for a BIBD $\lambda = r(k-1)/(t-1)$). So, if the trials at the various sites all have an E-CRD, a RCBD or a BIBD, then we can use univariate weights.

When the sites do not contain exactly the same set of varieties, the variety × site table becomes incomplete. However, this gives no problem in the analysis. To calculate the weight matrices, rows and columns with zeros have to be included in the $\mathbf{C}_k$ matrices at the positions that correspond to the varieties which are not present at site k; thereby creating $\mathbf{C}_k$ matrices of the same size. In order to calculate the outcomes of the BLUEs of $\mathbf{p}'\tau$ as $\Sigma\, \mathbf{p}'\, \mathbf{W}_k\tilde{\tau}_k$, null elements have to be included also in the $\tilde{\tau}_k$ vectors at the appropriate places.

To estimate the error variance we first have to determine the residuals. Analogous to the vector of observations, the vector of residuals $\hat{\mathbf{E}}$ can be subdivided into vectors $\hat{\mathbf{E}}_k$ $(k = 1,\ldots,m)$, with $\hat{\mathbf{E}}' = (\hat{\mathbf{E}}_1' \;\; \hat{\mathbf{E}}_2' \;\; \ldots \;\; \hat{\mathbf{E}}_m')$. For every k $(k = 1,\ldots,m)$ $\hat{\mathbf{E}}_k$ can be calculated as

$$\begin{aligned}\hat{\mathbf{E}}_k &= \mathbf{Y}_k - \mathbf{R}_{\mathbf{Z}_k}\mathbf{X}_k\hat{\tau} - \mathbf{Z}_k\beta_k^* \\ &= \mathbf{Y}_k - \mathbf{R}_{\mathbf{Z}_k}\mathbf{X}_k\hat{\tau} - \mathbf{Z}_k(\mathbf{Z}'_k\mathbf{Z}_k)^{-1}\mathbf{Z}'_k\mathbf{Y}_k \\ &= \mathbf{R}_{\mathbf{Z}_k}(\mathbf{Y}_k - \mathbf{X}_k\hat{\tau})\,.\end{aligned}$$

Notice that these residuals are not equal to the residuals calculated at the separate trials, because in general $\hat{\tau} \neq \tilde{\tau}_k$. The SS_E can now be calculated as $SS_E = \Sigma\, \hat{\mathbf{E}}_k'\hat{\mathbf{E}}_k$. The degrees of freedom for error can be calculated as $df_e = n - t - \Sigma\, b_k + 1$. Then σ^2 can be estimated as SS_E/df_e.

Examples

We wil demonstrate the estimation of $\mathbf{p}'\tau$ in three small examples. For each example the incidence scheme is given. Notation : S1 = site 1, B1 = block 1, V1 = variety 1, and so on. For the calculation of the weight matrices a Moore-Penrose pseudo-inverse is used. In addition to the characteristics of the pseudo-inverses chosen in *3.1.1*, the Moore-Penrose pseudo-inverse $\mathbf{A}^+$ of matrix $\mathbf{A}$ also satisfies $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$ and $\mathbf{A}\mathbf{A}^+ = (\mathbf{A}\mathbf{A}^+)'$. The Moore-Penrose inverse $\mathbf{A}^+$ of $\mathbf{A}$ is unique.

Example 1.

	S1		S2			S3			
	B1	B2	B3	B4	B5	B6	B7	B8	B9
V1	1	1	1	1	1	1	0	1	0
V2	0	0	1	1	1	1	0	0	1
V3	1	1	1	1	1	0	1	1	0
V4	1	1	0	0	0	0	1	0	1

$$\mathbf{C}_1=\frac{1}{3}\begin{pmatrix} 4 & 0 & -2 & -2 \\ 0 & 0 & 0 & 0 \\ -2 & 0 & 4 & -2 \\ -2 & 0 & -2 & 4 \end{pmatrix}, \quad \mathbf{W}_1=\begin{pmatrix} 0.2538 & 0 & -0.0615 & -0.1923 \\ 0.0615 & 0 & 0.0154 & -0.0769 \\ -0.0769 & 0 & 0.2308 & -0.1538 \\ -0.2385 & 0 & -0.1846 & 0.4231 \end{pmatrix}.$$

$$\mathbf{C}_2=\begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{W}_2=\begin{pmatrix} 0.3308 & -0.1885 & -0.1423 & 0 \\ -0.2077 & 0.4846 & -0.2769 & 0 \\ -0.1154 & -0.2308 & 0.3462 & 0 \\ -0.0077 & -0.0654 & 0.0731 & 0 \end{pmatrix}.$$

$$\mathbf{C}_3=\frac{1}{2}\begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}, \quad \mathbf{W}_3=\begin{pmatrix} 0.1654 & -0.0615 & -0.0462 & -0.0577 \\ -0.1038 & 0.2654 & 0.0115 & -0.1731 \\ -0.0577 & -0.0192 & 0.1731 & -0.0962 \\ -0.0038 & -0.1846 & -0.1385 & 0.3269 \end{pmatrix}.$$

Let $\mathbf{p}'=(1 \quad -1 \quad 0 \quad 0)$. Then

$$\mathbf{p}'\mathbf{W}_1=(0.1923 \quad 0.0000 \quad -0.0769 \quad -0.1154),$$

$$\mathbf{p}'\mathbf{W}_2=(0.5385 \quad -0.6731 \quad 0.1346 \quad 0.0000),$$

$$\mathbf{p}'\mathbf{W}_3=(0.2692 \quad -0.3269 \quad -0.0577 \quad 0.1154).$$

Notice that $\mathbf{p}'\mathbf{W}_k\tau$ at site k represents a contrast completely different than $\mathbf{p}'\tau$. So in this example multivariate weights are necessary.

Example 2.

	S1		S2		S3					
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
V1	1	1	1	1	1	0	1	0	1	0
V2	1	1	1	1	1	0	0	1	0	1
V3	1	1	1	1	0	1	1	0	0	1
V4	1	1	1	1	0	1	0	1	1	0

$$\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}_3 = \frac{1}{2}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}.$$

If we calculate $\mathbf{W}_k$, we find

$$\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W}_3 = \begin{pmatrix} 0.25 & -0.0833 & -0.0833 & -0.0833 \\ -0.0833 & 0.25 & -0.0833 & -0.0833 \\ -0.0833 & -0.0833 & 0.25 & -0.0833 \\ -0.0833 & -0.0833 & -0.0833 & 0.25 \end{pmatrix} = \frac{1}{3}\left(\mathbf{I}_4 - \frac{1}{4}\mathbf{1}_4\mathbf{1'}_4\right).$$

With $\mathbf{p'} = (1 \quad -1 \quad 0 \quad 0)$,

$$\mathbf{p'W}_1 = (0.3333 \quad -0.3333 \quad 0.0000 \quad 0.0000) = \frac{1}{3}\mathbf{p'},$$

$$\mathbf{p'W}_2 = (0.3333 \quad -0.3333 \quad 0.0000 \quad 0.0000) = \frac{1}{3}\mathbf{p'},$$

$$\mathbf{p'W}_3 = (0.3333 \quad -0.3333 \quad 0.0000 \quad 0.0000) = \frac{1}{3}\mathbf{p'}.$$

Notice that for this design, where all the **C** matrices are identical, univariate weights can be used, with equal weight $1/m$ for every site.

Example 3.

	S1		S2			S3		
	B1	B2	B3	B4	B5	B6	B7	B8
V1	1	1	1	1	1	1	1	0
V2	1	1	1	1	1	1	0	1
V3	1	1	1	1	1	0	1	1

$$\mathbf{C}_1 = \frac{2}{3}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad \mathbf{W}_1 = \begin{pmatrix} 0.2051 & -0.1026 & -0.1026 \\ -0.1026 & 0.2051 & -0.1026 \\ -0.1026 & -0.1026 & 0.2051 \end{pmatrix}.$$

$$\mathbf{C}_2 = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad \mathbf{W}_2 = \begin{pmatrix} 0.3077 & -0.1538 & -0.1538 \\ -0.1538 & 0.3077 & -0.1538 \\ -0.1538 & -0.1538 & 0.3077 \end{pmatrix}.$$

$$\mathbf{C}_3 = \frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \quad \mathbf{W}_3 = \begin{pmatrix} 0.1538 & -0.0769 & -0.0769 \\ -0.0769 & 0.1538 & -0.0769 \\ -0.0769 & -0.0769 & 0.1538 \end{pmatrix}.$$

Let $\mathbf{p}' = (1 \quad -1 \quad 0)$. Then

$$\mathbf{p}'\mathbf{W}_1 = (0.3077 \quad -0.3077 \quad 0.0000) = \frac{4}{13}\mathbf{p}',$$

$$\mathbf{p}'\mathbf{W}_2 = (0.4615 \quad -0.4615 \quad 0.0000) = \frac{6}{13}\mathbf{p}',$$

$$\mathbf{p}'\mathbf{W}_3 = (0.2308 \quad -0.2308 \quad 0.0000) = \frac{3}{13}\mathbf{p}'.$$

Notice that for this example $\mathbf{p}'\mathbf{W}_k\tau$ represents a contrast of the same form as $\mathbf{p}'\tau$. The weights are here univariate. We could have known this beforehand because the **C** matrices are a multiple of each other, hence a multiple of $\mathbf{C}_.$. The univariate weights can directly be calculated as (4/6)/(13/6)=4/13, (6/6)/(13/6)=6/13 and (3/6)/(13/6)=3/13.

3.2.3 The fixed interaction model

First assume that all t varieties are present at the m sites. Suppose we are of the opinion that variety × site interaction parameters should be included in the fixed model. Then the fixed additive model (3.8) can be extended to a fixed interaction model as

$$Y_{ijkl} = \mu + \tau_i + \beta_{j(k)} + \lambda_k + (\tau\lambda)_{ik} + E_{ijkl}, \tag{3.11}$$

with $(\tau\lambda)_{ik}$ the variety × site interaction parameters. With this model, the value of variety i is defined in terms of the parameters as

$$\mu + \tau_i + \sum_{k=1}^{m} w_k(\tau\lambda)_{ik} + \sum_{k=1}^{m} w_k(\overline{\beta}_{\cdot(k)} + \lambda_k),$$

with $\Sigma\, w_k = 1$. Now a contrast between variety values is not equal to the same contrast between variety parameters, but to this contrast between $\tau_i + \Sigma\, w_k(\tau\lambda)_{ik}$ $= \Sigma\, w_k(\tau_i + (\tau\lambda)_{ik})$. With the fixed interaction model this contrast depends on the given weights w_k (read : the definition of a variety value). A contrast between variety parameters τ_i is not estimable with this model because of the overparameterisation, but a contrast between variety values, as defined above, is. A variety value is a linear combination of $E(Y_{ijkl})$ and therefore estimable. Hence a contrast between variety values is also estimable. Note that a contrast between variety values in a fixed additive model is defined differently from the same contrast between variety values in a fixed interaction model.

In the analysis of the individual trial at site k, with model (3.1), the variety parameter corresponding to variety i is equivalent to $\tau_i + (\tau\lambda)_{ik}$ in model (3.11). So, the contrast estimates from the separate trials have to be averaged with weights given in the definition of a variety value in presence of interaction, in order to estimate contrasts between variety values. If the sites are of the same importance, equal weights $1/m$ are used. Sometimes it will be reasonable to give unequal weights to the various sites. If a site represents a part of the region which is very important, e.g. because of economical reasons, this site may be given a larger weight than the other sites. However, the choice of the weights should not be based on the results of the experiments.

Let us now return to matrix notation. Let the vector of variety values in presence of interaction be denoted by ξ. Then the least squares estimator of $\mathbf{p}'\xi$ can be written as $\mathbf{p}'\hat{\xi} = \Sigma\, w_k \mathbf{p}'\tilde{\tau}_k$, with $\mathbf{p}'\tilde{\tau}_k$ the estimator of a contrast between variety parameters at site k ($k = 1, \ldots, m$) and the w_k (with $\Sigma\, w_k = 1$) following from the definition of a variety value in the interaction model. The pseudo-variance/covariance matrix of $\hat{\xi}$ can be calculated as $\dot{\mathbf{D}}[\hat{\xi}] = (\Sigma\, w_k^2 \mathbf{C}_k^-)\sigma^2$.

The vectors with the parameters corresponding to site k ($k = 1, \ldots, m$) can be denoted by $\mathbf{1}\mu$, τ, $(\tau\lambda)_k$, β_k and $\mathbf{1}\lambda_k$. In the separate analysis of the trial at site k with the fixed model (3.1), the vector with variety parameters is equivalent with $\tau + (\tau\lambda)_k$ and the vector $\mathbf{1}\lambda + \beta$ is equivalent with $\mathbf{1}\mu + \beta_k + \mathbf{1}\lambda_k$ in the current model. Therefore the part of the vector of residuals $\hat{\mathbf{E}}$ corresponding to site k is equal to the vector of residuals from the separate analysis of that site. Hence we can calculate the estimate of the error variance by summation of the error sums of squares of the m sites and division by the degrees of freedom for error. With the fixed interaction model the degrees of freedom for error are equal to $df_e = n - \Sigma\, b_k - m(t-1)$, this is the sum of the df_e from the separate analyses of the m trials.

When will the estimate of a contrast between variety values, using a fixed additive model, be equal to the estimate of the same contrast between variety values using a fixed interaction model ? This will only be the case if the weights used with the additive model are univariate and equal to the weights following from the definition of a variety value in presence of interaction. If the latter weights are chosen to be $1/m$, the estimates from the fixed additive model are equal to the estimates from the fixed interaction model if and only if the **C** matrices at the various sites are equal to each other. This is e.g. the case if all the experiments at the different sites have the same design.

Now suppose that not all varieties are grown at all sites. So the variety × site table is incomplete. Then the estimate of a contrast between two variety values can only be calculated if the two varieties are present at all sites. Consequently, for incomplete variety × site tables some variety value contrasts cannot be estimated. Therefore, if we want to use a fixed interaction model, we must aim at a complete variety × site table. Since the variety × site tables in the plant breeding practice are mostly incomplete, the fixed interaction model appears to be not very useful in that field.

3.2.4 The mixed additive model

In the mixed additive model the terms for blocks within sites, or the site terms, or both are assumed to be random variables. If the terms for blocks within sites are assumed to be random variables, it is sometimes reasonable to extend the model by including fixed terms for groups of blocks. For instance, if resolvable designs are used at the various sites, fixed replication terms can be included in the model. Similarly, if the site terms are considered random but can be classified into a number of distinct groups (e.g. soil types), then the model can be extended by including fixed terms for groups of sites. We will elaborate estimation using the basic models and additionally mention the changes to be made when an extended model is used. Analogous to section *3.1.2*, we assume that variance ratios are known. For all models the variance/covariance matrix of the observations can be written as

$$\mathbf{D}\left[\begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\\ \vdots\\ \mathbf{Y}_m\end{pmatrix}\right]=\begin{pmatrix}\mathbf{V}_1 & & & \textbf{-0-}\\ & \mathbf{V}_2 & & \\ & & \ddots & \\ \textbf{-0-} & & & \mathbf{V}_m\end{pmatrix}\sigma^2,$$

so observations from different sites are uncorrelated.

First consider the situation where the terms for blocks within sites are random, and the site terms fixed. The model for the observations can then be written as

$$Y_{ijkl} = \mu + \tau_i + B_{j(k)} + \lambda_k + E_{ijkl} . \tag{3.12}$$

The $B_{j(k)}$ are uncorrelated random variables with expectation zero and those corresponding to site k $(k = 1,..,m)$ have a common variance $\sigma^2_{B(k)}$. Furthermore we assume that $\text{cov}(B_{j(k)}, E_{ijkl}) = 0$ for all i,j,k,l. The expectation of Y_{ijkl} is equal to $E[Y_{ijkl}] = \mu + \tau_i + \lambda_k$. The value of variety i is now defined as

$$\mu + \tau_i + \sum_{k=1}^{m} w_k\lambda_k , \quad \text{with} \sum_{k=1}^{m} w_k = 1 ,$$

so a contrast between variety values is equal to that contrast between variety parameters. If we change over to matrix notation, we can write the expectation of $\mathbf{Y}$ as

$$E\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} \\ \vdots \\ \mathbf{1}_{n_m} \end{pmatrix}\mu + \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}\tau + \begin{pmatrix} \mathbf{1}_{n_1} & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{1}_{n_m} \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix},$$

where n_k denotes the number of observations at site k $(k = 1, \ldots, m)$. The variance/covariance submatrix (divided by σ^2) of the observations of site k $(k = 1, \ldots, m)$ is equal to

$$\mathbf{V}_k = \mathbf{Z}_k\mathbf{Z'}_k \frac{\sigma^2_{B(k)}}{\sigma^2} + \mathbf{I}_{n_k} .$$

We can rewrite the expectation equations corresponding to site k as

$$\begin{aligned} E[\mathbf{Y}_k] &= \mathbf{1}_{n_k}\mu + \mathbf{X}_k\tau + \mathbf{1}_{n_k}\lambda_k \\ &= \boldsymbol{R}_{\mathbf{1}_k}\mathbf{X}_k\tau + \boldsymbol{P}_{\mathbf{1}_k}\mathbf{X}_k\tau + \mathbf{1}_{n_k}(\lambda_k + \mu) , \\ &\text{with } \boldsymbol{P}_{\mathbf{1}_k} = \mathbf{1}_{n_k}\left(\mathbf{1'}_{n_k}\mathbf{V}_k^{-1}\mathbf{1}_{n_k}\right)^{-1}\mathbf{1'}_{n_k}\mathbf{V}_k^{-1} \text{ and } \boldsymbol{R}_{\mathbf{1}_k} = \mathbf{I}_{n_k} - \boldsymbol{P}_{\mathbf{1}_k} , \\ &= \boldsymbol{R}_{\mathbf{1}_k}\mathbf{X}_k\tau + \mathbf{1}_{n_k}\lambda_k^* , \quad \text{with } \lambda_k^* = \left(\mathbf{1'}_{n_k}\mathbf{V}_k^{-1}\mathbf{1}_{n_k}\right)^{-1}\mathbf{1'}_{n_k}\mathbf{V}_k^{-1}\mathbf{X}_k\tau + \lambda_k + \mu . \end{aligned}$$

$\boldsymbol{R}_{1k}$ denotes a projection on the subspace spanned by the unit vector. The normal equations can now be written as

$$\begin{pmatrix} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{1}'_{n_1} \mathbf{V}_1^{-1} \mathbf{1}_{n_1} & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{1}'_{n_m} \mathbf{V}_m^{-1} \mathbf{1}_{n_m} \end{pmatrix} \begin{pmatrix} \hat{\tau} \\ \hat{\lambda}_1^* \\ \vdots \\ \hat{\lambda}_m^* \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{Y}_k \\ \mathbf{1}'_{n_1} \mathbf{V}_1^{-1} \mathbf{Y}_1 \\ \vdots \\ \mathbf{1}'_{n_m} \mathbf{V}_m^{-1} \mathbf{Y}_m \end{pmatrix} .$$

A (generalised least squares) solution for $\hat{\tau}$ is :

$$\hat{\tau} = \left(\sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k \right)^{-} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{Y}_k$$

$$= \sum_{k=1}^{m} \mathbf{W}_k \tilde{\tau}_k , \quad \text{with } \mathbf{W}_k = \left(\sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k \right)^{-} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k . \tag{3.13}$$

In (3.13) $\tilde{\tau}_k$ is the vector of least squares estimators of variety terms calculated at site k ($k = 1, \ldots, m$), when the mixed model (3.4) is used at this site. Analogous to section *3.2.2* we notice that the BLUE of $\mathbf{p}'\tau$ is in general not a univariately weighted average of the $\mathbf{p}'\tilde{\tau}_k$, but a multivariately weighted average. The pseudo-variance/covariance matrix of $\hat{\tau}$ is equal to

$$\dot{\mathbf{D}}[\hat{\tau}] = \left(\sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k \right)^{-} \sigma^2 .$$

To estimate the error variance, we first calculate the vector with residuals, denoted by $\hat{\mathbf{F}}$. For observations corresponding to site k ($k = 1, \ldots, m$) the vector of residuals is equal to

$$\hat{\mathbf{F}}_k = \mathbf{Y}_k - \boldsymbol{R}_{1_k} \mathbf{X}_k \hat{\tau} - \mathbf{1}_{n_k} \hat{\lambda}_k^*$$

$$= \boldsymbol{R}_{1_k} (\mathbf{Y}_k - \mathbf{X}_k \hat{\tau}) .$$

The degrees of freedom for error are equal to $df_f = n - t - m + 1$ and the estimate of σ^2 can be calculated as $s^2 = (\Sigma\, \hat{\mathbf{F}}'_k \hat{\mathbf{F}}_k)/df_f$.

The model can be extended by including fixed replication terms within sites. In that case the same formulae for $\hat{\tau}$, $\dot{\mathbf{D}}[\hat{\tau}]$ and $\hat{\mathbf{F}}_k$ as given above can be used, only $\mathbf{1}_k$ has to be replaced by $\mathbf{M}_k$ ($k = 1, \ldots, m$), with $\mathbf{M}_k$ the design matrix corresponding with the replications at site k.

Now assume that the block terms are fixed. We then use the following basic model for the observations :

$$Y_{ijkl} = \mu + \tau_i + \beta_{j(k)} + L_k + E_{ijkl} \,. \tag{3.14}$$

We assume that the L_k are uncorrelated random variables with expectation zero and common variance σ_L^2. Furthermore we assume that $\mathrm{cov}(L_k, E_{ijkl}) = 0$ for all i, j, k, l. The expectation of Y_{ijkl} is now equal to $E[Y_{ijkl}] = \mu + \tau_i + \beta_{j(k)}$, and the value of variety i is defined as

$$\mu + \tau_i + \sum_{k=1}^{m} w_k \overline{\beta}_{\cdot(k)}\,, \quad \text{with } \sum_{k=1}^{m} w_k = 1 \,.$$

Now a variety value has to be interpreted as an average over the population of sites. A contrast between variety values is equal to that contrast between variety parameters, and does not depend on the weights chosen.

First suppose the same varieties are present at every site. In matrix notation, we can write the expectation of the observations as

$$E\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} \\ \mathbf{1}_{n_2} \\ \vdots \\ \mathbf{1}_{n_m} \end{pmatrix}\mu + \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{pmatrix}\tau + \begin{pmatrix} \mathbf{Z}_1 & & & \text{-0-} \\ & \mathbf{Z}_2 & & \\ & & \ddots & \\ \text{-0-} & & & \mathbf{Z}_m \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix},$$

with n_k the number of observations at site k ($k = 1, \ldots, m$). Let the total number of observations be denoted by n. The variance/covariance submatrix (divided by σ^2) of the observations at site k $(k = 1, \ldots, m)$ is equal to

$$\mathbf{V}_k = \mathbf{1}_{n_k}\mathbf{1}'_{n_k}(\sigma_L^2/\sigma^2) + \mathbf{I}_{n_k} \,.$$

We can rewrite the expectation equations corresponding to site k as

$$E[\mathbf{Y}_k] = \mathbf{1}_{n_k}\mu + \mathbf{X}_k\tau + \mathbf{Z}_k\beta_k$$

$$= \boldsymbol{R}_{\mathbf{Z}_k}\mathbf{X}_k\tau + \boldsymbol{P}_{\mathbf{Z}_k}\mathbf{X}_k\tau + \mathbf{Z}_k\left(\beta_k + \mathbf{1}_{b_k}\mu\right),$$

$$\text{with } \boldsymbol{P}_{\mathbf{Z}_k} = \mathbf{Z}_k(\mathbf{Z}'_k\mathbf{V}_k^{-1}\mathbf{Z}_k)^{-1}\mathbf{Z}'_k\mathbf{V}_k^{-1} \text{ and } \boldsymbol{R}_{\mathbf{Z}_k} = \mathbf{I}_{n_k} - \boldsymbol{P}_{\mathbf{Z}_k}\,,$$

$$= \boldsymbol{R}_{\mathbf{Z}_k}\mathbf{X}_k\tau + \mathbf{Z}_k\beta_k^*\,, \quad \text{with } \beta_k^* = (\mathbf{Z}'_k\mathbf{V}_k^{-1}\mathbf{Z}_k)^{-1}\mathbf{Z}'_k\mathbf{V}_k^{-1}\mathbf{X}_k\tau + \beta_k + \mathbf{1}_{b_k}\mu \,.$$

$\boldsymbol{R}_{Zk}$ denotes a projection on the subspace spanned by the columns of $\mathbf{Z}_k$. Notice that $\boldsymbol{R}_{Zk}$ is different from $\mathbf{R}_{Zk}$, used in *3.2.2*. Now the normal equations have a convenient form :

$$\begin{pmatrix} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{\mathbf{Z}_k} \mathbf{X}_k & & & \text{-0-} \\ & \mathbf{Z}'_1 \mathbf{V}_1^{-1} \mathbf{Z}_1 & & \\ & & \ddots & \\ \text{-0-} & & & \mathbf{Z}'_m \mathbf{V}_m^{-1} \mathbf{Z}_m \end{pmatrix} \begin{pmatrix} \hat{\tau} \\ \beta_1^* \\ \vdots \\ \beta_m^* \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{\mathbf{Z}_k} \mathbf{Y}_k \\ \mathbf{Z}'_1 \mathbf{V}_1^{-1} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Z}'_m \mathbf{V}_m^{-1} \mathbf{Y}_m \end{pmatrix}.$$

A (generalised least squares) solution of these equations for $\hat{\tau}$ can be calculated as

$$\hat{\tau} = \left(\sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{\mathbf{Z}_k} \mathbf{X}_k \right)^{-} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{\mathbf{Z}_k} \mathbf{Y}_k \,. \tag{3.15}$$

We will now show that the solution given above is identical to the solution presented with the fixed additive model. The inverse of $\mathbf{V}_k$ can be calculated as

$$\mathbf{V}_k^{-1} = \mathbf{I}_{n_k} - \gamma_k \mathbf{1}_{n_k} \mathbf{1}'_{n_k}\,, \quad \text{with } \gamma_k = \frac{1}{n_k + (\sigma^2/\sigma_L^2)}\,.$$

This expression can be rewritten as

$$\begin{aligned} \mathbf{V}_k^{-1} &= \mathbf{P}_{\mathbf{Z}_k} + \mathbf{R}_{\mathbf{Z}_k} - \gamma_k \mathbf{P}_{\mathbf{Z}_k} \mathbf{1}_{n_k} \mathbf{1}'_{n_k} \\ &= \mathbf{P}_{\mathbf{Z}_k} \mathbf{V}_k^{-1} + \mathbf{R}_{\mathbf{Z}_k}\,, \quad \text{with } \mathbf{P}_{\mathbf{Z}_k} = \mathbf{Z}_k (\mathbf{Z}'_k \mathbf{Z}_k)^{-1} \mathbf{Z}'_k \text{ and } \mathbf{R}_{\mathbf{Z}_k} = \mathbf{I}_{n_k} - \mathbf{P}_{\mathbf{Z}_k}\,. \end{aligned}$$

So,

$$\begin{aligned} \mathbf{V}_k^{-1} \boldsymbol{R}_{\mathbf{Z}_k} &= \mathbf{P}_{\mathbf{Z}_k} \mathbf{V}_k^{-1} \boldsymbol{R}_{\mathbf{Z}_k} + \mathbf{R}_{\mathbf{Z}_k} \boldsymbol{R}_{\mathbf{Z}_k} \\ &= \mathbf{P}_{\mathbf{Z}_k} \boldsymbol{R}'_{\mathbf{Z}_k} \mathbf{V}_k^{-1} + \mathbf{R}_{\mathbf{Z}_k} \\ &= \mathbf{R}_{\mathbf{Z}_k}\,. \end{aligned}$$

Hence (3.15) reduces to (3.10). Consequently, if a mixed additive model is used with fixed block terms, the BLUEs of estimable functions of the variety parameters are equal to the corresponding estimators in the fixed additive model. Also the pseudo-variance/covariance matrix of the estimators of the variety parameters is equal in both models.

In order to estimate σ^2 we can calculate the residuals. The residuals corresponding to the observations at site k can be calculated as

$$\begin{aligned}\hat{\mathbf{F}}_k &= \mathbf{Y}_k - \boldsymbol{R}_{\mathbf{Z}_k}\mathbf{X}_k\hat{\boldsymbol{\tau}} - \mathbf{Z}_k\boldsymbol{\beta}^*_k \\ &= \mathbf{Y}_k - \boldsymbol{R}_{\mathbf{Z}_k}\mathbf{X}_k\hat{\boldsymbol{\tau}} - \mathbf{Z}_k(\mathbf{Z}'_k\mathbf{V}_k^{-1}\mathbf{Z}_k)^{-1}\mathbf{Z}'_k\mathbf{V}_k^{-1}\mathbf{Y}_k \\ &= \boldsymbol{R}_{\mathbf{Z}_k}(\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\tau}})\ .\end{aligned}$$

But if $\mathbf{V}_k^{-1}\boldsymbol{R}_{\mathbf{Z}_k} = \mathbf{R}_{\mathbf{Z}_k}$ then :

$$\begin{aligned}\boldsymbol{R}_{\mathbf{Z}_k} &= \mathbf{V}_k\mathbf{R}_{\mathbf{Z}_k} \\ &= \left(\mathbf{I}_{n_k} + (\sigma_L^2/\sigma^2)\mathbf{1}_{n_k}\mathbf{1}'_{n_k}\right)\mathbf{R}_{\mathbf{Z}_k} \\ &= \mathbf{R}_{\mathbf{Z}_k}\ ,\end{aligned}$$

and also $\boldsymbol{P}_{\mathbf{Z}k} = \mathbf{P}_{\mathbf{Z}k}$. Hence we can use the same sum of squares for error as calculated for a fixed additive model, namely SS_E. The degrees of freedom for error are equal to $df_f = n - t - \Sigma b_k + 1$, and finally σ^2 can be estimated as $s^2 = SS_E/df_f$.

Analogous to section *3.2.2*, the theory also applies if the variety × site table is incomplete. In that case columns with zeros enter the $\mathbf{X}_k$ matrices, but the $\mathbf{Z}_k$ matrices remain the same. In the proof that the estimators of variety parameters in the mixed additive model with fixed block terms are identical to the estimators of variety parameters in the fixed additive model $\mathbf{X}_k$ is not used.

Now assume that the model is extended by introducing fixed terms for groups of sites, with the sites within these groups considered a random sample. The variance/covariance matrix of the observations does not change because of this extension. Furthermore, the introduction of extra fixed group terms is merely an overparameterisation of the model. Therefore, they just as well can be removed from the model and the estimates obtained are identical to the ones described above.

Finally consider the situation where both block terms and site terms are assumed to be random variables. Then the model for the observations reads :

$$Y_{ijkl} = \mu + \tau_i + B_{j(k)} + L_k + E_{ijkl}\ . \qquad (3.16)$$

$B_{j(k)}$, L_k and E_{ijkl} are uncorrelated random variables with zero expectation. For a certain site k ($k = 1, \ldots, m$) all $B_{j(k)}$ have a common variance $\sigma^2_{B(k)}$. The L_k and the E_{ijkl} have common variance σ_L^2 and σ^2, respectively. Furthermore it is assumed that

$\mathrm{cov}(B_{j(k)}, L_k) = 0$, $\mathrm{cov}(B_{j(k)}, E_{ijkl}) = 0$ and $\mathrm{cov}(L_k, E_{ijkl}) = 0$ for all i, j, k, l. A variety value is defined as $\mu + \tau_i$, because $E[Y_{ijkl}] = \mu + \tau_i$. Hence a contrast between variety values is equal to that contrast between variety parameters. In matrix notation, the variance/covariance submatrix (divided by σ^2) of the observations at site k ($k = 1, \ldots, m$) can be written as

$$\mathbf{V}_k = \mathbf{Z}_k \mathbf{Z}'_k \frac{\sigma^2_{B(k)}}{\sigma^2} + \mathbf{1}_{n_k} \mathbf{1}'_{n_k} \frac{\sigma^2_L}{\sigma^2} + \mathbf{I}_{n_k},$$

$$\text{hence } \mathbf{V}_k^{-1} = \mathbf{I}_{n_k} - \mathbf{Z}_k \left(\mathbf{Z}'_k \mathbf{Z}_k + \mathbf{I}_{b_k} \frac{\sigma^2}{\sigma^2_{B(k)}} - \frac{\sigma^2}{\sigma^2_{B(k)}} \frac{\sigma^2_L}{\sigma^2_L b_k + \sigma^2_{B(k)}} \mathbf{1}_{b_k} \mathbf{1}'_{b_k} \right)^{-1} \mathbf{Z}'_k .$$

Analogous to the situation of model (3.12) we find as (generalised least squares) solution for $\hat{\tau}$:

$$\hat{\tau} = \left(\sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k \right)^{-} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{Y}_k .$$

It is known (Rao, 1973) that in the expression for the $\tilde{\tau}_k$ at site k with a mixed model with random block terms the variance/covariance matrix $\left(\mathbf{Z}_k \mathbf{Z}'_k \sigma^2_{B(k)}/\sigma^2 + \mathbf{I}_{n_k}\right)\sigma^2$ can be replaced by $\left(\mathbf{Z}_k \mathbf{Z}'_k \sigma^2_{B(k)}/\sigma^2 + \mathbf{I}_{n_k}\right)\sigma^2 + \mathbf{X}_k \mathbf{A}_k \mathbf{X}'_k \sigma^2$, with $\mathbf{A}_k$ any matrix. If $\mathbf{A}_k$ is chosen to be $\mathbf{1}_t \mathbf{1}'_t \sigma^2_L/\sigma^2$, then we obtain the $\mathbf{V}_k$ of the current model. Hence with model (3.16) the estimator of τ can be calculated as a multivariately weighted average of the $\tilde{\tau}_k$ from the separate sites, the latter calculated using a mixed model with random block terms. So,

$$\hat{\tau} = \sum_{k=1}^{m} \mathbf{W}_k \tilde{\tau}_k , \text{ with } \mathbf{W}_k = \left(\sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{X}_k \right)^{-} \sum_{k=1}^{m} \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1_k} \mathbf{Y}_k . \qquad (3.17)$$

For the weight matrices it is important to use the correct $\mathbf{V}_k$. The pseudo-variance/covariance matrix of $\hat{\tau}$ is equal to $\dot{\mathbf{D}}[\hat{\tau}] = (\Sigma \mathbf{X}'_k \mathbf{V}_k^{-1} \boldsymbol{R}_{1k} \mathbf{X}_k)^{-} \sigma^2$. If the variance ratios $\sigma^2_{B(k)}/\sigma^2$ approach infinity, the inverse of $\mathbf{V}_k$ becomes equal to $\mathbf{R}_{Zk}$. Hence in that case the same results are obtained as if we would have used the fixed additive model.

Remark

With this model it might be easier to calculate the estimates of the variety values instead of the variety parameters. The vector of variety values can be calculated as

$$\hat{\tau}+\mathbf{1}_t\hat{\mu} = \sum_{k=1}^{m} \mathbf{W}_k(\tilde{\tau}_k+\mathbf{1}_t\tilde{\lambda}_k)\,,\ \text{with}\ \mathbf{W}_k = \left(\sum_{k=1}^{m} \mathbf{X}'_k\mathbf{V}_k^{-1}\mathbf{X}_k\right)^{-1}\mathbf{X}'_k\mathbf{V}_k^{-1}\mathbf{X}_k\,.$$

The vector of residuals corresponding to the observations at site k $(k = 1,\ldots,m)$ is calculated as $\hat{\mathbf{F}}_k = \boldsymbol{R}_{1k}(\mathbf{Y}_k - \mathbf{X}_k\hat{\tau})$. The degrees of freedom for error are equal to $df_f = n - t$. Then the estimate of σ^2 is equal to $s^2 = (\Sigma\,\hat{\mathbf{F}}'_k\hat{\mathbf{F}}_k)/df_f$.

Sometimes it is reasonable to extend the model with fixed replication terms within sites. The model can be further extended by including fixed terms for groups of sites, but this is merely an overparameterisation. In this situation the formulae are equal to those for the situation of fixed site terms, fixed replication terms within sites and random block terms within each replication, only the variance/covariance matrix of the observations has to be equal to the one described above.

3.2.5 The mixed interaction model

With the mixed interaction model we can distinguish three basic situations : 1) the block terms are random and the site terms and the variety × site interaction terms are fixed, 2) the block terms are fixed and the site and interaction terms are random, 3) all terms except variety terms and general level are random. Extended models for trials with resolvable designs at the various sites assume fixed replication terms with random terms for blocks within these replications. If site terms and variety × site interaction terms are considered random, extended models can include fixed terms for groups of sites and fixed variety × group interaction terms. Within these groups the sites are considered random. We will study the three basic situations and make some comments about the extended model after each case.

First consider the situation where the terms for blocks within sites are assumed to be random variables, and site terms are assumed to be fixed. The corresponding model for the observations can be written as

$$Y_{ijkl} = \mu + \tau_i + B_{j(k)} + \lambda_k + (\tau\lambda)_{ik} + E_{ijkl}\,. \tag{3.18}$$

$B_{j(k)}$ and E_{ijkl} are uncorrelated random variables with expectation zero and variances $\sigma^2_{B(k)}$ and σ^2, respectively. Furthermore it is assumed that $\text{cov}(B_{j(k)}, E_{ijkl}) = 0$ for all

i,j,k,l. The expectation of Y_{ijkl} is equal to $\mu+\tau_i+\lambda_k+(\tau\lambda)_{ik}$, and therefore the variety value of variety i is defined as

$$\mu+\tau_i+\sum_{k=1}^{m} w_k(\tau\lambda)_{ik}+\sum_{k=1}^{m} w_k\lambda_k \ ,$$

with $\Sigma\, w_k = 1$. A contrast between variety values is equal to that contrast between $\tau_i+\Sigma\, w_k(\tau\lambda)_{ik}$ and not between τ_i. As explained in section *3.2.3*, we have to average the estimators from the separate trials in order to obtain the BLUEs of contrasts between variety values in presence of interaction : $\mathbf{p}'\hat{\xi}=\Sigma w_k\mathbf{p}'\tilde{\tau}_k$. Here the estimators from the separate trials correspond to the mixed additive model (3.4). Then $\dot{\mathbf{D}}[\hat{\xi}]=\Sigma w_k^2\dot{\mathbf{D}}[\tilde{\tau}_k]=\Sigma w_k^2(\mathbf{X}'_k\mathbf{V}_k^{-1}\boldsymbol{R}_{1k}\mathbf{X}_k)^-\sigma^2$. The SS_F can be calculated as the sum of the squared residuals from the separate trials. The degrees of freedom for error are equal to $df_f = n-tm$, this is the sum of the df_f from the separate analyses of the m trials.

The model can be extended by introducing fixed replication terms within sites. In that case the BLUEs are still equal to average estimators from the separate trials, but now the latter estimators correspond to a model with fixed replication terms and random block terms within the replications.

If the block terms are considered fixed, the model for the observations can be written as

$$Y_{ijkl}=\mu+\tau_i+\beta_{j(k)}+L_k+(TL)_{ik}+E_{ijkl}\ . \tag{3.19}$$

L_k, $(TL)_{ik}$ and E_{ijkl} are assumed to be uncorrelated random variables with expectation zero. The variables L_k have a common variance σ_L^2, the variables $(TL)_{ik}$ have a common variance σ_{TL}^2 and the E_{ijkl} have common variance σ^2. Furthermore we assume that $\mathrm{cov}(L_k,(TL)_{ik})=0$, $\mathrm{cov}(L_k,E_{ijkl})=0$ and $\mathrm{cov}((TL)_{ik},E_{ijkl})=0$ for all i,j,k,l. The expectation of Y_{ijkl} is equal to $\mu+\tau_i+\beta_{j(k)}$, and a variety value is defined as

$$\mu+\tau_i+\sum_{k=1}^{m} w_k\overline{\beta}_{\cdot(k)}\ ,$$

with $\Sigma\, w_k = 1$. A contrast between variety values is equal to that contrast between variety parameters. Written in matrix notation, the variance/covariance submatrix (divided by σ^2) of the observations at site k $(k=1,\ldots,m)$ is equal to

$$\mathbf{V}_k = \mathbf{X}_k\mathbf{X}'_k\frac{\sigma^2_{TL}}{\sigma^2} + \mathbf{1}_{n_k}\mathbf{1}'_{n_k}\frac{\sigma^2_L}{\sigma^2} + \mathbf{I}_{n_k}\,,$$

$$\text{hence } \mathbf{V}_k^{-1} = \mathbf{I}_{n_k} - \mathbf{X}_k\left(\mathbf{X}'_k\mathbf{X}_k + \mathbf{I}_t\frac{\sigma^2}{\sigma^2_{TL}} - \frac{\sigma^2}{\sigma^2_{TL}}\frac{\sigma^2_L}{t\sigma^2_L + \sigma^2_{TL}}\mathbf{1}_t\mathbf{1}'_t\right)^{-1}\mathbf{X}'_k\,.$$

In this case we can use solution (3.15), only the $\mathbf{V}_k$ matrices are different. $\mathbf{V}_k$ can be written as $\mathbf{I}_{n(k)} + \mathbf{X}_k\mathbf{A}_k\mathbf{X}'_k$, with $\mathbf{A}_k = \mathbf{I}_t\sigma^2_{TL}/\sigma^2 + \mathbf{1}_t\mathbf{1}'_t\sigma^2_L/\sigma^2$. At site k the solution $(\mathbf{X}'_k\mathbf{V}_k^{-1}\boldsymbol{R}_{\mathbf{Z}k}\mathbf{X}_k)^{-}\mathbf{X}'_k\mathbf{V}_k^{-1}\boldsymbol{R}_{\mathbf{Z}k}\mathbf{Y}_k$ is then equal to $(\mathbf{X}'_k\mathbf{R}_{\mathbf{Z}k}\mathbf{X}_k)^{-}\mathbf{X}'_k\mathbf{R}_{\mathbf{Z}k}\mathbf{Y}_k$, which is $\tilde{\tau}_k$ from the fixed additive model at site k. Hence,

$$\hat{\tau} = \sum_{k=1}^{m}\mathbf{W}_k\tilde{\tau}_k\,,\ \text{with } \mathbf{W}_k = \left(\sum_{k=1}^{m}\mathbf{X}'_k\mathbf{V}_k^{-1}\boldsymbol{R}_{\mathbf{Z}_k}\mathbf{X}_k\right)^{-}\sum_{k=1}^{m}\mathbf{X}'_k\mathbf{V}_k^{-1}\boldsymbol{R}_{\mathbf{Z}_k}\mathbf{Y}_k\,. \tag{3.20}$$

Here, $\tilde{\tau}_k$ is the estimator of τ at site k, using a fixed additive model. Furthermore, $\dot{\mathbf{D}}[\hat{\tau}] = (\Sigma\mathbf{X}'_k\mathbf{V}_k^{-1}\boldsymbol{R}_{\mathbf{Z}k}\mathbf{X}_k)^{-}\sigma^2$. If the variance ratio σ^2_{TL}/σ^2 approaches infinity, the inverse of $\mathbf{V}_k$ becomes equal to $\mathbf{R}_{\mathbf{X}k} = \mathbf{I}_{n(k)} - \mathbf{X}_k(\mathbf{X}'_k\mathbf{X}_k)^{-1}\mathbf{X}'_k$. But then $\mathbf{X}'_k\mathbf{V}_k^{-1} = \mathbf{X}'_k\mathbf{R}_{\mathbf{X}k} = 0$. This is to be expected, because for σ^2_{TL}/σ^2 approaching infinity, model (3.19) becomes equal to a model with fixed interaction terms, and in that situation $\mathbf{p}'\tau$ is not estimable.

Remark

In the Scheffé type of mixed models the random variables $(TL)_{ik}$ are not assumed uncorrelated. They are defined such that $\Sigma_{i=1}^{t}\,(TL)_{ik} = 0$ for all k and equal $\mathrm{cov}((TL)_{ik}, (TL)_{i'k})$ for all k and $i \neq i'$. In this case the variance/covariance matrix of these variables is not a diagonal matrix. Assume that this variance/covariance matrix is equal to $\mathbf{H}_k\sigma^2_{TL}$. Then $\mathbf{V}_k$ is equal to

$$\mathbf{V}_k = \mathbf{X}_k\mathbf{H}_k\mathbf{X}'_k\frac{\sigma^2_{TL}}{\sigma^2} + \mathbf{1}_{n_k}\mathbf{1}'_{n_k}\frac{\sigma^2_L}{\sigma^2} + \mathbf{I}_{n_k} = \mathbf{I}_{n_k} + \mathbf{X}_k\left(\mathbf{H}_k\frac{\sigma^2_{TL}}{\sigma^2} + \mathbf{1}_t\mathbf{1}'_t\right)\mathbf{X}'_k\,.$$

Suppose that model (3.19) is extended by including fixed terms for groups of sites and fixed variety × group interaction terms. The groups can for instance represent different soil types. Let $(\tau\gamma)_{ig}$ denote the fixed interaction contribution of variety i and group g. As with the fixed interaction model in *3.2.3*, the value of variety i $(i = 1, \ldots, t)$ must now be defined as $\mu + \tau_i + \Sigma\,\tilde{w}_g(\tau\gamma)_{ig} + \Sigma\,w_k\bar{\beta}_{\cdot(k)}$, with $\tilde{w}_g$ the weight of group g. This weight is equal to the sum of the weights of the sites within group g. In the separate analyses of the trials at the individual sites within a single group g the contribution of variety i is equivalent with $\tau_i + (\tau\gamma)_{ig}$. Then within a group of sites the best estimators for the variety parameters can be obtained as described above. Next the estimates from the different groups have to be averaged with weights $\tilde{w}_g$. This is only possible if a variety occurs in every group.

If the model contains random block terms, site terms and interaction terms, the model can be written as

$$Y_{ijkl} = \mu + \tau_i + B_{j(k)} + L_k + (TL)_{ik} + E_{ijkl}\,. \tag{3.21}$$

$B_{j(k)}$, L_k, $(TL)_{ik}$ and E_{ijkl} are (mutually) uncorrelated random variables with expectation zero and variances $\sigma^2_{B(k)}$, σ^2_L, σ^2_{TL}, σ^2, respectively. The expectation of Y_{ijkl} is equal to $\mu + \tau_i$, and the value of variety i is defined as $\mu + \tau_i$. Written in matrix notation, the variance/covariance submatrix (divided by σ^2) of the observations at site k is equal to

$$\mathbf{V}_k = \mathbf{X}_k\mathbf{X}'_k\frac{\sigma^2_{TL}}{\sigma^2} + \mathbf{1}_{n_k}\mathbf{1}'_{n_k}\frac{\sigma^2_L}{\sigma^2} + \mathbf{Z}_k\mathbf{Z}'_k\frac{\sigma^2_{B(k)}}{\sigma^2} + \mathbf{I}_{n_k}\,.$$

We now can proceed as described in the last part of section *3.2.4*, with $\mathbf{V}_k$ defined as above. Notice that $\mathbf{V}_k$ can be written as $\mathbf{Z}_k\mathbf{Z}'_k\sigma^2_{B(k)}/\sigma^2 + \mathbf{I}_{n(k)} + \mathbf{X}_k\mathbf{A}_k\mathbf{X}'_k$, with $\mathbf{A}_k = \mathbf{I}_t\sigma^2_{TL}/\sigma^2 + \mathbf{1}_t\mathbf{1}'_t\sigma^2_L/\sigma^2$. Consequently, $\hat{\tau}$ can be calculated as a multivariately weighted average of $\tilde{\tau}_k$, with $\tilde{\tau}_k$ the estimator at site k, using a mixed additive model with random block terms.

Sometimes it is reasonable to extend the model with fixed replication terms within each site, fixed terms of groups of sites and fixed variety × group interaction terms. Then first estimation of $\mu + \tau_i$ takes place for each separate group, using a model with fixed replication terms. Next these estimates are averaged with weights following from the definition of a variety value.

If there are two or more years of investigation, the breeder often wants to combine the results of these years also. The years are almost always considered to be random terms; the sampled years being a random set representing the climate of the region. In this case the models discussed can be extended by introducing random variables representing the year contribution and the variety × year interaction contributions. The introduction of extra random variables in the model causes the variance/covariance matrix to change. However, it is still possible to write this variance/covariance matrix as $\mathbf{V}\sigma^2$. Thus we can use the methods of analysis described above, of course using the correct $\mathbf{V}_k$ matrices.

Analogous to the discussion whether to use random or fixed site terms, we could classify certain groups of years as fixed, for instance dry , normal and wet years. These groups can then be included in the model as fixed terms. The years within each group can be seen as a random sample of years, and thus be included as random terms in the model.

3.2.6 Traditional analyses

In literature, most of the difficulties with respect to the combination of several experiments relate to the appropriateness of tests of significance for variety parameters and variety × site interactions (see e.g. Yates & Cochran, 1938; Cochran & Cox, 1957). However, in the plant breeding context these tests are of little importance. We know that the varieties differ from each other and we can hardly imagine that every variety reacts the same way to the different environments of the various sites. Less attention has been paid to the determination of best estimators of contrasts between variety parameters. The usual procedure is to first calculate least squares estimates of the variety values at each site. These variety value estimates are the entries of a variety × site table, which is often incomplete. The results of the separate sites have to be compared before they are combined. Cochran & Cox (1957) advise to check whether the differences among variety parameters are the same in each trial and whether there is a consistent superiority of certain varieties. If there are sites where the results are completely different from the other sites, then the constitution of the region should be reconsidered.

To combine the results of the various sites, the variety × site table is then analysed as an experiment with two factors : varieties and sites. Thus this table is analysed as if it reflects a trial with varieties and blocks, where the sites are equivalent with the blocks. The entries in the table are the observations, and it is assumed that the variance/covariance matrix of these observations is diagonal. So it is assumed that the observations are uncorrelated. For observations from different sites this is true, but for observations (read : variety value estimates) from the same site this is only true for orthogonal designs. With orthogonal designs the variety contrasts expectation subspace is orthogonal to the block contrasts expectation subspace. In the plant breeding practice the designs are almost never orthogonal, because of the chosen design itself or because of missing observations. The two stage procedure is convenient to work with because the joint observations from all sites do not have to be analysed as a whole. If also various years are involved, the statistical analysis of the data could be based on a model which includes both site and year terms and the interactions of both with variety. However, for convenience sake often the variety × site table for each year is analysed and a new variety × year table is produced and analysed (Silvey, 1978).

In the plant breeding practice the sites are often chosen in a strictly non-random way in order to reflect differences in soil type, climate conditions, etc.. As a result, the variability between sites is large. Therefore it is often assumed that the ratio σ_L^2/σ^2 approaches infinity. The variance/covariance matrix of the observations in case of random site terms and interaction terms is equal to

$$\mathbf{D}[\mathbf{Y}] = \mathbf{ZZ}'\sigma_L^2 + \mathbf{I}_n\sigma_{TL}^2 + \mathbf{I}_n\sigma^2 = \mathbf{V}\sigma^2,$$

with $\mathbf{V} = \mathbf{ZZ}'\dfrac{\sigma_L^2}{\sigma^2} + \mathbf{I}_n\dfrac{\sigma_{TL}^2 + \sigma^2}{\sigma^2}$.

$\mathbf{Z}$ is the design matrix of the "blocks", here equivalent with the sites. Now the inverse of the matrix $\mathbf{V}$ can be written as

$$\mathbf{V}^{-1} = \frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2}\left(\mathbf{I}_n - \frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2} + \frac{\sigma^2}{\sigma_L^2}\mathbf{I}_b\right)^{-1}\mathbf{Z}'\right).$$

If σ^2/σ_L^2 is neglectable with respect to $\sigma^2/(\sigma_{TL}^2 + \sigma^2)$, then $\mathbf{V}^{-1}$ simplifies to

$$\mathbf{V}^{-1} = \frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2}\left(\mathbf{I}_n - \frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2}\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2}\right)^{-1}\mathbf{Z}'\right)$$

$$= \frac{\sigma^2}{\sigma_{TL}^2 + \sigma^2}\mathbf{R}_\mathbf{Z}.$$

But then $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$ is equal to $(\mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}_\mathbf{Z}\mathbf{Y}$. Thus then the variety × site table also could have been analysed with the fixed additive model. Notice that this is the case if the ratio $(\sigma_{TL}^2 + \sigma^2)/\sigma_L^2$ approaches zero. Then both ratios σ_{TL}^2/σ_L^2 and σ^2/σ_L^2 have to approach zero. So it is not sufficient if the site variance is much larger than the interaction variance, as stated in Patterson & Silvey (1980). In the plant breeding practice the differences between the sites often provide little information on contrasts between variety values, and therefore the fixed model is mostly used. The least squares procedure on incomplete data sets was first applied by Yates (1933), using his fitting constants technique. An (at first sight) attractive option is to calculate weighted least squares estimates. The weights can then be chosen inversely proportional to the estimated error variance of the trial. Then experiments which are precise have a greater weight than experiments with a relatively large error variance. However, it is advised (Dyke, 1988) not to use weighted estimates with estimated weights, because (1) the weights themselves are estimates and therefore subject to sampling error, (2) there may be a correlation between the intrinsic variability of a site and its responsiveness to one of the varieties. This may lead to biased estimates of the variety values (Patterson & Silvey, 1980).

In the previous sections of this chapter we have shown that it is possible to obtain the best estimators of contrasts between variety values without analysing the joint observations of all sites as a whole. The analysis can be done in stages. But unlike the traditional analyses, we use the correct variance/covariance matrix of the observations. In the plant breeding practice, where the differences between varieties are only small, use of the best estimator is very important. We will give a short review of the results obtained in the preceding sections.

The first stage is to analyse the individual trials at the various sites, using a fixed additive model or a mixed additive model. For each site best estimators for contrasts between variety parameters can be determined. The second stage depends on the type of model used for the joint observations of the experiments. In case of an interaction model with fixed site terms, the BLUE of a contrast between variety values appears to be a univariately weighted average of the estimators from the separate sites, the weights given by the definition of a variety value in the presence of interaction. This definition has to be given by the breeder. If the sites are of equal importance to the plant breeder, equal weights $1/m$ will be chosen. However, if the variety × site table is incomplete, not all contrasts between variety values can be estimated. Since incomplete variety × site tables are almost inevitable in the plant breeding practice, we often cannot use an interaction model with fixed site terms.

For all other models the BLUE of a contrast between variety parameters (or variety values) in general appears to be a multivariately weighted average of estimators from the separate sites. Depending on the model used the latter estimators correspond to a fixed additive model or a mixed additive model with random block terms for the observations at the separate sites. The weight matrices $\mathbf{W}_k$ $(k = 1,\ldots,m)$ can be calculated with information from the individual trials at the various sites. Hence it is not necessary to analyse the experiment including varieties, blocks within sites and sites as a whole. There are situations when the multivariate weights reduce to univariate weights. In that case the BLUE of a contrast between variety parameters is a univariate weighted average of the estimators from the separate sites, the weights determined by the variance of the latter estimators. Univariate weights w_k $(k = 1,\ldots,m)$ can be used for the BLUE of a specific contrast $\mathbf{p}'\tau$ if and only if $\mathbf{p}$ is a common eigenvector of all $\mathbf{W}'_k$, with corresponding eigenvalue w_k.

For the fixed additive model we have shown that if the BLUEs of all contrasts between variety parameters have to be a univariately weighted average of the estimators from the separate sites, then the $\mathbf{C}_k$ matrices have to be proportional to each other. If all the trials at the various sites are variance-balanced, this is the case. The $\mathbf{C}_k$ matrices are also proportional (read : identical) if the trials at the various sites have identical designs. Then the univariate weights are equal to $1/m$.

If we compare the BLUE of a contrast between variety values in the interaction model with fixed site terms and all other models, we notice that in the latter models an estimator at a certain site with a small variance (e.g. caused by a large number of replications at a certain site) receives a larger weight than such an estimator with a large variance. This is not the case in the interaction model with fixed site terms, where the weights are determined by the definition of a variety value in the presence of interaction. This definition should not be based on variances of estimators at the individual sites.

3.3 Combining subtrials that are connected by control varieties only

In this section we return to the estimation of contrasts between variety parameters from the observations at a single site and year. As described in chapter 2, the trials in the sugar beet breeding practice often have a specific structure at one site. The set of new varieties is subdivided into disjunct subsets and the varieties belonging to a certain subset are tested in a separate trial. To connect all trials with each other, a small number of control varieties is included in each trial. We will further describe the design of the experiment in *3.3.1*, together with the standard analyses, using a fixed additive model. Because often the number of varieties to be tested is very large, it would be convenient if the BLUEs of contrasts between variety parameters could be determined from the analyses of the separate trials. That this is possible is described in *3.3.2*. There are situations where the estimators of specific contrasts between variety parameters, calculated at a separate trial, are identical to the estimators from the whole experiment. In *3.3.3* it is described for which designs of the trials this is the case. Examples illustrating the presented theory are given in *3.3.4*. In *3.3.5* we will briefly discuss the situation where the chosen model is not completely fixed, but where block terms are subdivided into fixed replication terms and random blocks terms within each replication. This situation can e.g. occur when the separate trials have resolvable designs.

3.3.1 Design of a concatenated trial and its traditional analysis

Suppose t new varieties have to be tested. These varieties are divided over $m < t$ trials, where t_k $(k = 1, \ldots, m)$ new varieties are included in trial k (with $\Sigma\, t_k = t$). Additionally, the same c control varieties are included in each trial. By including these control varieties the trials become connected, so it is possible to analyse the m trials as one large incomplete block experiment. This large trial is a concatenation of trials and therefore this type of trial will be denoted by 'concatenated trial'. To distinguish the separate trials from the concatenated trial we will denote these by 'subtrials'. The subtrials, with $t_k + c$ varieties, may have any design. In practice we often encounter the following easy going method of analysis : the subtrials are analysed separately and the least squares estimates of the variety values are ranked. The estimates of the values of the new varieties are expressed relatively to the estimate of the average value of the control varieties (in the separate subtrial), in order to take account of the different fertility levels of the subtrials.

It would be better to analyse the m subtrials as one large incomplete blocks experiment. Of course the variances of the estimators of contrasts between variety parameters will differ notably, but all comparisons between varieties can be made. This is important if we want to use statistical selection methods to make a selection. As mentioned before the variety parameters themselves are not uniquely estimable, but for ranking and selection purposes the estimation of contrasts between the variety parameters is sufficient. We will concentrate on the estimation of the contrast between the parameter of a new variety and the average of the parameters corresponding to the control varieties. This contrast is specifically informative because a plant breeder also wants to know whether a new variety is better than the average of the control varieties.

The concatenated trials found in the plant breeding practice are far too large to be used as an example. We will describe an introductory example of a concatenated trial. Assume a concatenated trial with $m = 3$ subtrials and $c = 2$ control varieties. Subtrial 1 has a randomised complete block design with 2 complete blocks and 4 varieties $(t_1 + c = 2 + 2 = 4)$ per block. The second subtrial has 4 varieties $(t_2 + c = 2 + 2 = 4)$ in a balanced 2x2 lattice design, so this design has 6 incomplete blocks each containing 2 varieties. There are 3 replications per variety. The third subtrial has 9 varieties $(t_3 + c = 7 + 2 = 9)$ in a partially balanced 3x3 lattice design with 2 replications. Consequently, there are 6 incomplete blocks each containing 3 varieties. In all three subtrials varieties 1 and 2 are the control varieties, and the three subtrials are connected by these 2 control varieties only. The schematical presentation of this concatenated trial is :

		subtrial 1		subtrial 2						subtrial 3					
variety	block :	1	2	3	4	5	6	7	8	9	10	11	12	13	14
control 1		•	•	•		•		•		•			•		
control 2		•	•	•			•		•	•				•	
3		•	•												
4		•	•												
5					•	•			•						
6					•		•	•							
7										•					•
8											•		•		
9											•			•	
10											•				•
11												•	•		
12												•		•	
13												•			•

If we want to analyse the subtrials as one large experiment with incomplete blocks, we can use a model for the observations that is analogous to model (3.1) :

$$Y_{ijk} = \lambda + \tau_i + \beta_j + E_{ijk} , \quad i = 1, \dots, t + c \ , \ j = 1, .., b \ , \ k = 1, \dots, n_{ij} .$$

Now we have $t + c$ varieties, with variety parameters τ_i. The λ, β_j and the E_{ijk} are defined as in (3.1). To calculate estimates of contrasts between variety parameters we could make use of the reduced normal equations as described in **3.1**. In matrix notation, the contrast between the parameter of a new variety and the average of the parameters of the control varieties will in the sequel be denoted by $\mathbf{b}'\tau$. Thus the BLUE of $\mathbf{b}'\tau$ is equal to $\mathbf{b}'\mathbf{C}^{-}\mathbf{Q}$. Often hundreds of varieties are included in a concatenated trial. This means that the calculation of a pseudo-inverse of **C** can become troublesome, especially when the calculation is done on a personal computer. To avoid these problems, we could estimate the contrast $\mathbf{b}'\tau$ only from the subtrial in which the new variety is included. This method results in an estimator of this contrast that is unbiased, but in general does not have minimum variance. We will denote this type of estimator by 'local estimator' and its outcome by 'local estimate'.

A different approach used to analyse large incomplete block trials is to calculate the estimates of the parameters by an iterative method described by Kuiper (1952, 1983) and later elaborated by Corsten (1967, 1976). These estimates have the restriction that $\Sigma\, n_{i}.\hat{\tau}_i = 0$, with $n_i. = \Sigma_j\, n_{ij}$ (see also the *Remark* in *3.1.1*). In practice the method is very simple; one subtraction and repeated weighted averaging are the only operations to be carried out. Let $\mathbf{R} = \mathbf{R}^{1/2}\mathbf{R}^{1/2}$, with **R** the replication matrix of the reduced normal equations. Then the solution of the reduced normal equations can also be written as

$$\hat{\tau} = \mathbf{R}^{-1/2}(\mathbf{I}_{t+c} - \mathbf{R}^{-1/2}\mathbf{N}\mathbf{K}^{-1}\mathbf{N}'\mathbf{R}^{-1/2})^{-}\mathbf{R}^{-1/2}(\mathbf{T} - \mathbf{N}\mathbf{K}^{-1}\mathbf{B})$$

$$= \mathbf{R}^{-1/2}(\mathbf{I}_{t+c} - \mathbf{S})^{-}\mathbf{R}^{-1/2}\mathbf{Q}\,.$$

The pseudo-inverse of $(\mathbf{I}_{t+c} - \mathbf{S})$ can be calculated as the sum of an infinite geometric progression : $(\mathbf{I}_{t+c} - \mathbf{S})^{-} = \mathbf{I}_{t+c} + \mathbf{S} + \mathbf{S}^2 + \ldots$, with the terms of the progression converging to zero. For connected designs it can be shown that **S** has one eigenvalue equal to 1 and that all other eigenvalues are less than 1, in absolute value. The convergence speed depends on the second largest eigenvalue, which is the largest eigenvalue less than one, of **S**. If this eigenvalue is close to one, the convergence speed will be slow. An eigenvalue close to one indicates that the design is very inorthogonal. The concatenated designs in the plant breeding practice are indeed very inorthogonal, because they are connected by control varieties only. Consider an experimental design that is a concatenation of two triple and two quadruple 5x5 lattices. The lattices are connected with each other by c control varieties. For c = 1, 2 and 3 the eigenvalues of **S** were calculated. The three largest eigenvalues, less than one, are

$c = 1$:	0.9681,	0.9333,	0.8950.
$c = 2$:	0.9677,	0.9328,	0.8942.
$c = 3$:	0.9673,	0.9320,	0.8933.

A concatenated design with only one control variety is more inorthogonal than a concatenated design with two or three control varieties, which results in larger eigenvalues. The largest eigenvalues (less than one) lie close to one, so the iterative calculation of the parameter estimates for concatenated trials is not very satisfactory. The size of the experiment does not cause any trouble here, but because of the inorthogonality of the design a large number of iterations would be necessary to calculate the estimates.

3.3.2 Combining the estimators from the separate subtrials

In practice the analysis of the joint observations of all the subtrials in one step is not convenient. Therefore we will describe a two stage procedure to obtain the BLUEs of $\mathbf{b}'\tau$. The first step in this two stage procedure is to calculate a solution of the normal equations at each individual subtrial. In the second stage these results are combined to calculate the best estimates.

First we will focus on a single subtrial, hence for the time being we will not use the subscript k to denote the subtrial. The calculation of a solution of the normal equations has already been explained in *3.1*. For later use we will now give a different way of calculating a solution. The vector τ with variety parameters at a single subtrial is subdivided into two vectors, namely vector α with c control variety parameters and vector γ with t new variety parameters. Let $\mathbf{G}$ be the design matrix associated with the control varieties, which appear at every subtrial. The rank of $\mathbf{G}$ is equal to c. Let $\mathbf{L}$ be the design matrix associated with the local new varieties, which only appear at the subtrial we focus on. The rank of $\mathbf{L}$ is equal to t. As in **3.1**, let $\mathbf{Z}$ be the design matrix of the blocks and let β be the vector of block parameters. The additive model for the n, say, observations at a single subtrial can now be written as

$$E[\mathbf{Y}] = \mathbf{1}_n\lambda + \mathbf{G}\alpha + \mathbf{L}\gamma + \mathbf{Z}\beta$$

$$= \mathbf{R_Z G}\alpha + \mathbf{P_Z G}\alpha + \mathbf{R_Z L}\gamma + \mathbf{P_Z L}\gamma + \mathbf{Z}(\beta + \mathbf{1}_b\lambda)\,,$$

with $\mathbf{P_Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$; $\mathbf{R_Z} = \mathbf{I}_n - \mathbf{P_Z}$,

$$
\begin{aligned}
E[\mathbf{Y}] &= \mathbf{R_Z G}\alpha + \mathbf{R_Z L}\gamma + \mathbf{Z}\{(\mathbf{Z'Z})^{-1}\mathbf{Z'G}\alpha + (\mathbf{Z'Z})^{-1}\mathbf{Z'L}\gamma + \beta + \mathbf{1}_b\lambda\} \\
&= \mathbf{R_Z G}\alpha + \mathbf{R_Z L}\gamma + \mathbf{Z}\beta^* \\
&= \mathbf{P}_{\mathbf{L/Z}}\mathbf{R_Z G}\alpha + \mathbf{R}_{\mathbf{L/Z}}\mathbf{R_Z G}\alpha + \mathbf{R_Z L}\gamma + \mathbf{Z}\beta^* ,
\end{aligned}
$$

with $\mathbf{L/Z} = \mathbf{R_Z L}$, $\mathbf{P}_{\mathbf{L/Z}} = \mathbf{R_Z L}(\mathbf{L'R_Z L})^{-1}\mathbf{L'R_Z}$ and $\mathbf{R}_{\mathbf{L/Z}} = \mathbf{I}_n - \mathbf{P}_{\mathbf{L/Z}}$,

$$
\begin{aligned}
&= \mathbf{R}_{\mathbf{L/Z}}\mathbf{R_Z G}\alpha + \mathbf{R_Z L}\{\gamma + (\mathbf{L'R_Z L})^{-1}\mathbf{L'R_Z G}\alpha\} + \mathbf{Z}\beta^* \\
&= \mathbf{R}_{\mathbf{L/Z}}\mathbf{R_Z G}\alpha + \mathbf{R_Z L}\gamma^* + \mathbf{Z}\beta^* . \qquad (3.22)
\end{aligned}
$$

Because of the orthogonalisation in (3.22), the normal equations are very simple:

$$
\begin{pmatrix} \mathbf{G'R_Z R_{L/Z} R_Z G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L'R_Z L} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z'Z} \end{pmatrix}
\begin{pmatrix} \tilde{\alpha} \\ \tilde{\gamma}^* \\ \tilde{\beta}^* \end{pmatrix}
=
\begin{pmatrix} \mathbf{G'R_Z R_{L/Z} Y} \\ \mathbf{L'R_Z Y} \\ \mathbf{Z'Y} \end{pmatrix} .
$$

With these normal equations a solution of $\tilde{\alpha}$ and $\tilde{\gamma}^*$ can be calculated as

$$\tilde{\alpha} = (\mathbf{G'R_Z R_{L/Z} R_Z G})^{-}\mathbf{G'R_Z R_{L/Z} Y} , \qquad (3.23)$$

$$\tilde{\gamma}^* = (\mathbf{L'R_Z L})^{-1}\mathbf{L'R_Z Y} ,$$

$$\tilde{\gamma} = (\mathbf{L'R_Z L})^{-1}\mathbf{L'R_Z Y} - (\mathbf{L'R_Z L})^{-1}\mathbf{L'R_Z G}\tilde{\alpha} .$$

The desired contrast $\mathbf{b}'\tau$, denoted here by $\mathbf{p}'\alpha + \mathbf{q}'\gamma$, can now be estimated as $\mathbf{p}'\tilde{\alpha} + \mathbf{q}'\tilde{\gamma}$. Notice that $\mathbf{p}'\alpha$ and $\mathbf{q}'\gamma$ themselves are not contrasts of parameters.

The **C** matrix of the subtrial (corresponding to the reduced normal equations) can be written as

$$
\mathbf{C} = \begin{pmatrix} \mathbf{G'R_Z G} & \mathbf{G'R_Z L} \\ \mathbf{L'R_Z G} & \mathbf{L'R_Z L} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} .
$$

So $\mathbf{G'R_Z R_{L/Z} R_Z G}$ can be calculated as

$$
\begin{aligned}
\mathbf{G'R_Z R_{L/Z} R_Z G} &= \mathbf{G'R_Z G} - \mathbf{G'R_Z L}(\mathbf{L'R_Z L})^{-1}\mathbf{L'R_Z G} \\
&= \mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21} .
\end{aligned}
$$

The inverse of $\mathbf{L'R_Z L} = \mathbf{C}_{22}$ exists, because $\mathbf{C}_{22}$ is non-singular. This can be seen as follows. If $\mathbf{C}_{22}$ was singular, there would exist a vector **d**, other than the null vector, for which $\mathbf{C}_{22}\mathbf{d} = \mathbf{0}$ is true. But then, since $\mathbf{C}_{22} = (\mathbf{0} \quad \mathbf{I}_t)\mathbf{C}(\mathbf{0} \quad \mathbf{I}_t)'$ and **C** is non-negative definite, $\mathbf{C}(\mathbf{0} \quad \mathbf{I}_t)'\mathbf{d} = \mathbf{0}$ must also be true. Because we are dealing with connected designs, $(\mathbf{0} \quad \mathbf{I}_t)'\mathbf{d}$ must then be a multiple of the unity vector, which is clearly not possible for $\mathbf{d} \neq \mathbf{0}$. We will now show that $\mathbf{G'R_ZR_{L/Z}R_ZG}$ is a singular, doubly-centred matrix analogous to $\mathbf{C}$. If singularities exist there is a vector $\mathbf{d}$, other than the null vector, for which $(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21})\mathbf{d} = \mathbf{0}$ is true. Then

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ -\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{d} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11}\mathbf{d} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{d} \\ \mathbf{C}_{21}\mathbf{d} - \mathbf{C}_{22}\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{d} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

A design with incomplete blocks is connected if and only if the rank of $\mathbf{C}$ is $t + c - 1$ (Dey, 1986). This means that, because $\mathbf{C1}_{t+c} = \mathbf{0}$, there exists no other vector $\mathbf{d}$ ($\mathbf{d} \neq a\mathbf{1}_{t+c}$, $a \in \mathcal{R}$) for which $\mathbf{Cd} = \mathbf{0}$ is true. Hence, because the $\mathbf{C}$ matrix corresponds to a connected design,

$$\begin{pmatrix} \mathbf{d} \\ -\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{d} \end{pmatrix}$$

must be a multiple of the unity vector. Hence $\mathbf{d}$ itself is also a multiple of $\mathbf{1}_c$. If $\mathbf{d}$ is a multiple of $\mathbf{1}_c$, say $d\mathbf{1}_c$, then

$$(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21})d\mathbf{1}_c = d(\mathbf{C}_{11}\mathbf{1}_c + \mathbf{C}_{12}\mathbf{1}_t) = \mathbf{0}\ ,$$

because with a connected design $\mathbf{C}_{21}\mathbf{1}_c + \mathbf{C}_{22}\mathbf{1}_t = \mathbf{0}$ and therefore $\mathbf{1}_t = -\mathbf{C}_{22}^{-1}\mathbf{C}_{21}\mathbf{1}_c$. Hence the matrix $\mathbf{G'R_ZR_{L/Z}R_ZG}$ has rank $c - 1$ and the only singularity is $\mathbf{G'R_ZR_{L/Z}R_ZG1}_c = \mathbf{0}$.

We now will describe the estimation of $\mathbf{b}'\tau$ from the joint observations of the m subtrials, without analysing the concatenated trial as a whole. We use the subscript k to indicate the subtrial. The vector of control variety parameters is identical for all subtrials. The vectors of parameters for new varieties are different at all subtrials, just as the vectors of block parameters. We can write the usual linear model for the joint observations of the m subtrials (thus for the concatenated trial as a whole) as

$$E\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_m \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{\mathbf{L}_1/\mathbf{Z}_1}\mathbf{R}_{\mathbf{Z}_1}\mathbf{G}_1 \\ \mathbf{R}_{\mathbf{L}_2/\mathbf{Z}_2}\mathbf{R}_{\mathbf{Z}_2}\mathbf{G}_2 \\ \vdots \\ \mathbf{R}_{\mathbf{L}_m/\mathbf{Z}_m}\mathbf{R}_{\mathbf{Z}_m}\mathbf{G}_m \end{pmatrix}\alpha + \begin{pmatrix} \mathbf{R}_{\mathbf{Z}_1}\mathbf{L}_1 & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{R}_{\mathbf{Z}_2}\mathbf{L}_2 & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{R}_{\mathbf{Z}_m}\mathbf{L}_m \end{pmatrix}\begin{pmatrix} \gamma_1^* \\ \gamma_2^* \\ \vdots \\ \gamma_m^* \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{Z}_2 & & \\ & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & \mathbf{Z}_m \end{pmatrix}\begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_m^* \end{pmatrix}. \tag{3.24}$$

The normal equations then read :

$$\begin{pmatrix} \sum\limits_{k=1}^{m}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k & & & & & & \text{-}\mathbf{0}\text{-} \\ & \mathbf{L}'_1\mathbf{R}_{\mathbf{Z}_1}\mathbf{L}_1 & & & & & \\ & & \ddots & & & & \\ & & & \mathbf{L}'_m\mathbf{R}_{\mathbf{Z}_m}\mathbf{L}_m & & & \\ & & & & \mathbf{Z}'_1\mathbf{Z}_1 & & \\ & & & & & \ddots & \\ \text{-}\mathbf{0}\text{-} & & & & & & \mathbf{Z}'_m\mathbf{Z}_m \end{pmatrix} \begin{pmatrix} \hat{\alpha} \\ \hat{\gamma}_1^* \\ \vdots \\ \hat{\gamma}_m^* \\ \beta_1^* \\ \vdots \\ \beta_m^* \end{pmatrix} = \begin{pmatrix} \sum\limits_{k=1}^{m}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{Y}_k \\ \mathbf{L}'_1\mathbf{R}_{\mathbf{Z}_1}\mathbf{Y}_1 \\ \vdots \\ \mathbf{L}'_m\mathbf{R}_{\mathbf{Z}_m}\mathbf{Y}_m \\ \mathbf{Z}'_1\mathbf{Y}_1 \\ \vdots \\ \mathbf{Z}'_m\mathbf{Y}_m \end{pmatrix}.$$

A solution of the normal equations for $\hat{\alpha}$ and $\hat{\gamma}_k$ $(k = 1, \ldots, m)$ can be calculated as

$$\hat{\alpha} = \left(\sum_{k=1}^{m}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\right)^{-}\sum_{k=1}^{m}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{Y}_k$$

$$= \sum_{k=1}^{m}\left(\sum_{k=1}^{m}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\right)^{-}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\tilde{\alpha}_k = \sum_{k=1}^{m}\mathbf{W}_k\tilde{\alpha}_k\,, \qquad (3.25)$$

$$\hat{\gamma}_k^* = \left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{Y}_k = \tilde{\gamma}_k^*\,,$$

$$\hat{\gamma}_k = \left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{Y}_k - \left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\hat{\alpha}$$

$$= \tilde{\gamma}_k + \left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k(\tilde{\alpha}_k - \hat{\alpha})\,.$$

So a solution for the least squares estimates of the control variety parameters can be calculated as a multivariately weighted average of the $\tilde{\alpha}_k$ from the separate subtrials. The weight matrices

$$\mathbf{W}_k = \left(\sum_{k=1}^{m}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\right)^{-}\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k$$

can be calculated with the use of the ordinary **C** matrices from the separate subtrials, because

$$\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k = \mathbf{C}_{11_k} - \mathbf{C}_{12_k}\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\,.$$

Analogous to the weight matrices in *3.2.2*, the weights matrices in this section also have the property that $\Sigma\,\mathbf{W}_k = \mathbf{I}_c - (1/c)\mathbf{1}_c\mathbf{1}'_c$. Because $\mathbf{G}'_k\mathbf{R}_{\mathbf{Z}k}\mathbf{R}_{\mathbf{L}k/\mathbf{Z}k}\mathbf{R}_{\mathbf{Z}k}\mathbf{G}_k$ is a doubly-centred matrix, $\mathbf{W}_k\mathbf{1}_c = \mathbf{0}$. The estimate of the contrast between the parameter of a new variety in subtrial k $(k = 1, \ldots, m)$ and the average of the parameters of the control varieties can be calculated as $\mathbf{p}'\hat{\alpha} + \mathbf{q}'\hat{\gamma}_k$.

The pseudo-variance/covariance matrix of $(\hat{\alpha} \quad \hat{\gamma}_k^*)'$ can be calculated as

$$\dot{\mathbf{D}}\left[\begin{pmatrix}\hat{\alpha}\\ \hat{\gamma}^{*}_{k}\end{pmatrix}\right]=\begin{pmatrix}\left(\sum\limits_{k=1}^{m}\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_{k}\right)^{-} & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\right)^{-1}\end{pmatrix}\sigma^{2}.$$

Because

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\gamma}_{k}\end{pmatrix}=\begin{pmatrix}\mathbf{I}_{c} & \mathbf{0}\\ -\left(\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\right)^{-1}\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_{k} & \mathbf{I}_{t}\end{pmatrix}\begin{pmatrix}\hat{\alpha}\\ \hat{\gamma}^{*}_{k}\end{pmatrix},$$

the pseudo-variance/covariance matrix of $(\hat{\alpha} \quad \hat{\gamma}_k)'$ can be calculated as

$$\dot{\mathbf{D}}\left[\begin{pmatrix}\hat{\alpha}\\ \hat{\gamma}_{k}\end{pmatrix}\right]=\begin{pmatrix}\mathbf{D}_{11} & \mathbf{D}_{12}\\ \mathbf{D}_{21} & \mathbf{D}_{22}\end{pmatrix},$$

with

$$\mathbf{D}_{11}=\left(\sum\limits_{k=1}^{m}\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_{k}\right)^{-}$$

$$=\left[\sum\limits_{k=1}^{m}\left(\mathbf{C}_{11_k}-\mathbf{C}_{12_k}\mathbf{C}^{-1}_{22_k}\mathbf{C}_{21_k}\right)\right]^{-},$$

$$\mathbf{D}_{12}=-\left(\sum\limits_{k=1}^{m}\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_{k}\right)^{-}\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\left(\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\right)^{-1}$$

$$=-\left[\sum\limits_{k=1}^{m}\left(\mathbf{C}_{11_k}-\mathbf{C}_{12_k}\mathbf{C}^{-1}_{22_k}\mathbf{C}_{21_k}\right)\right]^{-}\mathbf{C}_{21_k}\mathbf{C}^{-1}_{22_k},$$

$$\mathbf{D}_{21}=\mathbf{D}'_{12},$$

$$\mathbf{D}_{22}=\left(\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\right)^{-1}\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_{k}\left(\sum\limits_{k=1}^{m}\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_{k}\right)^{-}\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\left(\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\right)^{-1}+\left(\mathbf{L}'_{k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_{k}\right)^{-1}$$

$$=\mathbf{C}^{-1}_{22_k}\mathbf{C}_{12_k}\left[\sum\limits_{k=1}^{m}\left(\mathbf{C}_{11_k}-\mathbf{C}_{12_k}\mathbf{C}^{-1}_{22_k}\mathbf{C}_{21_k}\right)\right]^{-}\mathbf{C}_{21_k}\mathbf{C}^{-1}_{22_k}+\mathbf{C}^{-1}_{22_k}.$$

Analogous to the situation in *3.2.2*, the multivariate weights will reduce to univariate weights if all $\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}k}\mathbf{R}_{\mathbf{L}k/\mathbf{Z}k}\mathbf{R}_{\mathbf{Z}k}\mathbf{G}_{k}$ matrices are proportional to each other. In that case the weights can be calculated very easily in the same way as described in section *3.2.2*. Matrix $\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}k}\mathbf{R}_{\mathbf{L}k/\mathbf{Z}k}\mathbf{R}_{\mathbf{Z}k}\mathbf{G}_{k}$ is a symmetric matrix for which $\mathbf{G}'_{k}\mathbf{R}_{\mathbf{Z}k}\mathbf{R}_{\mathbf{L}k/\mathbf{Z}k}\mathbf{R}_{\mathbf{Z}k}\mathbf{G}_{k}\mathbf{1}_{c}=\mathbf{0}$ must be true. Therefore, in case of two control varieties this matrix must be of the following form :

$$d_{k}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.$$

Consequently, when there are two control varieties the weights are univariate and can be calculated as

$$w_k = \frac{d_k}{\sum\limits_{k=1}^{m} d_k},$$

or, using the well known variance criterion, as

$$w_k = \frac{[\mathrm{var}(\mathbf{c}'\tilde{\alpha}_k)]^{-1}}{\sum\limits_{k=1}^{m} [\mathrm{var}(\mathbf{c}'\tilde{\alpha}_k)]^{-1}}, \text{ with } \mathbf{c}' = (1 \quad -1)\,.$$

Besides estimation of $\mathbf{b}'\tau$ also the estimation of the error variance σ^2 is important. In general the SS_E of the concatenated trial is not equal to the sum of the error sums of squares from the individual subtrials. Similarly the degrees of freedom for error corresponding to the concatenated trial cannot be calculated in general by summation of the separate degrees of freedom for error from the subtrials. If we have m subtrials with c control varieties then $(m-1)(c-1)$ degrees of freedom corresponding to the between control varieties sum of squares in the separate subtrials correspond to the SS_E of the concatenated trial. So df_e in the concatenated trial is equal to $df_e = n - \Sigma t_k - c - \Sigma b_k + 1$. When there is only one control variety, there is no between control varieties sum of squares and the df_e for the concatenated trial can be calculated by summation of the df_e's from the subtrials. The vector of residuals corresponding to subtrial k $(k = 1, \ldots, m)$ can be calculated by subtracting the fitted values of $\mathbf{Y}_k$ from $\mathbf{Y}_k$ itself :

$$\begin{aligned}
\hat{\mathbf{E}}_k &= \mathbf{Y}_k - \mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\hat{\alpha} - \mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\hat{\gamma}_k^* - \mathbf{Z}_k\beta_k^* \\
&= \mathbf{Y}_k - \mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\hat{\alpha} - \mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{Y}_k - \mathbf{Z}_k(\mathbf{Z}'_k\mathbf{Z}_k)^{-1}\mathbf{Z}'_k\mathbf{Y}_k \\
&= \mathbf{R}_{\mathbf{L}_k/\mathbf{Z}_k}\mathbf{R}_{\mathbf{Z}_k}(\mathbf{Y}_k - \mathbf{G}_k\hat{\alpha})\,.
\end{aligned}$$

Now the error sum of squares can be calculated as $SS_E = \Sigma\, \hat{\mathbf{E}}'_k\hat{\mathbf{E}}_k$. The error variance can then be estimated by SS_E/df_e. The above steps can be carried out per subtrial, so the error variance can be estimated by analysis of the subtrials only.

The type of experiment studied in this section can be compared with the situation of variety trials at different sites, when some varieties are only grown at one site, because of their local importance. Then the local estimate can be improved by using information about the other variety parameters at other sites.

3.3.3 When is the local estimator equal to the BLUE ?

The contrast between the parameter of a new variety at subtrial k and the average of the parameters of the control varieties was denoted by $\mathbf{p}'\alpha+\mathbf{q}'\gamma$. This contrast can be estimated from the observations of subtrial k as $\mathbf{p}'\tilde{\alpha}_k+\mathbf{q}'\tilde{\gamma}_k$. From the concatenated trial as a whole the desired contrast can be estimated as $\mathbf{p}'\hat{\alpha}+\mathbf{q}'\hat{\gamma}_k$. The difference between these two estimates can be calculated as

$$\begin{aligned}\mathbf{p}'(\tilde{\alpha}_k-\hat{\alpha})+\mathbf{q}'(\tilde{\gamma}_k-\hat{\gamma}_k) &= \mathbf{p}'(\tilde{\alpha}_k-\hat{\alpha})-\mathbf{q}'\left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k(\tilde{\alpha}_k-\hat{\alpha})\\ &= \left(\mathbf{p}'-\mathbf{q}'\left(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k\right)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\right)(\tilde{\alpha}_k-\hat{\alpha})\\ &= \left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)\left([\mathbf{I}_c-\mathbf{W}_k]\tilde{\alpha}_k-\sum_{j=1,j\neq k}^{m}\mathbf{W}_j\tilde{\alpha}_j\right).\end{aligned}$$

Thus the outcome of the BLUE of $\mathbf{p}'\alpha+\mathbf{q}'\gamma_k$ can be calculated as $\mathbf{p}'\tilde{\alpha}_k+\mathbf{q}'\tilde{\gamma}_k-\left(\mathbf{p}'-\mathbf{q}'(\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{L}_k)^{-1}\mathbf{L}'_k\mathbf{R}_{\mathbf{Z}_k}\mathbf{G}_k\right)(\tilde{\alpha}_k-\hat{\alpha})$. We know that the expectation of the difference between the local estimator and the best estimator is zero, because both estimators are unbiased. We also know that $\Sigma\,\mathbf{W}_k=\mathbf{I}_c-(1/c)\mathbf{1}_c\mathbf{1}'_c$. Hence

$$\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)\left(\mathbf{I}_c-\sum_{j=1}^{m}\mathbf{W}_j\right)\alpha=\frac{1}{c}\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)\mathbf{1}_c\mathbf{1}'_c\alpha=0\,,$$

$$\mathbf{p}'\mathbf{1}_c-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\mathbf{1}_c=0\,.$$

Hence $(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22i}^{-1}\mathbf{C}_{21i})\alpha$ represents a contrast between control variety parameters. Because $\mathbf{W}_j\mathbf{1}_c=\mathbf{0}$, $\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)\mathbf{W}_j\quad(i\neq j)$ and $\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)(\mathbf{I}_c-\mathbf{W}_k)$ also denote a contrast. Further we notice that $(1/c)\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)\mathbf{1}_c\mathbf{1}'_c=\mathbf{0}'$. So we can conclude that the difference between the local estimator and the best estimator represents a contrast of contrasts between control variety parameters estimated at each subtrial.

If the difference between the local estimator and the best estimator is zero, then the local estimator is already the BLUE. We will show that the condition for equality of the two estimates is :

$$\operatorname{cov}(\mathbf{c}'\tilde{\alpha}_k,\mathbf{p}'\tilde{\alpha}_k+\mathbf{q}'\tilde{\gamma}_k)=0\,,$$

for all pairwise contrasts $\mathbf{c}'\tilde{\alpha}_k$. It is known that if an estimator T is BLU, then $\operatorname{cov}(T,z)=0$ for all z, where z is a function with expectation equal to zero (Rao, 1973). Hence if the local estimator is the BLUE,

$$\mathrm{cov}(\mathbf{p}'\tilde{\alpha}_k + \mathbf{q}'\tilde{\gamma}_k, \mathbf{d}'(\tilde{\alpha}_k - \hat{\alpha})) = 0\,,$$

$$\mathrm{cov}(\mathbf{p}'\tilde{\alpha}_k + \mathbf{q}'\tilde{\gamma}_k, \mathbf{d}'(\mathbf{I}_c - \mathbf{W}_k)\tilde{\alpha}_k) = 0\,,$$

for all $\mathbf{d}$ with $\mathbf{d}'\mathbf{1}_c = 0$. Now $\mathbf{d}'(\mathbf{I}_c - \mathbf{W}_k)$ is a contrast, and may be denoted by $\mathbf{c}'$. Hence $\mathrm{cov}(\mathbf{p}'\tilde{\alpha}_k + \mathbf{q}'\tilde{\gamma}_k, \mathbf{c}'\tilde{\alpha}_k) = 0$ for each $\mathbf{c}$ for which there exists a $\mathbf{d}$ such that $\mathbf{c} = (\mathbf{I}_c - \mathbf{W}'_k)\mathbf{d}$. Such a $\mathbf{d}$ exists for every $\mathbf{c}$, namely $\mathbf{d} = (\mathbf{I}_c - \mathbf{W}'_k)^{-1}\mathbf{c}$. Because the eigenvalues of $\mathbf{W}_k$ are less than 1, all the eigenvalues of $(\mathbf{I}_c - \mathbf{W}'_k)$ are larger than zero. So $(\mathbf{I}_c - \mathbf{W}'_k)$ is non-singular and the inverse exists. Now suppose that $\mathrm{cov}(\mathbf{p}'\tilde{\alpha}_k + \mathbf{q}'\tilde{\gamma}_k, \mathbf{c}'\tilde{\alpha}_k) = 0$ for all $\mathbf{c}$ with $\mathbf{c}'\mathbf{1}_c = 0$. Then the variance of the BLUE reduces to :

$$\mathrm{var}(\mathbf{p}'\tilde{\alpha}_k + \mathbf{q}'\tilde{\gamma}_k) + \mathrm{var}(\mathbf{d}'(\tilde{\alpha}_k - \hat{\alpha}))\,,$$

with $\mathbf{d}'\mathbf{1}_c = 0$. The minimum of this variance is reached when $\mathbf{d} = \mathbf{0}$, and in that case the BLUE is equal to the local estimator.

In practice the number of control varieties is rarely larger than 3. Using a different approach than described above, we will again study the necessary conditions for which the local estimator is equal to the BLUE, with $c = 1, 2$ or 3. If we want to analyse the concatenated trial, we could use the reduced normal equations : $\mathbf{C}\hat{\tau} = \mathbf{Q}$. For a connected design we know that $\mathbf{1}'_v\mathbf{Q} = \mathbf{1}'_v\mathbf{C}\hat{\tau} = 0$, with v the number of varieties. $\mathbf{Q}$ can be written as $\mathbf{Q} = \mathbf{MY}$, hence $\mathbf{1}'_v\mathbf{MY} = 0 \quad \forall\ \mathbf{Y}$; hence $\mathbf{M}'\mathbf{1}_v = \mathbf{0}$, so the rank of $\mathbf{M}$ has to be smaller than or equal to $v-1$. But if the rank of $\mathbf{M}$ is smaller than $v-1$, then there would exist a vector $\mathbf{d}$ ($\mathbf{d} \neq a\mathbf{1}_v$, $a \in \mathcal{R}$) that satisfies $\mathbf{M}'\mathbf{d} = \mathbf{0}$. Then $\mathbf{d}'\mathbf{MY} = 0 \ \ \forall\ \mathbf{Y}, \mathbf{d}'\mathbf{C}\hat{\tau} = 0 \ \ \forall\ \hat{\tau}$, hence $\mathbf{Cd} = \mathbf{0}$. This is in contradiction with the fact that the rank of $\mathbf{C}$ is $v-1$. So the rank of $\mathbf{M}$ has to be $v-1$ also. Hence $\mathbf{d}'\mathbf{Q} = 0$ only for $\mathbf{d} = a\mathbf{1}$ ($a \in \mathcal{R}$). In other words, the only relationship between the elements of $\mathbf{Q}$ is that they sum up to zero.

The estimate of $\mathbf{b}'\tau$ can be calculated from the reduced normal equations as $\mathbf{b}'\mathbf{C}^{-}\mathbf{Q}$. But because $\mathbf{1}'\mathbf{Q} = 0$, the (say) last $v-1$ normal equations also generate the first equation, hence the v equations are equivalent to the last $v-1$ equations. Hence all estimable contrasts from the v equations are also estimable and have the same solution from the last $v-1$ equations. We know that all contrasts are estimable from the v equations, so all contrasts are also estimable from the last $v-1$ equations.

The number of reduced normal equations for the concatenated trial is $v = c + t$, the total number of varieties. Let the first c normal equations correspond to the control varieties, the next t_1 equations correspond to the new varieties in the first subtrial, and so on. If we compare equations $c+1$ up to and including $c+t_1$ with the last t_1 reduced normal equations for the first subtrial alone, we notice that these equations are identical. So if we only have one control variety, then the local and the best estimates of $\mathbf{b}'\boldsymbol{\tau}$ are equal for any design, because the first normal equation can be generated by the last t_1. In general we can state that the difference between the local estimate and the best estimate of $\mathbf{b}'\boldsymbol{\tau}$ will be zero if the normal equations corresponding to the control varieties are not necessary for the estimation of this contrast. Because $\mathbf{1}'_v\mathbf{Q}=0$ is the only relationship among the elements of $\mathbf{Q}$, this is only possible if the first c elements of the $\mathbf{b}'\mathbf{C}^-$ vector are equal to each other.

Consider a subtrial with two control varieties. Only designs for which the first two elements of $\mathbf{b}'\mathbf{C}^-$ are equal result in the equality of the local and the best estimate of $\mathbf{b}'\boldsymbol{\tau}$. Let $\mathbf{e}_l$ be a null vector with a one in the l^{th} position. Then $\mathbf{b}'\mathbf{C}^-\mathbf{e}_1=\mathbf{b}'\mathbf{C}^-\mathbf{e}_2$ must be true. Let new variety i $(i>2)$ be the variety of interest, so $\mathbf{b}'=1/2\,(2\mathbf{e}_i-\mathbf{e}_1-\mathbf{e}_2)$. Now

$$2\mathbf{b}'\mathbf{C}^-(\mathbf{e}_1-\mathbf{e}_2)=0\,,$$

$$(\mathbf{e}_i-\mathbf{e}_1+\mathbf{e}_i-\mathbf{e}_2)'\mathbf{C}^-(\mathbf{e}_1-\mathbf{e}_i+\mathbf{e}_i-\mathbf{e}_2)=0\,,$$

$$(\mathbf{e}_i-\mathbf{e}_1)'\mathbf{C}^-(\mathbf{e}_1-\mathbf{e}_i)+(\mathbf{e}_i-\mathbf{e}_1)'\mathbf{C}^-(\mathbf{e}_i-\mathbf{e}_2)+(\mathbf{e}_i-\mathbf{e}_2)'\mathbf{C}^-(\mathbf{e}_1-\mathbf{e}_i)+(\mathbf{e}_i-\mathbf{e}_2)'\mathbf{C}^-(\mathbf{e}_i-\mathbf{e}_2)=0\,,$$

$$-(\mathbf{e}_i-\mathbf{e}_1)'\mathbf{C}^-(\mathbf{e}_i-\mathbf{e}_1)+(\mathbf{e}_i-\mathbf{e}_2)'\mathbf{C}^-(\mathbf{e}_i-\mathbf{e}_2)=0\,,$$

$$-\mathrm{var}(\hat{\tau}_i-\hat{\tau}_1)+\mathrm{var}(\hat{\tau}_i-\hat{\tau}_2)=0\,.$$

Here $\hat{\tau}_i$, $\hat{\tau}_1$ and $\hat{\tau}_2$ denote single variety parameters instead of vectors. Hence $\mathbf{b}'\mathbf{C}^-\mathbf{e}_1=\mathbf{b}'\mathbf{C}^-\mathbf{e}_2$ if and only if

$$\mathrm{var}(\hat{\tau}_i-\hat{\tau}_1)=\mathrm{var}(\hat{\tau}_i-\hat{\tau}_2)\,. \tag{3.26}$$

This is the case for randomised complete block designs and balanced incomplete block designs. For partially balanced incomplete block designs this is only the case if the new variety is both j^{th} associate with control variety 1 and j^{th} associate with control variety 2. For instance, if the subtrial has a partially balanced incomplete block design with 2 association classes 0 and 1, variety i has to appear never with a control variety in one block or once with control variety 1 and once with control variety 2 for the local estimate and the best estimate to be equal. Notice that the other subtrials may have any other design.

In case of three control varieties the local estimate is equal to the best estimate if and only if the first three elements of $\mathbf{b'C^-}$ are equal, so

$$\mathbf{b'C^-e_1} = \mathbf{b'C^-e_2}\,, \quad \mathbf{b'C^-e_1} = \mathbf{b'C^-e_3} \quad \text{and} \quad \mathbf{b'C^-e_2} = \mathbf{b'C^-e_3}\,.$$

If we elaborate the above equations (not given here), we arrive at the condition that has to be met if the first three elements of $\mathbf{b'C^-}$ are to be identical :

$$\begin{aligned} &3\,\mathrm{var}(\hat{\tau}_i - \hat{\tau}_1) + \mathrm{var}(\hat{\tau}_2 - \hat{\tau}_3) \ = \\ &3\,\mathrm{var}(\hat{\tau}_i - \hat{\tau}_2) + \mathrm{var}(\hat{\tau}_1 - \hat{\tau}_3) \ = \\ &3\,\mathrm{var}(\hat{\tau}_i - \hat{\tau}_3) + \mathrm{var}(\hat{\tau}_1 - \hat{\tau}_2)\,. \quad (i > 3) \end{aligned} \tag{3.27}$$

For subtrials with designs that are variance-balanced, this condition is always met. Hence for subtrials with a randomised complete block design or a balanced incomplete block design the local estimate of $\mathbf{b'}\tau$ is identical to the best estimate. In a partially balanced incomplete block design with h association classes, there are h different variances of pairwise contrast estimators. If there are two association classes (for instance, a square lattice design), condition (3.27) can only be fulfilled when $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_1) =$ $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_2) =$ $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_3)$ and $\mathrm{var}(\hat{\tau}_1 - \hat{\tau}_2) =$ $\mathrm{var}(\hat{\tau}_1 - \hat{\tau}_3) =$ $\mathrm{var}(\hat{\tau}_2 - \hat{\tau}_3)$. So the new variety has to be in the same association class with all three control varieties, and the control varieties mutually also have to be in the same association class. The first and the latter association class do not have to be the same. When there are more than two association classes, there theoretically are more possibilities to satisfy (3.27).

For a variance-balanced subtrial k it is known that the pseudo-inverse of the $\mathbf{C}$ matrix can be written as $\mathbf{C^-} = \theta^{-1}(\mathbf{I} - (t_k + c)^{-1}\mathbf{11'})$, with θ the unique non-zero eigenvalue of $\mathbf{C}$ (Dey, 1986). From this we see that for any number c of control varieties the first c elements of $\mathbf{b'C^-}$ are equal to each other. So if the variety is included in a subtrial with a variance-balanced design, the local estimate of $\mathbf{b'}\tau$ is identical to the best estimate, regardless of the number of control varieties.

3.3.4 Examples

We will discuss three examples. For each example the incidence scheme is given. Notation : CV1 = control variety 1, NV1 = new variety 1, B1 = block 1, ST1 = subtrial 1, and so on. For the pseudo-inverses we take Moore-Penrose inverses.

Example 1.

	ST1			ST2	
	B1	B2	B3	B4	B5
CV1	1	0	0	1	1
CV2	0	1	0	1	1
NV1	1	0	1		
NV2	0	1	1		
NV3				1	1
NV4				1	1

$$\mathbf{C}_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}, \quad \mathbf{C}_2 = \frac{1}{2}\begin{pmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{pmatrix}.$$

Suppose we want to estimate the contrast between the parameter of new variety NV1 and the average of the parameters of the two control varieties CV1 and CV2. Then this contrast is first estimated at subtrial 1, because NV1 is included in that subtrial. The contrast is estimated with use of the reduced normal equations :

$$(-1/2 \quad -1/2 \quad 1 \quad 0)(\tilde{\alpha}_1' \quad \tilde{\tau}_1')' = (-1/2 \quad -1/2 \quad 1 \quad 0)\mathbf{C}_1^- \mathbf{Q}_1 .$$

This estimate can be improved by subtracting from this local estimate :

$$\left(\mathbf{p}' - \mathbf{q}'\mathbf{C}_{22_1}^{-1}\mathbf{C}_{21_1}\right)(\tilde{\alpha}_1 - \hat{\alpha}) ,$$

where $\hat{\alpha} = \sum_{k=1}^{m} \mathbf{W}_k \tilde{\alpha}_k$,

with $\mathbf{W}_k = \left[\sum_{k=1}^{m}\left(\mathbf{C}_{11_k} - \mathbf{C}_{12_k}\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right)\right]^- \left(\mathbf{C}_{11_k} - \mathbf{C}_{12_k}\mathbf{C}_{22_k}^{-1}\mathbf{C}_{21_k}\right).$

$$\mathbf{C}_{11_1} - \mathbf{C}_{12_1}\mathbf{C}_{22_1}^{-1}\mathbf{C}_{21_1} = \frac{1}{6}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{W}_1 = \frac{1}{14}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

$$\mathbf{C}_{11_2} - \mathbf{C}_{12_2}\mathbf{C}_{22_2}^{-1}\mathbf{C}_{21_2} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \mathbf{W}_2 = \frac{1}{14}\begin{pmatrix} 6 & -6 \\ -6 & 6 \end{pmatrix}.$$

$$\left(\mathbf{p}' - \mathbf{q}'\mathbf{C}_{22_1}^{-1}\mathbf{C}_{21_1}\right)(\tilde{\alpha}_1 - \hat{\alpha}) = (1/7 \quad -1/7)\tilde{\alpha}_1 - (1/7 \quad -1/7)\tilde{\alpha}_2 .$$

Notice that in subtrial 2 only a contrast between the control varieties has to be estimated, and that the improvement is a contrast of two contrasts between control variety parameters.

Example 2.

	ST1			ST2			ST3	
	B1	B2	B3	B4	B5	B6	B7	B8
CV1	1	0	1	0	1	0	1	0
CV2	0	1	0	1	0	1	1	1
CV3	1	0	1	0	1	0	0	1
NV1	1	1	1					
NV2	1	1	1					
NV3				1	1	1		
NV4				1	1	1		
NV5							1	1
NV6							0	1

$$\mathbf{C}_1=\frac{1}{6}\begin{pmatrix} 9 & 0 & -3 & -3 & -3 \\ 0 & 4 & 0 & -2 & -2 \\ -3 & 0 & 9 & -3 & -3 \\ -3 & -2 & -3 & 13 & -5 \\ -3 & -2 & -3 & -5 & 13 \end{pmatrix},\quad \mathbf{C}_2=\frac{1}{12}\begin{pmatrix} 9 & 0 & -3 & -3 & -3 \\ 0 & 16 & 0 & -8 & -8 \\ -3 & 0 & 9 & -3 & -3 \\ -3 & -8 & -3 & 25 & -11 \\ -3 & -8 & -3 & -11 & 25 \end{pmatrix},$$

$$\mathbf{C}_3=\frac{1}{12}\begin{pmatrix} 8 & -4 & 0 & -4 & 0 \\ -4 & 17 & -3 & -7 & -3 \\ 0 & -3 & 9 & -3 & -3 \\ -4 & -7 & -3 & 17 & -3 \\ 0 & -3 & -3 & -3 & 9 \end{pmatrix}.$$

Suppose we want to estimate the contrast between the parameter of new variety NV1 and the average of the parameters of the three control varieties CV1, CV2 and CV3. First, this contrast is locally estimated in subtrial 1. Next this local estimate can be improved.

$$\mathbf{C}_{11_1}-\mathbf{C}_{12_1}\mathbf{C}_{22_1}^{-1}\mathbf{C}_{21_1}=\frac{1}{8}\begin{pmatrix} 9 & -2 & -7 \\ -2 & 4 & -2 \\ -7 & -2 & 9 \end{pmatrix},\quad \mathbf{W}_1=\frac{1}{1914}\begin{pmatrix} 599 & -154 & -445 \\ -154 & 308 & -154 \\ -445 & -154 & 599 \end{pmatrix},$$

$$\mathbf{C}_{11_2}-\mathbf{C}_{12_2}\mathbf{C}_{22_2}^{-1}\mathbf{C}_{21_2}=\frac{1}{14}\begin{pmatrix} 9 & -4 & -5 \\ -4 & 8 & -4 \\ -5 & -4 & 9 \end{pmatrix},\quad \mathbf{W}_2=\frac{1}{1914}\begin{pmatrix} 349 & -176 & -173 \\ -176 & 352 & -176 \\ -173 & -176 & 349 \end{pmatrix},$$

$$\mathbf{C}_{11_3}-\mathbf{C}_{12_3}\mathbf{C}_{22_3}^{-1}\mathbf{C}_{21_3}=\frac{1}{12}\begin{pmatrix} 7 & -6 & -1 \\ -6 & 12 & -6 \\ -1 & -6 & 7 \end{pmatrix},\quad \mathbf{W}_3=\frac{1}{1914}\begin{pmatrix} 328 & -308 & -20 \\ -308 & 616 & -308 \\ -20 & -308 & 328 \end{pmatrix}.$$

$$\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_1}^{-1}\mathbf{C}_{21_1}\right)(\tilde{\alpha}_1-\hat{\alpha})=0.0316(1\quad -2\quad 1)\tilde{\alpha}_1-0.0115(1\quad -2\quad 1)\tilde{\alpha}_2-0.0201(1\quad -2\quad 1)\tilde{\alpha}_3 .$$

Again the improvement represents a contrast of contrasts between control variety parameters estimated at the three subtrials. The contrasts at the various subtrials do not have to be of the same type, which is shown by the improvement of the local estimator of the contrast between the parameter of new variety NV6 and the average of the parameters of the control varieties :

$$\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_3}^{-1}\mathbf{C}_{21_3}\right)(\tilde{\alpha}_1-\hat{\alpha})=(-0.1389\quad 0.1264\quad 0.0125)\tilde{\alpha}_1-(-0.0684\quad 0.0459\quad 0.0225)\tilde{\alpha}_2-$$
$$(-0.0705\quad 0.0804\quad -0.0099)\tilde{\alpha}_3 .$$

Example 3.

This is the example introduced in *3.3.1*. Consider the following observations:

ST1

	B1	B2
CV1	13.20	14.11
CV2	15.68	10.53
NV1	11.00	11.31
NV2	11.96	12.58

ST2

	B3	B4	B5	B6	B7	B8
CV1	10.66		12.85		15.83	
CV2	10.85			14.21		13.10
NV3		11.41	13.58			13.63
NV4		11.91		15.56	16.41	

ST3

	B9	B10	B11	B12	B13	B14
CV1	14.51			14.66		
CV2	15.38				15.50	
NV5	16.10					14.40
NV6		16.73		14.46		
NV7		10.58			14.28	
NV8		11.28				14.04
NV9			11.91	15.66		
NV10			12.65		13.93	
NV11			14.06			14.40

With this small example we are able to calculate the outcome of the BLUE of $\mathbf{b}'\tau$ in one step, using the fixed additive model for the joint observations of the three subtrials. This can e.g. be done with the SAS package. We will give these outcomes, together with the outcomes of the local estimators and the difference between the local estimator and the BLUE.

variety	best estimate	local estimate	difference
NV1	-2.23	-2.23	0
NV2	-1.11	-1.11	0
NV3	0.59	0.59	0
NV4	1.01	1.01	0
NV5	0.57	0.57	0
NV6	1.86	1.81	-0.05
NV7	-1.99	-1.94	0.05
NV8	-1.09	-1.09	0
NV9	0.25	0.20	-0.05
NV10	-0.93	-0.88	0.05
NV11	0.68	0.68	0

Varieties NV1 and NV2 are present in a subtrial with a randomised complete block design, so the local estimate of $\mathbf{b}'\tau$ is also the best estimate. This is also true for varieties NV3 and NV4, because they are present in a subtrial with a balanced incomplete block design. Varieties NV5 up to and including NV11 are present in a subtrial with a partially balanced incomplete block design. Since varieties NV5, NV8 and NV11 are in the same association class with both control varieties (they occur once with control 1 and once with control 2 in the same block or they occur never with control 1 and never with control 2 in the same block), we know that for these varieties the local estimate of $\mathbf{b}'\tau$ is equal to the best estimate. For the remaining varieties the local estimate is not equal to the best estimate. The difference between these two estimates can be calculated from the separate analyses of the subtrials, as described in this section. The $\mathbf{W}_k$ matrices are :

$$\mathbf{W}_1=\mathbf{W}_2=\frac{2}{11}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix},\quad \mathbf{W}_3=\frac{3}{22}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}.$$

For variety NV6, for example, the difference can be calculated as

$$\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_3}^{-1}\mathbf{C}_{21_3}\right)(\tilde{\alpha}_3-\hat{\alpha})=\left(\mathbf{p}'-\mathbf{q}'\mathbf{C}_{22_3}^{-1}\mathbf{C}_{21_3}\right)([\mathbf{I}_c-\mathbf{W}_3]\tilde{\alpha}_3-\mathbf{W}_1\tilde{\alpha}_1-\mathbf{W}_2\tilde{\alpha}_2)$$

$$=(1/11 \quad -1/11)\tilde{\alpha}_3-(1/22 \quad -1/22)\tilde{\alpha}_1-(1/22 \quad -1/22)\tilde{\alpha}_2$$

$$=-0.05\,.$$

This difference can be used to improve the local estimate to the BLUE.

3.3.5 Using a mixed model

In the previous sections all the block terms in the model were considered fixed. Now consider the situation where the subtrials have resolvable designs, for instance lattice designs. Then one could introduce fixed terms for the replications, but further assume that the blocks within each replication form a random sample.

The corresponding model is (3.7). If we use this model, similar results as in *3.3.2* are obtained. Let $\mathbf{M}_k$ be the design matrix corresponding to the replications at subtrial k and let ρ_k denote the vector of replication parameters at subtrial k. With the current mixed model the variance/covariance matrix of the observations at subtrial k is equal to $\mathbf{D}[\mathbf{Y}_k] = (\mathbf{Z}_k\mathbf{Z'}_k\sigma_B^2/\sigma^2 + \mathbf{I})\sigma^2 = \mathbf{V}_k\sigma^2$. Using the well known Aitken transformation, the formulae given in *3.3.2* to calculate solutions of the parameter estimators and pseudo-variance/covariance matrix can be made valid for the current model. Let matrix $\mathbf{U}_k$ be defined by $\mathbf{U}_k\mathbf{U}_k = \mathbf{V}_k^{-1}$. This is possible because $\mathbf{V}_k$ is a positive definite matrix. Then in the given formulae $\mathbf{G}_k$ has to be replaced by $\mathbf{U}_k\mathbf{G}_k$, $\mathbf{L}_k$ has to be replaced by $\mathbf{U}_k\mathbf{L}_k$, $\mathbf{Z}_k$ has to be replaced by $\mathbf{U}_k\mathbf{M}_k$, β_k has to be replaced by ρ_k and $\mathbf{Y}_k$ has to be replaced by $\mathbf{U}_k\mathbf{Y}_k$. So also with a model with fixed replication terms and random blocks within replication terms, the best estimator of the desired contrasts between variety parameters can be determined in steps.

REFERENCES

Baksalary, J.K. & R. Kala (1983). On equalities between BLUEs, WLSEs and SLSEs. *The Canadian Journal of Statistics* 11, 119-123; Correction (1984). 12, 240.

Bartlett, M.S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Journal of the Royal Statistical Society* Suppl. 4, 158-159.

Cochran, W.G. & G.M. Cox (1957). *Experimental designs*. 2nd edition. Wiley & Sons, New York.

Corsten, L.C.A. (1967). Variantie-analyse, *Landbouwkundig Tijdschrift*. 79(5), 159-164.

Corsten, L.C.A. (1976). Canonical Correlation in Incomplete Blocks, In : *Essays in probability and statistics* p. 125-154. S. Ikeda *et al.* (eds.). Shinko Tsusho Co. Ltd. 1-7-1 Wakaba, Shinjukuku; Tokyo 160, Japan.

Dey, A. (1986). *Theory of block designs*. Wiley & Sons, New York.

Dyke, G.V. (1988). *Comparative experiments with field crops*. Butterworths, London.

John, P.W.M. (1971). *Statistical design and analysis of experiments*. The Macmillan Company, New York.

Kuiper, N.H. (1952). Variantie analyse. *Statistica Neerlandica*. 6(3), 149-194.

Kuiper, N.H. (1983). Analysis of variance. *Mededelingen Landbouwhogeschool Wageningen*. 83-10. (English translation of Kuiper (1952))

Levene, H. (1960). Robust tests for equality of variances. In : I. Olkin *et al.* (Eds.) *Contributions to probability and statistics*. Stanford : University Press, Stanford.

Mead, R. (1988). *The design of experiments, statistical principles for practical application*. Cambridge University Press, Cambridge.

Patterson, H.D. & V. Silvey. (1980). Statutory and recommended list trials of crop varieties in the United Kingdom. *Journal of the Royal Statistical Society* A 143, 219-252.

Rao, C.R. (1973). *Linear statistical inference and its applications*. 2nd. edition. Wiley & Sons, New York.

Silvey, V. (1978). Methods of analysing NIAB variety trial data over many sites and several seasons. *Journal of the National Institute for Agricultural Botany* 14, 385-400.

Verdooren, L.R. (1988). *Statistical inference on variance components*. PhD. thesis of the Agricultural University Wageningen.

Verdooren, L.R. (1989). Use of variance components in the analysis of incomplete block designs. *Biuletyn Oceny Odmian (Cultivar Testing Bulletin)* no. 21-22, 169-186.

Yates, F. (1933). The principles of orthogonality and confounding in replicated experiments. *Journal of Agricultural Science* 23, 108-145.

Yates, F. (1940). The recovery of interblock information in balanced incomplete block designs. *Annals of Eugenics* 10, 317-325.

Yates, F. & W.G. Cochran. (1938). The analysis of groups of experiments. *Journal of Agricultural Science* 28, 556-580.

CHAPTER 4

Selection of varieties

The main reason why variety trials are performed is to evaluate the varieties and to make a selection. For the time being we assume that there is a single quantitative character on the basis of which the selection is made. We assume that varieties with a high variety value are desired. Thence, after the analysis of the experiment the plant breeder selects varieties which have a high estimated value. Although the variety value is estimable in an experiment with a connected design, it is often more easy to make the selection on the basis of contrasts between variety parameters. Except for the case of a model including fixed interaction terms, this is possible because contrasts between variety parameters are equal to the corresponding contrasts between variety values. To rank the varieties we could estimate the pairwise contrasts between the variety parameter of variety i $(i = 1, \dots, t-1)$ and variety t. These estimates are equal to the solution of the reduced normal equations under the restriction that the parameter corresponding to variety t is equal to zero. However, contrasts between variety parameters can only be *estimated*, hence we are never absolutely sure what the true values are. This means that the variety with the largest estimated variety value is not necessarily also the variety with the largest true variety value. Plant breeders have perceived this and therefore select more than one variety, which are then often tested further in other years and/or at other sites. However, the number of selected varieties is often not determined on the basis of statistical motives, but by the prespecified selection programme itself.

Statistics richly provides the plant breeder with theory to analyse field trials. It gives techniques to determine the best estimators of contrasts between variety parameters and the distribution of these estimators. However, the classical approach to test the null-hypothesis that all variety parameters are equal (the so-called homogeneity test) is of no use in the plant breeding practice. The breeder takes the line that the varieties differ from each other, and if the null-hypothesis cannot be rejected this only indicates that the experimental error was too large for the differences to be detected. The multiple comparisons techniques give more information than the homogeneity test, but they are not designed for selection purposes. The plant breeder is not really interested in all possible comparisons, but more specifically in comparisons between the variety of interest and the other varieties. Statistical selection procedures are restricted to this kind of comparisons and are therefore better suited for selection problems. Statistical selection procedures are developed to aid the experimenter in making selection decisions. The procedures advice the plant breeder which varieties to select, when a certain criterium has to be satisfied. These selection procedures should not be used in order to replace the subjective opinion of the plant breeder, but as a supplement to this opinion.

The most important methods for plant breeding are described in section **4.1**. However, many selection procedures were not developed for large field trials with (unbalanced) incomplete block designs. Because the experiments in the plant breeding practice are often of this type, some selection procedures are not directly applicable in practice. In section **4.2** new selection procedures are described of which some are highly convenient in practical use. A selection procedure consists of a selection rule. This rule contains parameters, called selection constants, which have to be calculated in advance. When incomplete block designs with many varieties are used, the calculation of these selection constants may become very troublesome. Therefore we have to use simulation methods to approximate the selection constants. This is described in section **4.3**, together with the approximation of other important statistics by simulation. With computers becoming faster and faster, computer simulation becomes an attractive method to determine the selection constants in practice. With workable selection rules available and the selection constants approximated by simulation, statistical selection procedures can now be used in the plant breeding practice. This is further described in **4.4**. Finally, section **4.5** deals with modifications of the selection procedures that are made in order to make statistical selection procedures even more useful for the plant breeder.

4.1 Statistical selection procedures

We can distinguish two main streams in selection, called Model I selection and Model II selection. The difference lies in the assumptions about the variety terms in the models. In Model I selection we consider the variety terms fixed. The breeder is interested in these varieties included in the experiment. The variety values or contrasts between variety parameters are estimated using one of the models described in chapter 3. In these models the variety parameters, denoted by τ_i $(i = 1, \ldots, t)$, are fixed terms. After the estimation of contrasts between variety parameters, a limited number of the tested varieties is selected. To do this in a statistical sensible way, statistical selection procedures can be used.

In Model II selection the variety terms are assumed to be Normally distributed, independent random variables with common expectation μ, say, and common variance σ_T^2. The varieties included in the experiment are considered a random sample from an infinite Normal population with mean μ and standard deviation σ_T. Now the breeder is not interested in these sampled varieties themselves, but in the population which they represent. The selection aim is to increase the value of the population mean μ by means of selection and recombination. The top-ranking varieties are selected and allowed to mate at random. Then the offspring of the parents form a random sample from a new population, with expectation $\mu' > \mu$. The difference between μ' and μ is called 'Response to Selection' and depends on the selection percentage and the heritability in narrow sense, which is the proportion of the total variance that is attributable to the selectable effects of genes. However, since this situation is totally different from the one studied in this thesis, we will not pay attention to it. The interested reader is referred to Falconer (1986) or Bulmer (1985).

Nevertheless Model II selection is mentioned because of a different selection theory that we categorize as Model II selection, although we also could name it Model I selection. In this approach the interest of the plant breeder lies in the varieties actually used in the trial. So this corresponds to Model I selection. However, now we assume that the variety terms are a random sample from an infinite Normal population with mean μ and standard deviation σ_T, as we do in Model II selection. The mathematical theory used strongly resembles that of Model II selection, and so we classify this approach as such. It must be emphasized that no mating takes place in this situation, although we call it Model II selection.

4.1.1 Model I selection

The literature about statistical selection procedures is very extensive. The reader is referred to Gibbons, Olkin & Sobel (1977) [applied] and Gupta & Panchapakesan (1979) [theoretical]. For an overview of these selection procedures see Van Der Laan & Verdooren (1989). We will concentrate on some procedures which may be of interest in the plant breeding practice. In doing so we will use notation that corresponds to the notation used in chapter **3**, and present the selection procedures in the context of variety selection. The expensive and laborious creation of new varieties demands a thoroughly designed selection procedure in the selection phase of the breeding process. Inferior varieties should be discarded as soon as possible but on the other hand superior varieties should be retained. It is very important that the plant breeder knows how well he is selecting. Unfortunately this is often not the case since many breeders simply select a prespecified number (or percentage) of varieties that gave the best results in the experiment. This procedure has led to the following phrase often heard from plant breeders : 'The best variety I've made I've probably thrown away'. The Model I statistical selection procedures connect a statistical inference to the selection procedure. The two basic approaches of statistical selection are the Indifference Zone approach (Bechhofer, 1954) and the Subset approach (Gupta, 1956, 1965). These two approaches were combined in the Multiple Comparisons with the Best approach (Hsu, 1984).

In Model I selection procedures the variety terms are taken fixed. The theory was first developed for experiments with equi-replicated completely randomised designs. The model for the observations from such an experiment can be written as

$$Y_{ij} = \lambda + \tau_i + E_{ij}\,, \quad i = 1, \ldots, t\ ;\ j = 1, \ldots, n\ .$$

Here the number of plots (n) with variety i is equal for all varieties. The value of variety i is now defined as $\lambda + \tau_i$. Let the true variety values corresponding to the t varieties be ranked (in notation) from the smallest to the largest :
$\lambda + \tau_{(1)} \le \lambda + \tau_{(2)} \le \ldots \le \lambda + \tau_{(t)}$.

With the assumption that the E_{ij} are uncorrelated random variables with expectation zero and common variance σ^2, the variety value estimates are the solution of the normal equations. In an experiment with a completely randomised design the best linear unbiased estimator (BLUE) of the value of variety i is the unweighted average of the observations corresponding to this variety, denoted by $\overline{Y}_i$. These estimated variety values can now be ranked as
$\overline{Y}_{[1]} \le \overline{Y}_{[2]} \le \ldots \le \overline{Y}_{[t]}$.

If there are ties between estimated variety values the varieties concerned have to be 'ranked' at random. The estimated variety value corresponding to the variety with $\lambda + \tau_{(i)}$ is denoted by $\overline{Y}_{(i)}$. Suppose a breeder is interested in selecting the variety with the largest true variety value $\lambda + \tau_{(t)}$, called the best variety.

If only one variety is selected, it is logical to choose the variety with the largest estimated variety value, $\overline{Y}_{[t]}$. This selection is called correct if the selected variety is indeed the best variety. Then a quantitative measure for the correctness of the selection is the probability of correct selection, in the sequel denoted by $P(CS)$. This probability is equal to

$$
\begin{aligned}
P(CS) &= P(\overline{Y}_{[t]} = \overline{Y}_{(t)}) \\
&= P(\overline{Y}_{(i)} < \overline{Y}_{(t)}\,,\ i = 1, \ldots, t-1) \\
&= P\left(\frac{\overline{Y}_{(i)} - \lambda - \tau_{(i)}}{\sigma/\sqrt{n}} < \frac{\overline{Y}_{(t)} - \lambda - \tau_{(t)}}{\sigma/\sqrt{n}} + \frac{\tau_{(t)} - \tau_{(i)}}{\sigma/\sqrt{n}}\,,\ i = 1, \ldots, t-1 \right).
\end{aligned}
$$

Because we do not know the differences $\tau_{(t)} - \tau_{(i)}$, it is impossible to calculate the probability of correct selection. However, we can give the infimum of this probability, but then we have to make some extra assumptions. We now assume that the errors E_{ij} are independent random variables with a Normal distribution. Furthermore it is first assumed that the variance is known. The estimators $\overline{Y}_i$ are independent, hence the $P(CS)$ can now be calculated as

$$
P(CS) = \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{t-1} \Phi\left(X + \frac{\tau_{(t)} - \tau_{(i)}}{\sigma/\sqrt{n}} \right) \right\} \phi(X)\ dX\ , \tag{4.1}
$$

where $\Phi(.)$ is the Standard Normal cumulative distribution function and $\phi(.)$ is the Standard Normal probability density function. The $P(CS)$ lies between $1/t$ and 1. The value $1/t$ is attained if all $\tau_{(t)} - \tau_{(i)}$ are set to zero in (4.1). This is the probability of correct selection when a variety is randomly selected from the total number t of varieties.

A more useful minimum $P(CS)$ gives the Indifference Zone approach of Bechhofer (1954). The breeder has to specify a distance measure δ^*, which indicates the minimum difference between the true value of the best variety and the true value of the second best variety that he finds important. Now the parameter space is divided into an Indifference Zone, with configurations of parameters where $\tau_{(t)} - \tau_{(t-1)} < \delta^*$, and a Preference Zone, with configurations where $\tau_{(t)} - \tau_{(t-1)} \geq \delta^*$.

The plant breeder is only interested in making a correct selection if the actual configuration of parameters is included in the Preference Zone. Bechhofer (1954) showed that the infimum of the $P(CS)$, with the configuration of parameters included in the Preference Zone, is reached when all differences $\tau_{(t)} - \tau_{(i)}$ $(i \neq t)$ are equal to δ^*. This configuration of the parameters is called the Least Favourable Configuration (LFC). Then

$$P(CS) \geq \int_{-\infty}^{\infty} \Phi^{t-1}\left(X + \frac{\delta^* \sqrt{n}}{\sigma} \right) \phi(X) \; dX$$

$$= \int_{-\infty}^{\infty} \Phi^{t-1}(X + \gamma) \phi(X) \; dX \text{ , with } \gamma = \frac{\delta^* \sqrt{n}}{\sigma} .$$

This minimum value of the probability of correct selection over all configurations in the Preference Zone is denoted by $P_{LFC}(CS)$.

The aim of the Indifference Zone approach is to calculate the minimal number of observations for each variety, necessary to make the statement that the probability that the selected variety really is the best one is at least P^*, if the difference between the best variety parameter and the second best variety parameter is larger than δ^*. Thus this approach looks at selection from the design point of view. The breeder has to specify the desired P^* and the critical difference δ^*. Then the value of γ is determined as the solution of $P_{LFC}(CS) = P^*$. This can easily be done by numerical integration. The value of γ is tabulated for various values of t and P^*, e.g. by Gibbons, Olkin & Sobel (1977) and most extensively by Butler & Butler (1987). The latter have tabulated the values of γ for t = 1(1)400, 410 (10) 2000 and $P^* = 0.50$, 0.80, 0.90, 0.95, 0.975, 0.99, 0.995, 0.999. The common number of observations per variety must then be chosen as (at least) the smallest integer satisfying :

$$n \geq \left(\frac{\gamma\sigma}{\delta^*} \right)^2 .$$

When the requirements are met, we can state with probability of at least P^* that the variety parameter of the selected variety lies between $\tau_{(t)} - \delta^*$ and $\tau_{(t)}$. However, with a large number t of varieties to be tested the required number of observations is often too large. The number of observations in practice is often limited by shortage of seed, experimental fields, money, and so on. Therefore the Indifference Zone approach can only be useful in a selection phase where only a small number of varieties is tested.

Consider the following example. Assume there are $t = 100$ varieties, which we want to test in an experiment with a completely randomised design. With $P^* = 0.90$, γ can be found in Butler & Butler (1987) as $\gamma = 3.902021$. If we now choose $\delta^* = 1\ \sigma$, then $n \geq (3.902021)^2 = 15.23$, so $n = 16$. This means that we have $n.t = 1600$ plots in a completely randomised design, which is not feasible. If we have only $t = 4$ varieties, $\gamma = 2.451569$. If we choose $\delta^* = 1\ \sigma$, this would lead to a minimum number of observations $n = 7$. With only 4 varieties, 7 observations per variety, so a total number of 28 plots, is feasible.

In the above theory the assumption was first made that the variance is known. In the plant breeding practice the variance can only be considered known if the number of degrees of freedom for error is very large. If the variance is unknown, it is not possible to determine the required number of observations in one step, except when the distance δ^* is given as a multiple of σ. For the situation of unknown variance two-stage procedures have been proposed (Bechhofer, Dunnett & Sobel 1954; Dunnett & Sobel, 1954). However, two-stage procedures are very inconvenient to use in the plant breeding practice, because the selection procedure may not take too much time.

The Indifference Zone approach can also be used when the aim is to select the k ($k \geq 2$) best varieties. Then δ^* corresponds with the difference between $\tau_{(t-k+1)}$ and $\tau_{(t-k)}$. The formula of the probability of correct selection differs from the one given previously but is not stated here.

With the Indifference Zone approach a fixed number of varieties is selected. A different approach was suggested by Gupta (1956, 1965). Using his approach the breeder selects a non-empty subset of random size. Which varieties have to be included in the subset is prescribed by a selection rule. This selection rule is designed in such a way that the probability that the variety of interest is included in the subset is at least P^*. Because the number of selected varieties is not determined beforehand but is random, it is always possible to find a subset that satisfies the P^*-requirement. However, the subset size has to be as small as possible. Unlike the Indifference Zone approach, it is not necessary to specify a critical distance δ^*. The Subset approach looks at selection from the analysis point of view. Which varieties are of interest has to be indicated by the plant breeder, by specifying the aim of selection. Does he want to select a subset that includes the best variety, or a subset that includes at least one good variety (where 'good' has to be further specified), or a subset that comprises only good varieties, or perhaps a subset that includes all varieties better than a control variety ? The aim of selection indicates which type of selection rule we should use. Selection of the best variety enjoys most of the attention in literature till now.

Suppose the interest of the breeder lies in the smallest subset that includes the best variety with a certain confidence. The theory was first developed for an experiment with a completely randomised design with n observations per variety. We assume that the observations are independently Normally distributed with common variance σ^2. First assume that σ^2 is known. Gupta (1956, 1965) proposed the following selection rule :

Select variety i $(i = 1, \ldots, t)$ if and only if

$$\overline{Y}_i \geq \overline{Y}_{[t]} - \frac{\gamma}{\sqrt{n}}\sigma\,, \tag{4.2}$$

with γ the so-called selection constant. Thus on the basis of the experimental results the subset is determined. The probability of correct selection can be written as

$$\begin{aligned} P(CS) &= P\left(\overline{Y}_{(t)} \geq \overline{Y}_{[t]} - \frac{\gamma}{\sqrt{n}}\sigma\right) \\ &= P\left(\overline{Y}_{(t)} \geq \overline{Y}_{(i)} - \frac{\gamma}{\sqrt{n}}\sigma\,,\ i = 1, \ldots, t-1\right) \\ &= P\left(\frac{\overline{Y}_{(i)} - \lambda - \tau_{(i)}}{\sigma/\sqrt{n}} \leq \frac{\overline{Y}_{(t)} - \lambda - \tau_{(t)}}{\sigma/\sqrt{n}} + \frac{\tau_{(t)} - \tau_{(i)}}{\sigma/\sqrt{n}} + \gamma\,,\ i = 1, \ldots, t-1\right) \\ &= \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{t-1} \Phi\left(X + \frac{\tau_{(t)} - \tau_{(i)}}{\sigma/\sqrt{n}} + \gamma\right)\right\} \phi(X)\, dX\,. \end{aligned} \tag{4.3}$$

Gupta showed that, since $\tau_{(t)} - \tau_{(i)} \geq 0$, the infimum of this probability is reached when all pairwise contrasts $\tau_{(t)} - \tau_{(i)}$ are taken equal to zero. The configuration where all τ_i are equal to each other is the Least Favourable Configuration for subset selection. Hence an absolute minimum of the probability of correct selection over all configurations possible is equal to

$$P(CS) \geq \int_{-\infty}^{\infty} \Phi^{t-1}(X+\gamma)\phi(X)\,dX$$

$$= P_{LFC}(CS)\,.$$

If the breeder wants P^* to be e.g. 0.90, then γ can be found by solving $P_{LFC}(CS) = P^*$. These selection constants can e.g. be found in Butler & Butler (1987).

For the equi-replicated completely randomised design the expected subset size can easily be calculated. Consider the random variable Z_i, with $Z_i = 1$ if variety i is selected and $Z_i = 0$ if variety i is not selected. Then the expected subset size, denoted by $E(|S|)$, can be written as

$$E(|S|) = \sum_{i=1}^{t} E(Z_i)$$

$$= \sum_{i=1}^{t} P(\text{variety } i \text{ selected})$$

$$= \sum_{i=1}^{t} P\left(\overline{Y}_i \geq \overline{Y}_{[t]} - \frac{\gamma}{\sqrt{n}}\sigma\right)$$

$$= \sum_{i=1}^{t} P\left(\overline{Y}_i \geq \overline{Y}_j - \frac{\gamma}{\sqrt{n}}\sigma\,,\ \forall j \neq i\right)$$

$$= \sum_{i=1}^{t} \int_{-\infty}^{\infty} \left\{ \prod_{\substack{j=1 \\ j \neq i}}^{t} \Phi\left(X + \frac{\tau_i - \tau_j}{\sigma/\sqrt{n}} + \gamma\right)\right\} \phi(X)\,dX\,. \tag{4.4}$$

Notice that we only can calculate the expected subset size for a given configuration of the τ_i. Gupta (1965) proved that the maximum subset size over all possible configurations is reached if we calculate $E(|S|)$ for the Least Favourable Configuration. So

$$E(|S|) \leq \sum_{i=1}^{t} \int_{-\infty}^{\infty} \Phi^{t-1}(X+\gamma)\phi(X)\,dX$$

$$= \sum_{i=1}^{t} P_{LFC}(CS) = tP_{LFC}(CS)\,.$$

If we cannot assume that σ^2 is known, we replace σ in the selection rule by s, the root of the mean squared error. For the equi-replicated completely randomised design, the number of degrees of freedom for error is equal to $t(n-1)$. Then the infimum of the probability of correct selection can be calculated as

$$P(CS) \geq \int_{0}^{\infty} \int_{-\infty}^{\infty} \Phi^{t-1}(X + \gamma Z)\phi(X) q_{\nu}(Z)\, dX dZ \ ,$$

with $q_\nu(.)$ the density of $\sqrt{\chi^2_\nu/\nu}$ with ν degrees of freedom, where χ^2_ν is a Chi-square distributed random variable independent of χ, the standard Normal distributed variable. It must be understood that although we denote the selection constant in both the situation of known variance and unknown variance by γ, the values of γ are different. Some γ values have been tabulated in table A.4 of Gibbons, Olkin & Sobel (1977). The table entries have to be multiplied by $\sqrt{2}$, because the authors explicitly include $\sqrt{2}$ in the selection rule. More γ values are tabulated in Bechhofer & Dunnett (1988).

We have assumed that the observations are independent, Normally distributed variables with common variance σ^2. The common variance assumption was also used for the estimation techniques described in chapter 3. However, Driessen, Van Der Laan & Van Putten (1988) showed, for known variance, that the robustness of the selection procedures against heterogeneity of variances is not good, for both the Indifference Zone procedure of Bechhofer and the Subset procedure of Gupta. If the variances are not equal, the $P_{LFC}(CS)$ can become smaller than the desired P^*, especially when many varieties are compared. Nevertheless, in practice we have strong reasons to believe that the real configuration of variety parameters is not equal to the Least Favourable Configuration. For configurations unequal to the Least Favourable Configuration, the probability of correct selection is predominantly determined by the distances between the variety parameters, and not by the variances. Therefore, in practical situations the robustness of the selection procedures will probably be good enough to allow unequal variances. This will probably also be the case if the variance is unknown. However, more research on this topic seems necessary. The robustness of the selection procedures against departures from the Normality assumption is satisfactory (Van Der Laan & Van Putten, 1988).

Unfortunately, the completely randomised design is seldomly used in the plant breeding practice. As mentioned in the previous chapters, often used designs are (in)complete block designs. Driessen (1991) extended the theory to this type of designs. He used the fixed additive model for the observations from an experiment with blocks, given in section *3.1.1* :

$$Y_{ijk} = \lambda + \tau_i + \beta_j + E_{ijk}\ , \quad i = 1,\ldots,t\ ;\ j = 1,\ldots,b\ ;\ k = 1,\ldots,n_{ij}\ .$$

In section *3.1.1* the analysis of an experiment with a connected block design was described. The variety parameters τ_i are not uniquely estimable, but contrasts between variety parameters are. We denote the variance of the BLUE of the difference between τ_i and τ_j by $\mathrm{var}(\hat{\tau}_i - \hat{\tau}_j) = v_{ij}^2\sigma^2$. Driessen (1991) proposed the following selection rule :

Select variety i ($i = 1,\ldots,t$) if and only if
$$\hat{\tau}_i \geq \hat{\tau}_j - \delta_i v_{ij} s\ ,\ \ \forall j \neq i\ . \tag{4.5}$$

Here δ_i is the selection constant for variety i, calculated for a particular P^* and the experimental design used. Notice that every variety has its own selection constant. Although $\hat{\tau}_i$ is not a unique value, it can still be used in the selection rule. This because only differences between variety parameters are important to make the selection. So it is not necessary to estimate and use the variety values for selection purposes. With the proposed selection rule the probability of correct selection can be written as

$$\begin{aligned} P(CS) &= P(\hat{\tau}_{(t)} \geq \hat{\tau}_{(j)} - \delta_{(t)} v_{(t)(j)} s\ ,\ \ \forall j \neq t) \\ &= P\left(\frac{\hat{\tau}_{(j)} - \hat{\tau}_{(t)} - (\tau_{(j)} - \tau_{(t)})}{v_{(t)(j)} s} \leq \frac{\tau_{(t)} - \tau_{(j)}}{v_{(t)(j)} s} + \delta_{(t)}\ ,\ \ \forall j \neq t \right). \end{aligned} \tag{4.6}$$

For the Least Favourable Configuration, i.e. $\tau_1 = \tau_2 = \ldots = \tau_t$, all $\tau_{(t)} - \tau_{(j)} = 0$ and the $P(CS)$ is minimal. Because it is not known which variety is the best, the separate selection constants δ_i are calculated as if variety i ($i = 1,\ldots,t$) is the best, hence $\delta_i = \delta_{(t)}$. Thus selection constant δ_i, corresponding to a specific minimum probability of correct selection P^*, is calculated by solving this equation :

$$P\left(\frac{\hat{\tau}_j - \hat{\tau}_i - (\tau_j - \tau_i)}{v_{ij} s} \leq \delta_i\ ,\ \ \forall j \neq i \right) = P^*$$

In case of variance-balanced designs, the separate selection constants δ_i are equal to each other. For the completely randomised design with a common number of observations n per variety $v_{ij} = \sqrt{2}/\sqrt{n}$, so $\delta_i = \gamma/\sqrt{2}$. Driessen (1992) proved that for an experiment with a partially balanced incomplete block design with two association classes, based on the group divisible association scheme or the triangular scheme or the L_2 Latin square scheme, the selection constants are also equal to each other.

In case the variance is known, s is replaced by σ in the above equations and selection rule (4.5). Of course the selection constants will differ for both situations.

Lam (1989) constructed upper as well as lower confidence bounds for $\tau_{(t)} - \tau_{(i)}$ $(i \neq t)$, for the situation of an equi-replicated completely randomised design. The lower confidence bounds are of special interest, because with these bounds it is possible to calculate a lower bound of the achieved probability of correct selection, using the Bechhofer procedure or the Gupta procedure. Consider the constant Δ^c, in the sequel denoted by 'confidence constant', which is the solution of the following equation :

$$\int_{-\infty}^{\infty} \{\Phi(X+\Delta^c) - \Phi(X-\Delta^c)\}^{t-1} \phi(X)\, dX = P^* .$$

Lam (1989) proved that in case of known variance σ^2, simultaneous $P^* \times 100\%$ confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ $(i \neq t)$, denoted by L_i, are

$$L_i = \max\left(0, \min_{j \in S} \overline{Y}_j - \overline{Y}_{[i]} - \frac{\Delta^c \sigma}{\sqrt{n}}\right),$$

with S the selected Gupta subset with $\gamma = \Delta^c$. Now a $P^* \times 100\%$ lower bound of the $P(CS)$, denoted by $P(CS)_L$, for the Indifference Zone approach is obtained by substituting the $\tau_{(t)} - \tau_{(i)}$ in (4.1) by the L_i. So with confidence $P^* \times 100\%$

$$P(CS)_L = \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{t-1} \Phi\left(X + \frac{L_i}{\sigma/\sqrt{n}}\right) \right\} \phi(X)\, dX . \tag{4.7}$$

For the Subset approach a $P^* \times 100\%$ confidence lower bound of the probability of correct selection is analogously obtained as

$$P(CS)_L = \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{t-1} \Phi\left(X + \frac{L_i}{\sigma/\sqrt{n}} + \gamma\right) \right\} \phi(X)\, dX . \tag{4.8}$$

Simultaneous confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ were also obtained for the situation of unknown variance. Now the confidence constant is the solution of

$$\int_0^{\infty}\int_{-\infty}^{\infty} \{\Phi(X+\Delta^c Z) - \Phi(X - \Delta^c Z)\}^{t-1} \phi(X) q_\nu(Z)\, dX dZ = P^*,$$

and σ is replaced by s to calculate L_i.

For the Subset approach, Driessen (1991) extended the theory of Lam to (in)complete block designs. Consider separate confidence constants Δ_i^c for each variety, for the situation of a common unknown variance. These confidence constants are defined by the following equations :

$$P(\hat{\tau}_j - \tau_j - \Delta_i^c v_{ij} s \le \hat{\tau}_i - \tau_i \le \hat{\tau}_j - \tau_j + \Delta_i^c v_{ij} s\ ,\ \forall j \ne i) = P^*.$$

Driessen (1991) proved that simultaneous $P^* \times 100\%$ confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ $(i \ne t)$ can be calculated as

$$L_i = \max\left(0, \min_{j\in S}\{\hat{\tau}_j - U^j_{[i]}\}\right),$$

with $U^j_{[i]}$ the i^{th} ordered statistic of $U_1^j, \ldots, U_t^j$, where $U_l^j = \hat{\tau}_l + \Delta_j^c v_{jl} s\ ,\ l = 1, \ldots, t$. S is the subset selected with selection rule (4.5) and $\delta_i = \Delta_i^c$. In case of variance-balanced designs, v_{ij} is a common value for all i, j $(i \ne j)$, say v, and $\Delta_1^c = \ldots = \Delta_t^c = \Delta^c$, so the calculation of L_i simplifies to

$$L_i = \max\left(0, \min_{j\in S} \hat{\tau}_j - \hat{\tau}_{[i]} - \Delta^c v\ s\right).$$

The probability of correct selection is given by (4.6). We now can replace $\tau_{(t)} - \tau_{(j)}$ in this equation by L_j to construct a confidence lower bound of $P(CS)$. However, $v_{(t)(j)}$ is in general not a common value for all $j \ne t$. It depends on the position of the ranked varieties (t) and (j) in the design. Furthermore, $\delta_{(t)}$ has to be known. Since we do not know the ranks of the true variety parameters, it is impossible to calculate the probability of correct selection. A way out is randomisation, which will be described in **4.2**. For an experiment with a variance-balanced design all v_{ij} and all δ_i are equal to each other ($v_{ij} = v$ and $\delta_i = \delta$). In that case the $P(CS)$ can be calculated. If the variance is unknown, a $Q \times 100\%$ confidence upper bound of σ can be calculated as $s/\sqrt{\chi^2_{\nu,1-Q}/\nu}$. (Notice the difference between the Roman letter v and the Greek letter ν.) Then the $(P^* + Q - 1) \times 100\%$ confidence lower bound of $P(CS)$ can be calculated as

$$P(CS)_L = \int_0^{\infty} \int_{-\infty}^{\infty} \left\{ \prod_{i=1}^{t-1} \Phi\left(X + \frac{L_i}{\nu} \frac{\sqrt{\chi^2_{\mathrm{v},1-\varrho}/\mathrm{v}}}{s} + \delta Z \right) \right\} \phi(X) q_{\mathrm{v}}(Z)\, dX dZ \,. \tag{4.9}$$

In case of known variance equation (4.8) can be used, with $1/\sqrt{n}$ replaced by ν and the L_i as used in (4.9) (Driessen, 1991).

The above theory is related to selection of the best variety. The subset selection rules can however be modified in order to be used for other selection goals. We will mention two alternative selection goals : 1) selection of at least one good variety, and 2) selection of all varieties sufficiently better than the average of control varieties. Although a plant breeder prefers to select the best variety, he is often already satisfied if at least one good variety is selected. This selection goal is probably more realistic than selection of the best variety, because the latter goal is rather exacting and often results in a disappointingly large subset size. First it has to be defined what is meant by a good variety. We will classify variety i as good if the variety parameter τ_i is not smaller than $\tau_{(t)} - \delta^*$, so at most δ^* smaller than the parameter of the best variety.

First, suppose that the design is completely randomised and equi-replicated, and that the variance is known. To select the smallest subset that includes at least one good variety with probability at least P^*, the following selection rule can be used (Butler & Butler, 1987) :

Select variety i ($i = 1, \ldots, t$) if and only if

$$\overline{Y}_i \geq \overline{Y}_{[t]} - \frac{\gamma}{\sqrt{n}}\sigma + \delta^* \quad \text{when } \frac{\gamma}{\sqrt{n}} - \delta^* \geq 0\,,$$

$$\overline{Y}_i = \overline{Y}_{[t]} \quad \text{otherwise}\,, \tag{4.10}$$

with the same selection constant γ as in (4.2). In case the variance is unknown, σ in (4.10) is replaced by s, and the same selection constant as in (4.2), associated with unknown variance, is used. The distance measure δ^* indicates which difference is of importance to the plant breeder. It is a matter of indifference to the breeder whether the best variety or a good variety is selected. Following that line of reasoning selection of at least one good variety can be put in the Indifference Zone framework. It can be interpreted as selection of the best variety, with a

Preference Zone taken into account. As in the Bechhofer approach, the Preference Zone is defined as the part of the parameter space with configurations where $\tau_{(t)} - \tau_{(t-1)} \geq \delta^*$. Then the infimum of the $P(CS)$, with the configuration of parameters included in the Preference Zone, is obtained when all differences $\tau_{(t)} - \tau_{(i)}$ $(i \neq t)$ are taken equal to δ^* in (4.3). Consequently,

$$P(CS) \geq \int_{-\infty}^{\infty} \Phi^{t-1}\left(X + \gamma + \frac{\delta^*}{\sigma/\sqrt{n}}\right)\phi(X)dX$$

$$= \int_{-\infty}^{\infty} \Phi^{t-1}(X + \Delta)\phi(X)dX \text{ , with } \Delta = \gamma + \frac{\delta^*}{\sigma/\sqrt{n}} .$$

Then selection rule (4.2) can be written as

$$\overline{Y}_i \geq \overline{Y}_{[t]} - \frac{\gamma}{\sqrt{n}}\sigma$$

$$= \overline{Y}_{[t]} - \frac{\Delta}{\sqrt{n}}\sigma + \delta^* ,$$

and thus becomes equal to rule (4.10) with γ substituted by Δ.

If $\frac{\gamma}{\sqrt{n}}\sigma - \delta^* < 0$, then

$$n > \left(\frac{\gamma\sigma}{\delta^*}\right)^2 ,$$

which satisfies the requirements for the number of observations for Indifference Zone selection. Hence when only the variety with the largest variety value estimate is selected the P^*-requirement is met.

For the situation of an experiment with an (in)complete block design, and variance unknown, selection rule (4.5) can be modified to :

Select variety i $(i = 1, \ldots, t)$ if and only if

$$\hat{\tau}_i \geq \hat{\tau}_j - \delta_i v_{ij} s + \delta^* , \; \forall j \neq i . \qquad (4.11)$$

Only for variance-balanced designs this selection rule is equal to rule (4.5) with the Preference Zone taken into account (like selection rules (4.10) and (4.2)), because then v_{ij} is a common value. Using selection rule (4.11) and the Preference Zone, the probability of correct selection can be written as

$$
\begin{aligned}
P(CS) &= P\left(\frac{\hat{\tau}_{(j)} - \hat{\tau}_{(t)} - (\tau_{(j)} - \tau_{(t)})}{v_{(t)(j)}s} \le \delta_{(t)} + \frac{\tau_{(t)} - \tau_{(j)}}{v_{(t)(j)}s} - \frac{\delta^*}{v_{(t)(j)}s} \quad , \forall \ j \ne t \right) \\
&\ge P\left(\frac{\hat{\tau}_{(j)} - \hat{\tau}_{(t)} - (\tau_{(j)} - \tau_{(t)})}{v_{(t)(j)}s} \le \delta_{(t)} + \frac{\delta^*}{v_{(t)(j)}s} - \frac{\delta^*}{v_{(t)(j)}s} \quad , \forall \ j \ne t \right) \\
&\ge P\left(\frac{\hat{\tau}_{(j)} - \hat{\tau}_{(t)} - (\tau_{(j)} - \tau_{(t)})}{v_{(t)(j)}s} \le \delta_{(t)} \quad , \forall \ j \ne t \right),
\end{aligned}
$$

so the selection constants of rule (4.5) can be used in rule (4.11). If the subset is empty, we can select the variety that we would have selected if δ^* was such that the subset size was equal to 1. Usually this will be the variety with the largest variety parameter estimate, but this is not self-evident.

Often potential varieties are compared with varieties that are currently on the market. This is often done in an advanced selection stage. The currently used varieties are included in the experiment together with the new varieties and are called control varieties. The parameters of the control varieties are not known but are estimated from the data of the experiment. The aim of the breeder can be to select all varieties that are sufficiently better than the average of the control varieties. Let the estimated average variety parameter of the control varieties be denoted by $\hat{\tau}_0$ and let a variety be considered 'sufficiently better' if the variety parameter of that variety is at least δ^* larger than the average control parameter. Gupta & Sobel (1958) assumed a design comparable with the equi-replicated completely randomised design, and a single control variety. Assume that the variance is known. Then the selection rule analogous to selection rule (4.2) reads:

Select variety i ($i = 1, \ldots, t$) if and only if

$$\overline{Y}_i \ge \overline{Y}_0 - \frac{\gamma}{\sqrt{n}}\sigma + \delta^* . \qquad (4.12)$$

It is not known how many varieties are sufficiently better than the control variety. To calculate the minimum probability of correct selection we then have to assume that all the new varieties are sufficiently better than the control variety and so all have to be selected. Then the minimum probability of correct selection is equal to

$$P(CS) \geq P\left(\frac{\overline{Y}_0 - \tau_0}{\sigma/\sqrt{n}} \leq \frac{\overline{Y}_i - \tau_i}{\sigma/\sqrt{n}} + \gamma ,\ i = 1, \ldots, t \right)$$

$$= \int_{-\infty}^{\infty} \Phi^t(X+\gamma)\phi(X)\, dX \ .$$

So the selection constant γ is equal to the selection constant for the selection-of-the-best rule (4.2), calculated for $t+1$ varieties. Analogous results can be obtained for the situation of unknown variance by replacing σ by s.

For (in)complete block designs, and the variance unknown, the selection rule analogous to (4.5) reads :

Select variety i $(i = 1, \ldots, t)$ if and only if

$$\hat{\tau}_i \geq \hat{\tau}_0 - \delta_0 v_{i0} s + \delta^* \ . \qquad (4.13)$$

In the selection-of-the-best situation we do not know which variety is the really best, therefore we have to calculate a selection constant for each variety. In case we make a selection with respect to control varieties, the variety numbers of these controls are known beforehand, so we only have to calculate a single selection constant δ_0. If a single control variety is used, this selection constant is identical to the selection constant associated with this control variety, corresponding to the selection-of-the-best rule with separate selection constants for each variety; calculated for the complete design with $t+1$ varieties. This is not the case with several control varieties (see also **4.3**). Using selection rule (4.13) we can make the inference that the probability that all varieties sufficiently better than the average of the control varieties are included in the subset is at least P^*, with the subset size as small as possible.

All the above mentioned selection procedures make no inferences about the ranking of the varieties. For the situation of an equi-replicated completely randomised design Hsu (1981) introduced simultaneous $P^* \times 100\%$ confidence lower bounds for $\tau_i - \tau_{(t)}$ $(i = 1, \ldots, t)$, with which it is possible to rank the varieties. Lower bounds for $\tau_i - \tau_{(t)}$ are also lower bounds for $\tau_i - \max_{j \neq i} \tau_j$. Hsu (1984) showed that it is possible to add simultaneous confidence upper bounds (say U_i) for $\tau_i - \max_{j \neq i} \tau_j$ to the lower bounds (say L_i) without decreasing the confidence level. These bounds are equal to :

$$L_i = \min\left(0, \overline{Y}_i - \max_{j \neq i} \overline{Y}_j - \gamma\right),$$

$$U_i = \max\left(0, \overline{Y}_i - \max_{j \neq i} \overline{Y}_j + \gamma\right),$$

with γ the Gupta-rule selection constant for known or unknown variance, depending on the knowledge about σ^2.

Driessen (1991) extended the theory of Hsu (1984), which is known as the (constrained) Multiple Comparisons with the Best approach, to experiments with blocks. The lower limits and the upper limits can be calculated as

$$L_i = \begin{cases} 0 & \text{if } S = \{i\} \\ \min\limits_{\substack{j \in S \\ j \neq i}} \{\hat{\tau}_i - \hat{\tau}_j - \delta_j v_{ij} s\} & \text{else,} \end{cases}$$

$$U_i = \max\left(0, \min_{j \neq i} \{\hat{\tau}_i - \hat{\tau}_j + \delta_i v_{ij} s\}\right), \tag{4.14}$$

with S the subset selected by selection rule (4.5) and δ_i ($i = 1, \ldots, t$) the selection constants associated with this rule (Driessen, 1991). Both Hsu and Driessen stress the fact that the construction of the confidence intervals and the selection of the subset including the best variety can be performed with a simultaneous confidence level P^*.

The Multiple Comparisons with the Best approach offers the plant breeder the possibility to achieve yet another selection goal : the selection of a subset that consists of good varieties only. Now the subset size is chosen as large as possible, given the P^*-requirement. Remember that a variety is called 'good' if $\tau_i - \max_{j \neq i} \tau_j \geq -\delta^*$. The selection rule then reads :

Select variety i ($i = 1, \ldots, t$) if and only if

$$L_i \geq -\delta^* . \tag{4.15}$$

The probability that all selected varieties are good is then at least P^*, with the subset size as large as possible.

4.1.2 Model II selection

As already mentioned, original Model II selection aims at the improvement of genetic populations as a whole by means of selection and random mating of the selected varieties. So the population is of primary interest and not the varieties, which differ in every generation and are considered to be a random sample from the population of all varieties possible. However, we will describe a selection approach that uses theory corresponding to Model II selection, but in which the varieties themselves are of primary interest. Assume t varieties, included in variety trials at one or more sites in a particular year. After the performance of the trials the variety values can be estimated. Suppose the experimental design used is completely randomised, so that the estimator of the value of variety i ($i = 1, \ldots, t$) is equal to $\overline{Y}_i$, the average of the observations of variety i. While estimating, the variety terms in the model are taken fixed. We now assume that all $\overline{Y}_i$ have equal variance σ^2/n (or that the differences are so small that they can be considered equal).

Although the variety terms were taken fixed for estimation purposes, it is assumed that they are a random sample from a near-infinite Normal population with mean 0 and variance σ_T^2. If we intend to improve the group of varieties with respect to the mean variety value, we can select a predetermined number, say t_k, of varieties. The varieties corresponding to the largest estimates are selected. Hence the ultimate selection percentage is equal to $\pi = t_k/t$. The reduction from t to t_k varieties is sometimes made in a single selection stage ($k = 1$), but often two ($k = 2$) or three ($k = 3$) selection stages are used, with each selection stage a different year of experimentation. If more than one selection stage is used, the number of varieties is gradually reduced to t_k after the final stage. If only one selection stage is used, it could be worthwile to test only a fraction p_0 of the total number of varieties available, using more replications per variety. The selection percentages in the i^{th} selection stage are denoted by p_i ($i = 0, 1, \ldots, k$), with $\prod p_i = \pi$. If k stages of selection are used, then in a particular year k cohorts of varieties are tested, each of which entered the selection process $k-1$, $k-2$, ..., 0 years ago. Assume that all these varieties must be grown at a fixed area of land, denoted by A. Then the area assigned to the i^{th} selection stage is denoted by A_i, with $\Sigma A_i = A$. Finney (1958) and Curnow (1961) devised selection schemes (values of p_i and A_i/A) for programmes to select varieties for further development. The selection goal they used can be defined as follows : maximise the difference between the expected variety value in the selected group of varieties and the expected variety value in the original group. This difference is called the expected gain of selection, and is often presented in standardised form (hence divided by σ_T).

First consider the situation of a single stage of selection. The estimators $\overline{Y}_i$ can be written as

$$\overline{Y}_i = T_i + \frac{\sigma}{\sqrt{n}}\chi ,$$

with T_i $(i = 1, \ldots, t)$ the random variety value and χ the standard Normal variable. Now the variance of the estimators is equal to $\text{var}(\overline{Y}_i) = \sigma_T^2 + \sigma^2/n$. Curnow (1961) assumed that σ^2/n is proportional to the area of land available :

$$\frac{\sigma^2}{n} = \frac{c\,t}{A}\sigma_T^2 ,$$

with c the constant of proportionality. Without loss of generality we can assume that $E(T_i) = 0 \;\; \forall\, i$. The expected value of $\overline{Y}_i$ for the selected varieties can then be approximated by :

$$E(\text{selected } \overline{Y}_i) = \sqrt{\sigma_T^2 + \sigma^2/n}\,\frac{z(p_1)}{p_1} ,$$

with $z(p_1)$ the ordinate of the standard Normal distribution at the point above which lies a proportion p_1 of the distribution. It can be shown that $z(p_1)/p_1$ is the expected value of the top p_1 of the standard Normal distribution (Cochran, 1951). The given formula of the expected value of $\overline{Y}_i$ for the selected varieties gives a slight over-estimation, because in practice we are selecting a number of varieties from a finite population and not a fraction of an infinite population. The regression coefficient associated with regression of the true variety values on the estimated variety values is equal to :

$$\beta = \frac{\text{cov}(\overline{Y}_i\,, T_i)}{\text{var}(T_i)} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2/n} ,$$

so we can calculate the expectation of the variety value of the selected varieties as

$$E(\text{selected } T_i) = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2/n} \sqrt{\sigma_T^2 + \sigma^2/n} \frac{z(p_1)}{p_1}$$

$$= \frac{\sigma_T^2}{\sqrt{\sigma_T^2 + \sigma^2/n}} \frac{z(p_1)}{p_1} .$$

With $E(T_i) = 0$, this is also the expected gain of selection. Finney (1958) studied the effect of an initial random discard of a fraction $(1 - p_0)$ on the gain of selection. Since only $p_0 t$ varieties are tested, we write :

$$\frac{\sigma^2}{n} = \frac{c\, p_0 t}{A} \sigma_T^2 .$$

Because fewer varieties are tested on the fixed area A, they can be tested with increased replication. Furthermore p_1 now becomes $p_1 = \pi/p_0$, because t_k is a predetermined value. Then the standardised gain can be written as

$$\frac{E(\text{selected } T_i) - \mu}{\sigma_T} = \frac{1}{\sqrt{1 + (c\, p_0 t)/A}} \frac{z(\pi/p_0)}{\pi/p_0} .$$

For known ct/A and π, the p_0 that maximises the standardised gain can be calculated. Finney (1958) found that initial discarding often increases the gain of selection, especially when the heritability is small. However, we will not make further use of this approach, because it is not likely that a plant breeder will discard varieties *a priori*.

Also for the situation of $k > 1$ stages, Finney (1958) evaluated the standardised gain, now with the aim to find optimum values of p_i and A_i/A. He assumed that in every stage new estimates are calculated from the results of that year, and that the selection is based on these estimates only. Furthermore he assumed $p_0 = 1$. To evaluate the gain of selection for two-stage selection, Finney used formulae for the cumulants after one selection from a Normal distribution and a series expansion for the mean after selection from a general distribution. For k-stage selection in general, he gave a very useful practical recommendation to use the following scheme :

$$p_1 = p_2 = \ldots = p_k = \pi^{1/k} ,$$

$$A_1 = A_2 = \ldots = A_k = A/k .$$

So when the number of tested varieties decreases, the number of replications increases. Because all p_i and A_i are equal, this scheme is called a symmetric scheme. Young (1972) found that more than three stages of selection is not worthwile. If there is much variety × year interaction, each selection stage should be based on the results of several years (Curnow, 1961).

4.1.3 Model I selection versus Model II selection

In this section we will compare the Subset selection approach described in *4.1.1* and the approach of selecting a predefined number of varieties, described in *4.1.2*. First we will make some general remarks.

The approach of Finney takes into consideration that a breeding programme often contains several stages of selection. If we look at the breeding procedure of new pollinators of sugar beets (Table 2.2), we see that in year 2 experimental hybrids are selected, which are tested again in year 4 and further selected. This is an example of two-stage selection. The recommendation that Finney gives is clear and easy to use in practice. A plant breeder is able to plan his selection programme in advance, which is highly convenient. He is free to choose the final number of selected varieties, and probably he will choose this number rather small. However, in this freedom there might be some danger. For how large should he choose t_k ? Often used is a selection percentage of about 10 %, depending also on the total number of varieties tested. A weakness in this approach is that we have no idea about the probability of correct selection. No quantitative measure for this is given.

This weakness of the Finney approach is the strength of the Subset approach. With probability at least P^* the best variety is included in the subset. However, this probability statement is valid for one stage of selection. Also, a difference between the two approaches is that the subset approach makes no use of an assumption about the distribution of the variety terms. The $P_{LFC}(CS)$ is the probability of correct selection calculated for a configuration where all variety parameters are equal to each other. This is a very unfavourable configuration indeed, which will not be found in practice. The Normality assumption seems more realistic. In section **4.5** we will make use of this assumption for subset selection. As for the Finney approach, the symmetric scheme is also nearly optimal for distributions other than the Normal one. Because the Subset selection rule is designed so that it guarantees a probability of correct selection of at least P^* for all configurations of variety parameters, the real probability of correct selection will be much higher if a Normal distribution of the variety terms is assumed. To decrease the subset sizes, selection of at least one good variety could be a useful selection goal. In addition to the often unpleasant value of the subset size, it is also random. Hence the plant breeder cannot plan the selection programme in advance.

In the comparison, different selection schemes will be used. In each situation we assume that the varieties are tested in a completely randomised design. Then selection takes place in one, two or three stages. This is done using the approach of predefined t_k and equal selection percentages in each stage, and by Subset selection in every stage. For the latter approach, σ^2 is assumed known and $P^* = 0.80$. Using simulation, the expected final number of selected varieties, the expected standardised gain, the expected standardised difference between the variety parameter of the best variety and the variety parameter of the best selected variety and the expected probability of correct selection was approximated, based on 10000 iterations. 'Standardised' means that it is expressed in units of σ_T. These statistics are calculated for 4 values of the heritability h^2 in the starting population : $h^2 = 0$, 0.1, 0.3 and 0.5. The heritability h^2 is here defined as

$$h^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma^2},$$

and should not be confused with the 'effective' heritability $\tilde{h}^2 = \sigma_T^2/(\sigma_T^2 + \sigma^2/n)$.

1) Assume 100 varieties are grown in an experiment with a completely randomised design with 2 observations per variety (case (i)), or 4 observations per variety (case (ii)). The selection programme consists only of one stage. Using the Finney approach, 10 varieties are selected, hence $\pi = 0.10$. We assume that the variety terms are a random sample from a Normal distribution. The results of the simulation are presented in Table 4.1.

From Table 4.1 it becomes clear that with Subset selection we have to select more than 10 varieties; much more if the heritability is small. The fact that more varieties have to be selected, implicates that also some less good varieties are selected with the Subset approach. This results in a low expected gain of selection, compared with the 10 % rule. The less good varieties decrease the average of the selected group. However, if we look at the expected difference between the best variety and the selected best variety, we notice that with Subset selection this expected difference is much smaller than with the 10 % selection. With the latter type of selection we seem to loose the very good varieties more often than with

Subset selection. This also results in a lower probability of correct selection, especially if the heritability is small. The minimum probability of correct selection is 0.80 for the Subset approach and only 0.10 for the 10 % rule.

Table 4.1. Estimates of : expected number of selected varieties ($\hat{E}(|S|)$), expected standardised gain ($\hat{E}(G/\sigma_T)$), expected standardised difference between the variety parameter of the best variety and the variety parameter of the best selected variety ($\hat{E}(D/\sigma_T)$) and expected probability of correct selection ($\hat{E}(PCS)$); approximated by simulation (10000 runs) for Subset selection ($P^* = 0.80$, σ^2 known) and 10% selection in one stage, starting with 100 varieties. Assumption : Variety terms are a random sample from a Normal distribution, and heritability (h^2) is given. Design : CRD with (i) 2 and (ii) 4 replications.

		(i)		(ii)	
	h^2	Subset selection	10 % selection	Subset selection	10 % selection
$\hat{E}(\|S\|)$	0	80.0	10	80.0	10
	0.1	70.8	10	62.6	10
	0.3	50.5	10	34.9	10
	0.5	31.4	10	18.5	10
$\hat{E}(G/\sigma_T)$	0	0	0	0	0
	0.1	0.211	0.739	0.347	0.962
	0.3	0.562	1.176	0.896	1.375
	0.5	0.983	1.413	1.396	1.548
$\hat{E}(D/\sigma_T)$	0	0	0	0	0
	0.1	0.015	0.345	0.009	0.210
	0.3	0.005	0.108	0.003	0.039
	0.5	0.002	0.029	0.001	0.007
$\hat{E}(PCS)$	0	0.80	0.10	0.80	0.10
	0.1	0.95	0.41	0.97	0.56
	0.3	0.98	0.71	0.99	0.85
	0.5	0.99	0.88	0.99	0.96

In plant breeding the difference between the best varieties currently on the market and the better variety to be found is very small. But *if* a better variety is found and commercialized, this gives large financial revenues. Therefore, a plant breeder is anxious to select the few varieties at the top. It is obvious that a measure like the probability of correct selection is of great value to him, and he is probably willing to test the varieties more intensively if this probability appears to be small. Selection with predefined t_k and maximisation of the expected gain of selection is associated with true Model II selection, where the plant breeder is not interested in the varieties themselves. If the interest lies in the actual varieties tested, then the selection should be aimed at selection of the best variety, or at least one good variety, or all varieties better than a control, so a selection procedure must be used that guarantees the probability of correct selection. Table 4.1 shows that if we can assume that the variety terms are a random sample from a Normal distribution, then the expected probability of correct selection is much larger than the guaranteed P^*. So if we use a selection rule that satisfies the P^*-requirement, more varieties then strictly necessary are selected.

The probability of correct selection also depends on the effective heritability. The higher the effective heritability, the easier it is to select the best variety. For 2 replications, the effective heritability corresponding with $h^2 = 0, 0.1, 0.3, 0.5$ is $\tilde{h}^2 = 0, 0.18, 0.46, 0.67$, respectively. With 4 replications, the effective heritability increases to $\tilde{h}^2 = 0, 0.31, 0.63, 0.80$, respectively. If we can assume that the variety terms are a random sample from a Normal distribution and if the heritability is known, we can use results as in Table 4.1 to design the experiment. For instance, if it is known that $h^2 = 0.1$, then two replications are sufficient to reach an expected $P(CS)$ of 0.95.

2) We will now study a situation where the selection programme consists of two stages. Assume 400 varieties are grown in an experiment with a completely randomised design with 2 observations per variety in the first stage. Following Finney's advice, we can choose the number of observations per variety in the second stage so that the areas used in both selection stages are approximately equal (case (i)). If the selection fractions in both stages are chosen equal to $\sqrt{0.025}$, the ultimate selection percentage is equal to 2.5 %, which corresponds with $t_k = 10$. For case (i) and this Finney approach the number of observations in the second stage will then be 13. For the Subset approach the number of selected varieties is random, so the number of observations in the second stage is also random. A slightly different situation is created if the number of observations per variety in the second stage is fixed to 4 (case (ii)), for both selection procedures. We assume that the variety terms are a random sample from a Normal distribution. The results of the simulation are presented in Table 4.2.

From Table 4.2 we get the same impression as from Table 4.1. Especially for small heritabilities, selection of a predefined number of varieties can result in a low expected probability of correct selection and a large expected difference between the best variety and the selected best variety. Because there are two stages, the minimum probability of correct selection guaranteed with the Subset rule is now 0.8 × 0.8 = 0.64. It is however remarkable that the expected probability of correct selection lies already above 0.9 for $h^2 = 0.1$.

Table 4.2. Estimates of : expected number of selected varieties ($\hat{E}(|S|)$), expected standardised gain ($\hat{E}(G/\sigma_T)$), expected standardised difference between the variety parameter of the best variety and the variety parameter of the best selected variety ($\hat{E}(D/\sigma_T)$) and expected probability of correct selection ($\hat{E}(PCS)$); approximated by simulation (10000 runs) for Subset selection ($P^* = 0.80$, σ^2 known) and $\sqrt{0.025}\times 100\%$ selection in two stages, starting with 400 varieties. Assumption : Variety terms are a random sample from a Normal distribution, and heritability (h^2) is given. Design : CRD with 2 replications in the first stage and (i) approximately equal area per stage, (ii) 4 replications in the second stage.

		(i)		(ii)	
	h^2	Subset selection	$\sqrt{0.025}\times 100\%$ selection	Subset selection	$\sqrt{0.025}\times 100\%$ selection
$\hat{E}(\mid S\mid)$	0	256.0	10	256.0	10
	0.1	183.3	10	170.3	10
	0.3	67.6	10	69.1	10
	0.5	17.0	10	27.4	10
$\hat{E}(G/\sigma_T)$	0	0	0	0	0
	0.1	0.482	1.684	0.543	1.380
	0.3	1.342	2.102	1.263	1.909
	0.5	2.226	2.229	1.865	2.106
$\hat{E}(D/\sigma_T)$	0	0	0	0	0
	0.1	0.016	0.195	0.014	0.322
	0.3	0.003	0.028	0.004	0.068
	0.5	0.001	0.003	0.002	0.014
$\hat{E}(PCS)$	0	0.64	0.025	0.64	0.025
	0.1	0.93	0.53	0.94	0.39
	0.3	0.98	0.88	0.98	0.75
	0.5	0.99	0.98	0.99	0.92

3) Finally, we will study a situation where the selection programme consists of three stages. Assume 400 varieties are grown in an experiment with a completely randomised design with 2 observations per variety in the first stage, 4 observations

per variety in the second stage and 6 observations per variety in the third stage. In case (i) it is assumed that the variety terms are a random sample from a Normal distribution. In case (ii) it is assumed that the variety terms are a random sample from a Uniform distribution. For the Finney type of selection the selection fractions in all stages are chosen equal to $(0.025)^{1/3}$, hence the ultimate selection percentage is equal to 2.5 %, corresponding with $t_k = 10$. The results of the simulation are presented in Table 4.3.

Table 4.3. Estimates of : expected number of selected varieties ($\hat{E}(|S|)$), expected standardised gain ($\hat{E}(G/\sigma_T)$), expected standardised difference between the variety parameter of the best variety and the variety parameter of the best selected variety ($\hat{E}(D/\sigma_T)$) and expected probability of correct selection ($\hat{E}(PCS)$); approximated by simulation (10000 runs) for Subset selection ($P^* = 0.80$, σ^2 known) and $(0.025)^{1/3} \times 100\%$ selection in three stages, starting with 400 varieties. Assumption : Variety terms are a random sample from a (i) Normal distribution and (ii) Uniform distribution, and heritability (h^2) is given. Design : CRD with 2 replications in the first stage, 4 replications in the second stage and 6 replications in the third stage.

		(i)		(ii)	
	h^2	Subset selection	$(0.025)^{1/3} \times 100\%$ selection	Subset selection	$(0.025)^{1/3} \times 100\%$ selection
$\hat{E}(\lvert S\rvert)$	0	204.8	10	204.8	10
	0.1	98.7	10	117.1	10
	0.3	28.5	10	68.4	10
	0.5	10.5	10	47.5	10
$\hat{E}(G/\sigma_T)$	0	0	0	0	0
	0.1	0.895	1.617	0.834	1.265
	0.3	1.778	2.049	1.243	1.464
	0.5	2.347	2.184	1.403	1.539
$\hat{E}(D/\sigma_T)$	0	0	0	0	0
	0.1	0.019	0.196	0.003	0.046
	0.3	0.007	0.027	0.003	0.024
	0.5	0.005	0.005	0.002	0.015
$\hat{E}(PCS)$	0	0.51	0.025	0.51	0.025
	0.1	0.92	0.54	0.75	0.15
	0.3	0.97	0.87	0.80	0.26
	0.5	0.97	0.97	0.81	0.36

With three selection stages, the minimum probability of correct selection reduces to $(0.80)^3 = 0.51$ for the Subset selection rule, but if the heritability in the starting population is 0.1 the expected $P(CS)$ is already 0.92 for case (i); see Table 4.3. The minimum probability of correct selection for the Finney type of rule is 0.025 and the expected $P(CS)$ remains low for small heritabilities. For case (ii), where the variety terms are a random sample from a Uniform distribution, the conclusions made before remain valid. Because a Uniform distribution has no long tails, it is more difficult to select the best variety. Therefore the expected probabilities of correct selection are not as high as with the Normal distribution assumption. The expected probabilities of correct selection for the Finney type of rule are very low.

Recapitulating, if the plant breeder wants to select varieties in a way which guarantees a minimum probability of correct selection, he has to use Subset selection instead of selecting a predetermined number of varieties. Then he has to take it for granted that the number of selected varieties is random. The subset size depends on the chosen P^*, the experimental design used, the experimental error and the actual configuration of the variety parameters. If one (or more) of these factors is (are) such that it is simply not possible to select a small subset, the plant breeder has to face this. If he does not, and still selects a small number of varieties, the probability that he has selected the desired varieties is low. The Finney approach does not advice us how many varieties to select. In the studied cases we just as well could have used a different selection percentage. Furthermore, the goal of maximising the expected gain of selection is closely connected with maximising the Response to Selection, which is the goal in real Model II selection. This goal is more suited in case the breeder is not interested in the actual varieties tested, but in the selected group of varieties as a whole.

4.2 Subset selection procedures for randomised experiments [1]

Suppose we have t ($t > 1$) varieties and we are mainly interested in selecting the best variety, which is defined here as the variety with the largest variety parameter. Thence, we conduct an experiment in which the t varieties are tested and make a selection on the basis of the data resulting from the experiment. For this purpose subset selection can be used, as described in **4.1.** In this section, we introduce selection rules for randomised experiments. First, in *4.2.1*, we will explain what is meant by 'randomised experiment'. Second, in *4.2.2*, two types of subset selection procedures (for selection of the best variety) that satisfy the P^*-requirement, if the experiment is randomised, will be given. The first type consists of rules that need a separate selection constant for each individual variety. Rules of the second type need only one selection constant for all varieties. We will illustrate selection rules from both types by application to a variety trial in *4.2.3* and compare them in *4.2.4*. The developed ideas can also be applied to selection procedures which aim at selecting at least one good variety. This is described in *4.2.5*.

4.2.1 Randomised experiments

If we want to test t (e.g. sugar beet) varieties in an experiment and draw conclusions from the results obtained, we have to decide on the experimental design and the (statistical) model to be used. For this purpose we make use of t imaginary varieties, which at this moment have nothing to do with the varieties we want to test. To avoid confusion we will call the imaginary varieties 'design varieties' (because they are associated with the experimental design) and the really existing varieties 'actual varieties'. The actual (sugar beet) varieties have promising names like 'Univers', 'Regina' or 'Herald', or in an earlier stage they have a code name. For notational convenience we are forced to give the actual varieties the in-expressive names $V1$, $V2$, ..., Vt. The design varieties are denoted by numbers $1, 2, \dots, t$. Until now we have tacitly used the notation associated with the design varieties, using variety numbers $1, 2, \dots, t$. For (in)complete blocks experiments we have described a fixed additive model on the basis of design varieties in *3.1.1*. In this section we will introduce a randomisation procedure that 'assigns' the actual

[1] Co-author : Stefan Driessen

varieties to the design varieties. The model for observations of design varieties combined with this assignment procedure then gives a model for the observations of the actual varieties. We will restrict ourselves to completely randomised and connected randomised block designs (with equal block sizes). The combination of a randomised design and randomisation procedure that assigns actual varieties to design varieties leads towards a randomised experiment.

If we want to use a completely randomised design for the experiment, we have to choose and randomly select a number of experimental units n_i $(i = 1, 2, \ldots, t)$ for each individual design variety. For this design we assume the following model:

$$Y_{ij} = \lambda + \tau_i + E_{ij}\,, \quad i = 1, 2, \ldots, t\,;\; j = 1, 2, \ldots, n_i\,. \tag{4.16}$$

We assume that the E_{ij}'s are uncorrelated, Normally distributed variables with zero expectation and common variance σ^2.

For a randomised block design, we have to decide on the number (b) and size of the blocks, and the allocation of the design varieties to the blocks. We first randomise the blocks and next randomly assign the units of a block to the design varieties of that block. As already described in *3.1.1*, for a randomised block design the model can be written as

$$Y_{ijk} = \lambda + \tau_i + \beta_j + E_{ijk}\,,\; i = 1, 2, \ldots, t\,;\; j = 1, 2, \ldots, b\,;\; k = 1, 2, .., n_{ij}\,, \tag{4.17}$$

where the number of experimental units for each individual design variety i in block j is denoted by n_{ij}. Notice that both models (4.16) and (4.17) have additive parameters, so we assume that there are no interactions between design varieties and units (or blocks). To estimate the variety parameters of design varieties 1, 2, ..., t we will use the least squares estimators and denote the latter by $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_t$.

The choice of the experimental design determines the variances of the estimators of the contrasts between design variety parameters. The estimated variance of the estimator of the difference between the parameters of design varieties i and j was denoted by $v_{ij}^2 s^2$ $(i, j = 1, 2, \ldots, t)$, with s^2 the usual unbiased estimate of the error variance. For a given design matrix, we can calculate v_{ij}^2 for each pair of design varieties. In case of a balanced design, v_{ij}^2 is a common value for all pairs (i, j) ,$i, j = 1, 2, \ldots, t$ $(i \neq j)$. Unbalanced designs give unequal values for v_{ij}^2; some pairwise contrasts will be estimated more precisely than others. Sometimes we want to estimate particular contrasts of variety parameters more precisely than others, and in that case we deliberately select an unbalanced design with which some contrasts of design variety parameters are estimated more precisely. This is often done in case a control variety is involved in the experiment, and this control is replicated more often than the other varieties. When the set of varieties does not have a special structure, we usually prefer a design which is as near as possible (under the restrictions of block size and number of replications) to a completely balanced design. Often we use an already developed design, which sometimes asks adjustment of the number of varieties. For example, a square lattice design needs a square number of varieties.

We may use the information about the variances and the structure of the set of varieties when we assign the t actual varieties to the t design varieties, which is the next important step in designing the experiment. When we deliberately want to estimate particular contrasts of actual variety parameters more precisely than others, we assign the relevant actual varieties to the design varieties of which the parameter contrasts are estimated more precisely. When all contrasts are of the same importance, the approach is to randomly assign the actual varieties to the design varieties. Whichever method is used, there always must be a one-to-one correspondence between the actual varieties and the design varieties. Let H be the set of one-to-one 'functions' from the set of actual varieties to the set of design varieties. For sake of convenience we call these 'functions' assignments, of which there are $t!$. Let $\eta(V1) = 2$ mean that variety $V1$ is assigned to design variety 2 by the assignment η, $\eta \in$ H. So $\eta(Vj)$ represents the name of the design variety to which the actual variety Vj is assigned by assignment η. Let for any assignment $\eta \in$ H, $P_r[\eta]$ be the probability that varieties Vj are assigned to design varieties $\eta(Vj)$, $j = 1, 2, \ldots, t$, by assignment η. Of course $\Sigma_{\eta \in \mathrm{H}} P_r[\eta] = 1$. We will call an experiment **randomised** if the design of this experiment is randomised and the actual varieties are assigned to the design varieties by means of a defined randomisation process, i.e. the probabilities $P_r[\eta]$ are given for all $\eta \in$ H.

Remark.

The procedure described above has already been summarized by Cochran and Cox (1957, pp. 442/443) in three steps:

1. Rearrange the blocks at random. (If the design is arranged in complete replications, the blocks are randomised only within each replication and the replications are kept separate).

2. Randomise the position of the variety numbers separately and independently within each block.
3. Assign the varieties at random to the variety numbers in the plan.
Here of course variety number is a synonym for design variety.

Given assignment η, (4.16) can be rewritten for the completely randomised design as

$$Y_{\eta(Vi)j} = \lambda + \tau_{\eta(Vi)} + E_{\eta(Vi)j} ,$$

and (4.17) for the randomised block design as

$$Y_{\eta(Vi)jk} = \lambda + \tau_{\eta(Vi)} + \beta_j + E_{\eta(Vi)jk} .$$

We have to keep in mind that these are models for a given assignment. The variance of the estimator of $\tau_{Vi} - \tau_{Vj}$ (a contrast between parameters of actual varieties) will be denoted by $v^2_{(Vi)(Vj)}\sigma^2$ and is equal to $v^2_{\eta(Vi)\eta(Vj)}\sigma^2$ given the assignment η of the actual varieties to the design varieties.

A different way of giving names to varieties (irrespective of the names 'design' and 'actual' varieties) is to describe the value of this variety with respect to the other varieties. This notation has already been used in *4.1.1* and concerns the subscripts (i) and $[i]$ $(i = 1, 2, \ldots, t)$. Let (1) denote the worst variety, ..., (t) denote the best variety. Hence the (unknown) variety with rank number i is denoted by (i). The ranking of the varieties is based on the variety parameters, such that $\tau_{(1)} \le \tau_{(2)} \le \ldots \le \tau_{(t)}$. Analogous to the notation of the design variety to which variety Vi is assigned ($\eta(Vi)$), the design variety to which variety (i) is assigned by the same assignment η is denoted by $\eta((i))$. From the above mentioned notation it follows that correct selection occurs when variety (t) is included in the selected subset. Of course we do not know which variety is really the best, which variety is really the best but one, and so on. This leads to a third way of naming the varieties, based on the ranking as given by the results of the experiment. Let [1] denote the, according to the data, worst variety, ..., $[t]$ denote the, according to the data, best variety. This ranking is based on the estimated variety parameters, such that $\hat{\tau}_{[1]} \le \hat{\tau}_{[2]} \le \ldots \le \hat{\tau}_{[t]}$.

4.2.2 Selection rules for selection of the best variety

We want to construct selection rules in such a way that the probability of correct selection using a randomised design ($P_R(CS)$), calculated for the Least Favourable Configuration (LFC) of the parameters, is equal to a prespecified value P^*. For the studied designs, the assignment of the actual varieties to the design varieties with use of assignment $\eta \in \mathrm{H}$ has an influence on the probability of correct selection, because this probability depends on the position of the best variety, the position of the best but one variety, ..., the position of the worst variety in the experimental design. In general, we cannot calculate the probability of correct selection without knowing these positions (Driessen, 1991), but in case of a randomised experiment we can. The probability of correct selection for a randomised experiment is equal to

$$P_R(CS) = \sum_{\eta \in \mathrm{H}} P(CS \mid \eta) P_r(\eta) . \tag{4.18}$$

Thus we calculate the probability of correct selection by conditioning on the assignment of the actual varieties to the design varieties. As we will see, this implies that two types of selection rules that meet the P^*-requirement exist.

For the first type of selection rules we use a separate selection constant for each individual design variety i. Of this type we present two rules, denoted by *R1* and *R2*, which will now be specified in detail. Intuitively it seems wise to take account of variance differences in case of unbalanced designs. Therefore in rule *R1* v_{ij} is included.

Rule *R1* is defined as follows :
Randomly select an assignment $\eta \in \mathrm{H}$ and select the actual variety assigned to design variety i $(i = 1, 2, \ldots, t)$ if and only if
$\hat{\tau}_i \geq \hat{\tau}_j - \delta_i v_{ij} s$, $\forall j \neq i$.

To execute the above selection rule all the contrast variances have to be calculated. It is, however, also possible to define a rule that satisfies the P^*-requirement but does not contain v_{ij}. This rule, rule *R2*, is attractive because it is relatively simple. Later we will study the effect of the omission of v_{ij} on the expected subset size.

Rule $R2$ is given by :
Randomly select an assignment $\eta \in \mathrm{H}$ and select the actual variety assigned to design variety i $(i = 1,2,\dots,t)$ if and only if
$\hat{\tau}_i \geq \hat{\tau}_{[t]} - \Delta_i s$.

The values δ_i, Δ_i $(i = 1,2,\dots,t)$ are the selection constants for the rules $R1$ and $R2$, respectively, and will be defined later. Rule $R1$ is an extension to randomised experiments of selection rule (4.5).

We only have a correct selection when variety (t) is selected. So, using $R1$, we can work out (4.18) as follows :

$$\begin{aligned}
P_R(CS) &= \sum_{\eta \in \mathrm{H}} P(CS \mid \eta) P_r(\eta) \\
&= \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{(t)} \geq \hat{\tau}_j - \delta_{(t)} v_{(t)j} s \ , \ \forall j \neq (t) \mid \eta) P_r(\eta) \\
&= \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{(t)} \geq \hat{\tau}_{(j)} - \delta_{(t)} v_{(t)(j)} s \ , \ \forall j \neq t \mid \eta) P_r(\eta) \\
&= \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{(j)} - \hat{\tau}_{(t)} - (\tau_{(j)} - \tau_{(t)}) \leq \delta_{(t)} v_{(t)(j)} s + \tau_{(t)} - \tau_{(j)} \, , \ \forall j \neq t \mid \eta) P_r(\eta) \\
&= \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{\eta((j))} - \hat{\tau}_{\eta((t))} - (\tau_{\eta((j))} - \tau_{\eta((t))}) \leq \delta_{\eta((t))} v_{\eta((t))\eta((j))} s + \tau_{(t)} - \tau_{(j)} \, , \ \forall j \neq t) P_r(\eta) \, .
\end{aligned} \tag{4.19}$$

Without loss of generality, let the set of assignments H be subdivided into the following subsets $\mathrm{H}_1, \mathrm{H}_2, \dots, \mathrm{H}_t$: the $\eta \in \mathrm{H}_l$ are assignments with which the best variety (t) is assigned to design variety l, $l = 1,2,\dots,t$. Because $\tau_{(t)} - \tau_{(j)} \geq 0 \ \forall j$, it is clear that the $P_R(CS)$ is minimal when $\tau_{(1)} = \tau_{(2)} = \dots = \tau_{(t)}$, as such forming the LFC. Therefore the minimum probability of correct selection is the $P_R(CS)$ for the LFC, denoted by $P_{R,LFC}(CS)$, and we have

$$\begin{aligned}
P_{R,LFC}(CS) &= \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{\eta((j))} - \hat{\tau}_{\eta((t))} - (\tau_{\eta((j))} - \tau_{\eta((t))}) \leq \delta_{\eta((t))} v_{\eta((t))\eta((j))} s \ , \ \forall j \neq t) P_r(\eta) \\
&= \sum_{l=1}^{t} \left\{ P(\hat{\tau}_j - \hat{\tau}_l - (\tau_j - \tau_l) \leq \delta_l v_{lj} s \ , \ \forall j \neq l) \sum_{\eta \in \mathrm{H}_l} P_r(\eta) \right\} .
\end{aligned} \tag{4.20}$$

Let the t selection constants $\delta_1, \delta_2, \dots, \delta_t$ be such that

$$P(\hat{\tau}_j - \hat{\tau}_l - (\tau_j - \tau_l) \leq \delta_l v_{lj} s \ , \ \forall j \neq l) = P^* \, , \ l = 1,2,\dots,t \, . \tag{4.21}$$

Then, by substituting (4.21) in (4.20), the minimum probability of correct selection becomes

$$\begin{aligned} P_{R,LFC}(CS) &= P^* \sum_{l=1}^{t} \sum_{\eta \in \mathrm{H}_l} P_r(\eta) \\ &= P^* \sum_{\eta \in \mathrm{H}} P_r(\eta) \\ &= P^*, \end{aligned}$$

which was our probability requirement. Notice that no extra assumptions are necessary for the randomisation process that assigns the actual varieties to the design varieties. The calculation of the selection constants is described in Driessen (1991), resulting in the evaluation of a multivariate t probability, for which no simple numerical procedures are available. In case of unbalanced designs with more than a few varieties a numerical integration procedure is not feasible, but we can use computer simulation to calculate the selection constants. This is described in **4.3.**

For rule *R2* we get the same type of results. The selection constants $\Delta_1, \Delta_2, \ldots, \Delta_t$, corresponding to this rule, are defined by the equations

$$P(\hat{\tau}_j - \hat{\tau}_l - (\tau_j - \tau_l) \leq \Delta_l s \ , \ \forall j \neq l) = P^* \ , \ l = 1, 2, \ldots, t \ . \tag{4.22}$$

This leads, in the same way as selection rule *R1*, to a prespecified minimum probability of correct selection P^* $(= P_{R,LFC}(CS))$, without any extra assumption about the randomisation process.

The second type of selection rules uses a single selection constant for all varieties. This is of course very convenient for practical use. Two rules will be presented, denoted by *R3* and *R4*.

Rule *R3* is defined as follows :
Randomly select an assignment $\eta \in \mathrm{H}$ and select the actual variety assigned to design variety i $(i = 1, 2, \ldots, t)$ if and only if
$\hat{\tau}_i \geq \hat{\tau}_j - \delta v_{ij} s \ , \ \forall j \neq i \ .$

Notice that *R3* uses only a single selection constant δ and that this selection rule is the same as *R1* if $\delta_1 = \delta_2 = \ldots = \delta_t = \delta$.

Rule *R4* is given by :
Randomly select an assignment $\eta \in \mathrm{H}$ and select the actual variety assigned to the design variety i $(i = 1, 2, \ldots, t)$ if and only if
$\hat{\tau}_i \geq \hat{\tau}_{[t]} - \Delta s$.

This rule corresponds with the second rule *R2* of the first type and is identical to it if $\Delta_1 = \Delta_2 = \ldots = \Delta_t = \Delta$.

Following the same reasoning as with *R1* we find that the LFC for *R3* again is the set $\{\tau : \tau_{(1)} = \tau_{(2)} = \ldots = \tau_{(t)}\}$ with the corresponding probability of correct selection :

$$P_{R,LFC}(CS) = \sum_{l=1}^{t} \left\{ P(\hat{\tau}_j - \hat{\tau}_l - (\tau_j - \tau_l) \leq \delta v_{lj} s \ , \ \forall j \neq l) \sum_{\eta \in \mathrm{H}_l} P_r(\eta) \right\} . \tag{4.23}$$

Let the selection constant δ be such that

$$\sum_{l=1}^{t} \left\{ P(\hat{\tau}_j - \hat{\tau}_l - (\tau_j - \tau_l) \leq \delta v_{lj} s \ , \ \forall j \neq l) \sum_{\eta \in \mathrm{H}_l} P_r(\eta) \right\} = P^* , \tag{4.24}$$

then clearly *R3* meets the P^*-requirement. The left-hand function is monotonically increasing in δ and so for $1/t < P^* < 1$ there always exists a solution of (4.24).

For *R4* we have to solve the following equation in Δ :

$$\sum_{l=1}^{t} \left\{ P(\hat{\tau}_j - \hat{\tau}_l - (\tau_j - \tau_l) \leq \Delta s \ , \ \forall j \neq l) \sum_{\eta \in \mathrm{H}_l} P_r(\eta) \right\} = P^*, \tag{4.25}$$

to satisfy the P^*-requirement. For the same reasons as for selection rule *R3* a solution for Δ exists.

If selection of the best variety is our purpose, a randomisation procedure for which each assignment η has the same probability, i.e.

$$P_r(\eta) = \frac{1}{t!} , \ \forall \eta \in \mathrm{H} ,$$

seems appropriate.

4.2.3 Example

We will demonstrate the four selection rules, using the results of a trial conducted by the plant breeding company The Royal Vanderhave Group (The Netherlands). The experimental design of the trial is a 5x5 lattice design with four replications. This design can be found in Cochran and Cox (1957). The aim of the experiment was to compare 25 new sugar beet hybrids with respect to their white sugar yield and to select a subset that includes the best hybrid with a certain minimum probability. The best sugar beet hybrid is defined here as the hybrid with the largest white sugar yield. The design varieties were randomised as described in section *4.2.1* and the hybrids were assigned completely at random to the design varieties. Unfortunately, the plants in the last three blocks of the fourth replication were killed during the growing season. So, if we use the design mentioned in Cochran and Cox (1957), then the hybrids that were assigned to design varieties 3, 4, 5, 6, 7, 8, 11, 14, 15, 17, 18, 19, 21, 22 and 25 only have 3 replications. Due to this accident the experimental design became more unbalanced. We assume the randomised block design model described in section *4.2.1*. With the design of the trial given, we can calculate the selection constants of rules *R1* and *R2*. These selection constants have been approximated by means of simulation (see section **4.3**), for $P^* = 0.80$ and $P^* = 0.90$. They read :

	δ_i for *R1*		Δ_i for *R2*	
design variety number	$P^* = 0.80$	$P^* = 0.90$	$P^* = 0.80$	$P^* = 0.90$
1	2.13	2.50	1.77	2.07
2	2.14	2.49	1.77	2.07
3	2.04	2.42	1.81	2.17
4	2.05	2.45	1.82	2.18
5	2.07	2.45	1.84	2.19
6	2.04	2.44	1.82	2.18
7	2.04	2.42	1.81	2.16
8	2.05	2.42	1.83	2.18
9	2.14	2.51	1.77	2.08
10	2.12	2.49	1.75	2.07
11	2.03	2.42	1.81	2.16
12	2.13	2.52	1.76	2.08
13	2.11	2.51	1.75	2.08
14	2.04	2.44	1.82	2.17
15	2.04	2.45	1.82	2.18
16	2.12	2.50	1.76	2.06
17	2.05	2.44	1.83	2.18
18	2.04	2.42	1.81	2.16
19	2.03	2.41	1.81	2.15
20	2.15	2.54	1.77	2.10
21	2.04	2.42	1.82	2.16
22	2.04	2.43	1.82	2.17
23	2.13	2.50	1.76	2.07
24	2.12	2.50	1.76	2.07
25	2.04	2.43	1.82	2.17

For the calculation of the selection constant of rule *R3* and *R4* we must assume that the randomisation process is known. We assume that the sugar beet hybrids have been assigned to the design varieties completely at random. The selection constants of rules *R3* and *R4* are approximately equal to :

δ for *R3*		Δ for *R4*	
$P^* = 0.80$	$P^* = 0.90$	$P^* = 0.80$	$P^* = 0.90$
2.08	2.46	1.80	2.12

After the assignment of the sugar beet hybrids to the design varieties, the experiment can be performed. The estimated standard deviation was 0.358 ton/ha. The least squares estimates of the variety values were :

design variety	actual variety [1]	white sugar yield (ton/ha)	design variety	actual variety [1]	white sugar yield (ton/ha)
1	VDH05	13.44	13	VDH21	12.61
2	VDH12	12.00	14	VDH16	12.97
3	VDH23	13.16	15	VDH04	12.39
4	VDH11	12.64	16	VDH13	9.77
5	VDH18	12.22	17	VDH22	12.07
6	VDH01	12.14	18	VDH09	11.44
7	VDH20	12.50	19	VDH02	12.22
8	VDH14	12.04	20	VDH24	12.48
9	VDH17	12.45	21	VDH15	12.00
10	VDH07	11.73	22	VDH19	12.90
11	VDH25	12.44	23	VDH03	12.08
12	VDH10	12.83	24	VDH08	12.51
			25	VDH06	12.14

[1] The actual hybrid names are coded to maintain trade secrecy.

The, according to the trial, ordered sugar beet hybrids are VDH05 > VDH23 > VDH16 > VDH19 > VDH10 > VDH11 > VDH21 > VDH08 > VDH20 > VDH17 > VDH24 > VDH25 > VDH04 > VDH18 = VDH02 > VDH01 = VDH06 > VDH03 > VDH22 > VDH14 > VDH12 = VDH15 > VDH07 > VDH09 > VDH13. The selected subsets corresponding to the above mentioned experiment can now be calculated. To demonstrate the selection procedures, we will work out the selection rules for sugar beet hybrid VDH10, with $P^* = 0.80$. Hybrid VDH10 was assigned to design variety 12. The values of $v_{12\,1}, \ldots, v_{12\,11}, v_{12\,13}, \ldots, v_{12\,25}$ are equal to 0.7771, 0.7766, 0.8647, 0.8647, 0.8373, 0.8367, 0.8373, 0.8647, 0.7766, 0.7973, 0.8367, 0.7766, 0.8380, 0.8367, 0.7973, 0.8367, 0.8373, 0.8647, 0.7766, 0.8647, 0.8373, 0.7771, 0.7766, 0.8647.

- For selection rule *R1* the selection constant associated with hybrid VDH10 is equal to $\delta_{12} = 2.13$ for $P^* = 0.80$. Now

$12.83 < \max_{j \neq 12}(\hat{\tau}_j - 2.13 \times v_{12j} \times 0.358) = 13.44 - 2.13 \times 0.7771 \times 0.358 = 12.85$.

Hence VDH10 is not selected at $P^* = 0.80$. The same way it is checked whether the other varieties have to be selected or not. The result is that selection rule *R1* selects the hybrids that were assigned to design varieties 1, 3, 14 and 22 (the hybrids VDH05, VDH23, VDH16 and VDH19) with $P^* = 0.80$ and to design varieties 1, 3, 12, 14 and 22 (the hybrids VDH05, VDH23, VDH10, VDH16 and VDH19) with $P^* = 0.90$.

- For selection rule *R2* the selection constant associated with hybrid VDH10 is equal to $\Delta_{12} = 1.76$ for $P^* = 0.80$. Now

$12.83 > 13.44 - 1.76 \times 0.358 = 12.81$.

Hence VDH10 is selected at $P^* = 0.80$. After the rule is executed for every hybrid, selection with rule *R2* results in a subset that includes the hybrids that were assigned to design varieties 1, 3, 12, 14 and 22 (these are VDH05, VDH23, VDH10, VDH16 and VDH19) for both $P^* = 0.80$ and $P^* = 0.90$.

- For selection rule *R3* the selection constant for hybrid VDH10, as for all other hybrids, is equal to $\delta = 2.08$ for $P^* = 0.80$. Now

$12.83 < \max_{j \neq 12}(\hat{\tau}_j - 2.08 \times v_{12j} \times 0.358) = 13.44 - 2.08 \times 0.7771 \times 0.358 = 12.86$.

Hence VDH10 is not selected at $P^* = 0.80$. Making the selection decision for every hybrid, it turns out that selection rule *R3* selects exactly the same subset as selection rule *R1* .

- For selection rule *R4* the selection constant for hybrid VDH10, as for all other hybrids, is equal to $\Delta = 1.80$ for $P^* = 0.80$. Now

$12.83 > 13.44 - 1.80 \times 0.358 = 12.80$.

Hence VDH10 is selected at $P^* = 0.80$. The ultimate subset corresponding to selection rule *R4* is exactly the same as the subset obtained with selection rule *R2* .

So the two rules that make use of the variances of the estimators of the contrasts between variety parameters are better in this example. Because it is much easier to work with a common selection constant for all design varieties, we prefer selection rule *R3* . A further advantage of selection rule *R3* is that the selection constant for this rule can be calculated much faster than the selection constants of selection rule *R1* (See **4.3**).

4.2.4 Comparison of the selection rules

In this section we will compare the four selection rules that have been described in *4.2.2*. All the rules are designed in such a way that the infimum (over all configurations of the variety parameters) of the probability of correct selection is equal to P^*. In other aspects, however, they most probably will differ from each other.

For prespecified values of P^* the selection rules $R1$, $R2$, $R3$ and $R4$ will be defined such that they meet the P^*-requirement. Next, the rules will be compared with respect to the expected subset size ($E_R(|S|)$). (In hypothesis testing, tests are often compared with respect to their powers provided they have the same significance level.) $E_R(|S|)$ depends on the true values of the variety parameters and the experimental design used. The ranking of the selection rules needs not be the same for different configurations of variety parameter values and/or different designs, and therefore an integral comparison is not feasible. We will therefore confine ourselves to the discussion of two experimental designs and a limited number of configurations of variety parameter values.

The selection constants are determined (by simulation, see **4.3**) so that the corresponding rules satisfy the P^*-requirement approximately. In order to calculate the selection constant of rule $R3$ and $R4$, the randomisation process must be known. We assume a completely random assignment of the actual varieties to the design varieties ($P_r(\eta) = 1/t!$). This also holds for the calculation of the expected subset size. The expected subset sizes are approximated by simulation (see **4.3**).

The different types of configurations of variety parameters, which we will use throughout this section, are the following :

- A so-called slippage configuration, with distance parameter q. This is the set of configurations that satisfy $\tau_{(1)} = \tau_{(2)} = \ldots = \tau_{(t-1)} = \tau_{(t)} - q\sigma$, with σ the root of the error variance. We denote this configuration by SL(q).
- An equidistant configuration, with distance parameter q. This is the set of configurations that satisfy $\tau_{(1)} = \tau_{(2)} - q\sigma = \tau_{(3)} - 2q\sigma = \ldots = \tau_{(t)} - (t-1)q\sigma$. We denote this configuration by EQ(q).
- A configuration of variety parameters which are drawn from a Normal distribution with zero mean and standard deviation $q\sigma$. We denote this configuration by NO(q).

Most of the configurations used are rather artificial. In plant breeding, it is sometimes assumed that the variety terms are a random sample from a Normal distribution with variance σ_T^2. The ratio between σ_T^2 and σ^2 is expressed in the heritability (in broad sense) :

$$h^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2} = \frac{\sigma_T^2/\sigma_e^2}{\sigma_T^2/\sigma_e^2 + 1}.$$

With the q parameter corresponding to the NO-configurations, the heritability can be written as

$$h^2 = \frac{q^2}{q^2 + 1}.$$

Now we can give the following translation table :

h^2	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
$q(NO)$	0.10	0.23	0.33	0.50	0.65	0.82	1.00	1.22	1.52	2.00	3.00

All entries of the tables presented in this section are calculated by simulation, based on 100,000 iterations per entry. Two designs will be considered.

The first design for which we have calculated the comparison criterium for all selection rules is a completely randomised design with 5 varieties and $n_1 = 100$, $n_2 = \ldots = n_5 = 2$. The results, for $P^* = 0.90$, are presented in Table 4.4. The expected subset sizes of the selection rules indicate that the rules which make use of v_{ij} (*R1* and *R3*) are slightly better than the other two rules, although rule *R4* is the best one in case of a slippage configuration with a very large distance parameter. Selection rule *R1* seems to be the best rule for the other configurations, but the differences between the results of rule ***R1*** and rule *R3* are very small.

The second experimental design that we will discuss is a 7x7 triple lattice of which three blocks (belonging to one replication) are missing. We studied the comparison criterium for rules *R3* and *R4* only and for $P^* = 0.75, 0.80, 0.90$. The results are presented in Table 4.5. This table shows that for the 7x7 triple lattice design with missing values the expected subset size for selection rule *R3* is smaller than for rule *R4*, except for slippage configurations with a very large difference parameter.

Table 4.4. Expected subset size, estimated by simulation (100000 iterations), for selection rules *R1*, *R2*, *R3* and *R4*, satisfying the probability requirement at level $P^* = 0.90$. Completely randomised design ; $t = 5$, $n_1 = 100$ and $n_2 = \ldots = n_5 = 2$.

	R1	*R2*	*R3*	*R4*
SL(0.1)	4.496	4.496	4.498	4.497
SL(1.0)	3.860	3.957	3.861	3.988
SL(2.0)	2.276	2.383	2.296	2.391
SL(3.0)	1.296	1.297	1.313	1.275
SL(4.0)	1.035	1.031	1.039	1.026
SL(5.0)	1.002	1.002	1.002	1.001
EQ(0.10)	4.439	4.449	4.439	4.453
EQ(0.25)	4.114	4.171	4.116	4.193
EQ(0.50)	3.249	3.370	3.263	3.383
EQ(0.75)	2.536	2.635	2.550	2.632
EQ(1.00)	2.089	2.158	2.100	2.157
NO(0.1)	4.465	4.470	4.466	4.473
NO(0.23)	4.302	4.331	4.302	4.347
NO(0.33)	4.102	4.159	4.103	4.182
NO(0.50)	3.665	3.766	3.672	3.791
NO(0.65)	3.267	3.385	3.279	3.396
NO(0.82)	2.867	2.981	2.880	2.983
NO(1.0)	2.525	2.623	2.538	2.621
NO(1.5)	1.924	1.981	1.933	1.977
NO(2.0)	1.594	1.633	1.602	1.633

Other results have shown, as to be expected, that it is impossible to draw general conclusions and to point out one selection rule as being the best one. The ranking of the selection rules according to expected subset size depends on the configuration of the variety parameters, the minimum probability of correct selection (read : P^*) and the experimental design. However, a conclusion may be that the selection rules which make use of the root of the variances of estimators of contrasts between variety parameters (i.e. *R1* and *R3*) are on the average better than the other selection rules, but the differences are small. This was also found in the example in *4.2.3*. For the NO-configuration, meaningful distance parameters for the plant breeding practice are given by the heritability. For a heritability between 0.01 and 0.80 we see in the tables that the best selection rules for Normal configurations are those that use v_{ij}. In practice the heritability is almost always within this range. In Table 4.4 for the completely randomised design with five varieties we see that rule *R1* is the best one. However, the difference with rule *R3* is very small. It should also be noticed that for variance-balanced designs and some partially balanced designs selection rules *R1* and *R3* are identical. For unbalanced designs with many varieties, for example the triple 7x7 lattice with missing values, the calculation of the selection constants and the expected subset sizes for selection rule *R1* becomes very laborious, and the rule is not convenient to work with in such a situation. Therefore a general practical conclusion could be to use selection rule *R3*. Since the differences between the expected subset sizes for the various selection rules are small, one could also decide to use selection rule *R4*, because this rule is very convenient to work with.

Table 4.5. Expected subset size, estimated by simulation (100000 iterations), for selection rules *R3* and *R4*, satisfying the probability requirement at levels $P^* = 0.75, 0.80, 0.90$. Randomised triple 7x7 lattice design ($t = 49$), of which three blocks belonging to one replication are missing.

	P^*	*R3*	*R4*		P^*	*R3*	*R4*
SL(0.5)	0.75	36.449	36.484	NO(0.01)	0.75	36.763	36.769
	0.80	38.876	38.947		0.80	39.167	39.204
	0.90	43.851	43.880		0.90	44.055	44.056
SL(1.0)	0.75	34.856	35.021	NO(0.1)	0.75	36.364	36.397
	0.80	37.402	37.606		0.80	38.800	38.860
	0.90	42.790	42.952		0.90	43.792	43.821
SL(2.0)	0.75	24.313	24.898	NO(1.0)	0.75	16.491	16.807
	0.80	27.083	27.804		0.80	18.265	18.678
	0.90	34.071	34.951		0.90	23.355	24.012
SL(4.0)	0.75	2.583	2.491	NO(5.0)	0.75	2.814	2.826
	0.80	3.107	3.027		0.80	2.902	2.916
	0.90	5.241	5.254		0.90	3.182	3.204
EQ(0.01)	0.75	35.932	35.989				
	0.80	38.392	38.480				
	0.90	43.496	43.552				
EQ(0.1)	0.75	14.332	14.551				
	0.80	15.588	15.861				
	0.90	19.073	19.479				
EQ(1.0)	0.75	2.532	2.561				
	0.80	2.657	2.692				
	0.90	3.006	3.053				
EQ(2.0)	0.75	1.563	1.577				
	0.80	1.625	1.642				
	0.90	1.800	1.820				

4.2.5 Selection of at least one good variety

The approach that uses knowledge about the randomisation process that assigns actual varieties to design varieties to create a selection rule with a single selection constant can also be applied to selection rules that do not aim at selection of the best variety. Also rules for selection of at least one good variety can be defined such that they only need a single selection constant. In *4.1.1* it was explained that selection of at least one good variety can be seen as selection of the best variety with an Indifference Zone taken into account. There, the selection-of-the-best rule was redefined into a selection-of-at-least-one-good rule. Consequently, the selection-of-the-best rules in *4.2.2* with a single selection constant can be redefined into selection-of-at-least-one-good rules.

Variety i $(i = 1, 2, \ldots, t)$ is classified as 'good' if the variety parameter τ_i is not less than $\tau_{(t)} - \delta^*$, with δ^* a distance measure defined by the plant breeder. Then rule *R3* can be redefined into a selection rule for selection of at least one good variety as :

Randomly select an assignment $\eta \in$ H and select the actual variety assigned to design variety i $(i = 1, 2, \ldots, t)$ if and only if

$$\hat{\tau}_i \geq \hat{\tau}_j - \delta v_{ij} s + \delta^*, \ \forall j \neq i .$$

If the subset is empty, we can select the actual variety assigned to the design variety that we would have selected if δ^* was such that the subset size was equal to 1.

Analogous to the above situation rule *R4* can be redefined as :

Randomly select an assignment $\eta \in$ H and select the actual variety assigned to design variety i $(i = 1, 2, \ldots, t)$ if and only if

$$\hat{\tau}_i \geq \hat{\tau}_{[t]} - \Delta s + \delta^* \text{ when } \Delta s - \delta^* \geq 0 ,$$

$$\hat{\tau}_i = \hat{\tau}_{[t]} \text{ otherwise .}$$

4.3 The use of simulation in statistical selection

To execute statistical selection rules the selection constants first have to be determined. In *4.3.1* it is described how to determine the selection constants by means of simulation. In *4.3.2* we compare the results with those obtained by numerical integration. In order to compare different selection rules, we would like to determine the expected subset size. In *4.3.3* it is shown how this can be done by simulation. Also the probability of correct selection can be determined that way. We will compare simulation results with computations by numerical integration.

4.3.1 Approximation of selection constants

Suppose we have t varieties which are tested in a randomised experiment. Let the model for the observations be equal to (4.17). To select the best variety we can use subset selection rules *R1*, *R2*, *R3* and *R4*, given in *4.2.2*. In **4.2** we have given the equations that have to be solved to determine the selection constants corresponding to the four selection rules. We now will rewrite them in vector notation. We define $\mathbf{T}_i$ $(i = 1, 2, \ldots, t)$ as

$$\mathbf{T}_i = \Big(\frac{\hat{\tau}_1 - \hat{\tau}_i - (\tau_1 - \tau_i)}{v_{i\,1}\sigma}, \ldots, \frac{\hat{\tau}_{i-1} - \hat{\tau}_i - (\tau_{i-1} - \tau_i)}{v_{i\,i-1}\sigma},$$

$$\frac{\hat{\tau}_{i+1} - \hat{\tau}_i - (\tau_{i+1} - \tau_i)}{v_{i\,i+1}\sigma}, \ldots, \frac{\hat{\tau}_t - \hat{\tau}_i - (\tau_t - \tau_i)}{v_{i\,t}\sigma}\Big)'.$$

Given the experimental design, we can calculate each selection constant δ_i $(i = 1, 2, \ldots, t)$ for selection rule *R1* with σ^2 unknown as the solution of

$$P^* = P\left(\frac{\mathbf{T}_i}{s/\sigma} \le \delta_i \mathbf{1}_{t-1}\right) \quad \text{(Driessen, 1991)}. \tag{4.26}$$

From the model it follows that $\mathbf{T}_i/(s/\sigma)$ has a standard multivariate t-distribution with a particular correlation matrix. Note that if σ^2 is known, then $s = \sigma$ and we have a multivariate Normal distribution. For selection rule *R2* the selection constants Δ_i are the solutions of

$$P^* = P\left[\frac{\mathbf{T}_i}{s/\sigma} \le \left(\frac{\Delta_i}{v_{i\,1}}, \ldots, \frac{\Delta_i}{v_{i\,i-1}}, \frac{\Delta_i}{v_{i\,i+1}}, \ldots, \frac{\Delta_i}{v_{i\,t}}\right)'\right]. \tag{4.27}$$

Assume that the actual varieties are assigned at random to the design varieties. Then for selection rule *R3* the selection constant δ is the solution of

$$P^* = \frac{1}{t}\sum_{i=1}^{t} P\left(\frac{\mathbf{T}_i}{s/\sigma} \leq \delta \mathbf{1}_{t-1}\right). \tag{4.28}$$

For selection rule *R4* the selection constant Δ can be found by solving

$$P^* = \frac{1}{t}\sum_{i=1}^{t} P\left[\frac{\mathbf{T}_i}{s/\sigma} \leq \left(\frac{\Delta}{v_{i\,1}}, \ldots, \frac{\Delta}{v_{i\,i-1}}, \frac{\Delta}{v_{i\,i+1}}, \ldots, \frac{\Delta}{v_{i\,t}}\right)\right]. \tag{4.29}$$

If we want to select all varieties sufficiently better than the average of the control varieties, we can use selection rule (4.13), which in the sequel will be denoted by *R5*. We now define selection rule *R6*, which resembles *R5* and can be used for the same purpose :

Select variety i $(i = 1, .., t)$ if and only if

$\hat{\tau}_i \geq \hat{\tau}_0 - \Delta_0 s + \delta^*$,

where $\hat{\tau}_i$ is the estimate of the parameter corresponding to new variety i, $\hat{\tau}_0$ is the average of the estimates of the parameters corresponding to the control varieties, Δ_0 is the selection constant and δ^* is used to define when a variety is considered 'sufficiently better' than the control varieties (see also *4.1.1*). For known variance the s in the rule is substituted by σ.

Using the same notation as in **4.2**, the minimum probability of correct selection for rule *R5* is equal to :

$$P_{R,LFC}(CS) \geq \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{\eta(0)} - \hat{\tau}_{\eta(i)} - (\tau_{\eta(0)} - \tau_{\eta(i)}) \leq \delta_0 v_{\eta(0)\eta(i)} s, \quad \forall i \neq 0) P_r(\eta).$$

Suppose we have two control varieties a and b. If we equate $P_{R,LFC}(CS)$ to P^*, we can calculate the selection constant δ_0 by solving the equations :

$$P^* = P\left(\frac{\mathbf{T}_{\frac{\eta(a)+\eta(b)}{2}}}{s/\sigma} \leq \delta_0 \mathbf{1}_t\right). \tag{4.30}$$

Here $\mathbf{T}_{\frac{\eta(a)+\eta(b)}{2}}$ is defined as

$$\mathbf{T}_{\frac{\eta(a)+\eta(b)}{2}} = \left(\frac{\hat{\tau}_{\eta(1)} - \frac{\hat{\tau}_{\eta(a)}+\hat{\tau}_{\eta(b)}}{2} - \left(\tau_{\eta(1)} - \frac{\tau_{\eta(a)}+\tau_{\eta(b)}}{2}\right)}{v_{\frac{\eta(a)+\eta(b)}{2}\eta(1)}\sigma}, \ldots, \frac{\hat{\tau}_{\eta(t)} - \frac{\hat{\tau}_{\eta(a)}+\hat{\tau}_{\eta(b)}}{2} - \left(\tau_{\eta(t)} - \frac{\tau_{\eta(a)}+\tau_{\eta(b)}}{2}\right)}{v_{\frac{\eta(a)+\eta(b)}{2}\eta(t)}\sigma}\right)'.$$

If there is more than one control variety, we cannot use the selection constants calculated for the selection-of-the-best rule for $t+1$ varieties. For selection rule *R6* the equations that we have to solve in order to calculate the selection constant Δ_0 are

$$P^* = P\left[\frac{\mathbf{T}_{\frac{\eta(a)+\eta(b)}{2}}}{s/\sigma} \leq \left(\frac{\Delta_0}{v_{\frac{\eta(a)+\eta(b)}{2}\eta(1)}}, \ldots, \frac{\Delta_0}{v_{\frac{\eta(a)\eta(b)}{2}\eta(t)}} \right)' \right]. \tag{4.31}$$

Note that again we cannot use the selection constants calculated for selection-of-the-best rules. This can only be done in case we have one control variety.

As shown above the calculation of the selection constant boils down to the evaluation of a multivariate t-distribution. For an equi-replicated completely randomised design the parameter v_{ij} is equal to $\sqrt{2/n}$ for all $i \neq j$. Then the vectors $\mathbf{T}_i$ are identically distributed, hence the selection constants of selection rules *R1* and *R2* are equal to the selection constant of rules *R3* and *R4*, respectively. This is in general true for variance-balanced designs. In that case it is possible to calculate the selection constant by numerical integration (see also *4.1.1*). The selection constants of rule *R1* are also equal to each other in case of a partially balanced incomplete block design with two association classes, based on the group divisible association scheme or the triangular scheme or the Latin square scheme (Driessen, 1991). See for tables of PBIB designs with 2 association classes Bose, Clatworthy & Shrikhande (1954). An example of such a design is the partially balanced lattice design. (Finney, 1960)

As contrasted with variance-balanced designs, it is in general too troublesome to calculate selection constants for partially balanced and unbalanced experimental designs by numerical integration. Only for a very small number of varieties numerical integration will work. However, a lot of the experimental designs used in practice are unbalanced or become unbalanced during the testing period because of accidents. For practical use the research worker is willing to accept approximate selection constants, as long as the approximation is accurate enough. We will describe a computer program that makes it possible to approximate the selection constants for all kinds of designs, using simulation methods.

A computer program, named SELCON, was written by the author in Fortran 77, with additional use of subroutines from IMSL (International Mathematical & Statistical Libraries, version 1.0). We will describe the program SELCON, at the same time explaining the simulation method. The computer program makes it possible to calculate the selection constants for the selection rules mentioned above. Thus we can use four rules for selection of the best variety and two rules for selection of all varieties sufficiently better than the control variety (or average of the control varieties). We have to choose which selection rule we want to use. Next we have to give the design of the experiment or the pseudo-variance/covariance matrix (divided by σ^2) of the variety parameter estimators and information whether the variance must be assumed known or unknown. The program approximates the minimum probability of correct selection (P^*) for a series of selection constants, which have to be given. Afterwards, we can find the selection constant for a given P^* by interpolation.

The experimental design is reflected by the incidence matrix **N**. This matrix is read from an input file. We will restrict ourselves to connected designs. As in chapter **3** let $\mathbf{p}'\tau$ denote a contrast between design variety parameters. Then we have the reduced normal equations $\mathbf{C}\hat{\tau} = \mathbf{Q}$, and we may use any pseudo-inverse $\mathbf{C}^-$ as if it were the covariance matrix (divided by σ^2) of $\hat{\tau}$ to calculate the variance of the estimator of an estimable contrast $\mathbf{p}'\tau$. Therefore the variance of the estimator of $\mathbf{p}'\tau$ can be calculated as $\mathbf{p}'\mathbf{C}^-\mathbf{p}\sigma^2$ (John, 1971). This estimator $\mathbf{p}'\hat{\tau}$, which is a linear combination of Normally distributed observations, has a normal distribution with expectation $\mathbf{p}'\tau$ and variance $\mathbf{p}'\mathbf{C}^-\mathbf{p}\sigma^2$. Consider the $(t-1)*(t)$ matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} -1 & 1 & & -0- \\ -1 & & 1 & \\ \vdots & & \ddots & \\ -1 & -0- & & 1 \end{pmatrix}.$$

Then $\mathbf{A}\tau$ gives a vector with all pairwise variety parameter contrasts with respect to design variety 1. The estimator $\mathbf{A}\hat{\tau}$ of these contrasts is distributed as a random variable with a $(t-1)$ dimensional multivariate normal distribution with expectation $\mathbf{A}\tau$ and covariance matrix $\mathbf{A}\mathbf{C}^-\mathbf{A}'\sigma^2$.

In the computer program SELCON a pseudo-inverse of **C** is calculated as the inverse of $(\mathbf{C}+\mathbf{1}_t\mathbf{1}_t')$, because the null space of **C** has basis $\mathbf{1}_t$. With this pseudo-inverse of **C** the $v_{i\,j}$ are calculated in case selection rule *R1*, *R3* or *R5* has been chosen. For instance, $v_{1\,2}$ is calculated as $(-1,1,0,\ldots,0)\mathbf{C}^-(-1,1,0,\ldots,0)'$.

Next the covariance matrix of $\mathbf{A}\hat{\tau}$ is calculated. For selection of the best variety, $\mathbf{A}$ has the form as described above. In case of selection with respect to c $(c \geq 1)$ control varieties, we substitute matrix $\mathbf{A}$ by the $t \times (t+c)$ matrix $\mathbf{B}$:

$$\mathbf{B} = \begin{pmatrix} -1/c & \dots & -1/c & 1 & & -0- \\ -1/c & \dots & -1/c & & 1 & \\ \vdots & \cdot & \vdots & & & \ddots \\ -1/c & \dots & -1/c & -0- & & 1 \end{pmatrix}.$$

So $\mathbf{B}\tau$ is a vector with t variety parameter contrasts, which are all estimable. Because we want to calculate the minimum probability of correct selection, we are working with the Least Favourable Configuration. So $\mathbf{A}\tau = \mathbf{0}_{t-1}$ and $\mathbf{B}\tau = \mathbf{0}_t$. For the calculation of the selection constants we may, without loss of generality, assume that the common known variance σ^2 is equal to 1, because the selection constant has no dimension.

The next part of the simulation program is repeated very often (e.g. 10000 times). In the situation of selection of the best variety, a realisation of the $t-1$ variety contrasts from the $t-1$ dimensional multivariate normal distribution with zero expectation and covariance matrix $\mathbf{AC}^{-}\mathbf{A}'$ is generated. Then we have a solution of the reduced normal equations with the assumption that $\hat{\tau}_1 = 0$. In case of selection with respect to a control variety (or an average of control varieties), the t contrasts are generated from the t dimensional multivariate normal distribution with zero expectation and covariance matrix $\mathbf{BC}^{-}\mathbf{B}'$. If the error variance σ^2 is unknown, we estimate this variance by s^2 with ν degrees of freedom. From the model it follows that s^2 is distributed as $\sigma^2\chi^2_\nu/\nu$, with χ^2_ν the Chi-square distributed random variable with ν degrees of freedom. Hence s is distributed as $s \sim \sigma\sqrt{\chi^2_\nu/\nu}$. If we calculate selection constants for the situation of unknown variance, we generate a realisation of s from the $\sqrt{\chi^2_\nu/\nu}$ distribution, because σ is assumed to be 1. This realisation is independent from the contrast realisations. After generating the realisations of the variety contrasts and s we can actually execute the chosen selection rule.

- Selection rule $R1$. This selection rule needs a selection constant δ_i for each individual design variety i $(i = 1,2,\dots,t)$. Therefore we have to calculate P^* corresponding to δ_i for every i, assuming that the best variety (t) is assigned to design variety i. To calculate P^* corresponding to δ_1, we assume that the best variety (t) is assigned to design variety 1. Next we execute selection rule $R1$ with the realisations of the estimated variety parameters and s (if the variance is unknown). If $\hat{\tau}_1 \geq \max_j\{\hat{\tau}_j - \delta_1 v_{1\,j} s\}$, then the best variety is selected and we have a correct selection. This is checked for a series of values of the selection constant. For the situation of known variance s is substituted by σ, which is assumed equal to 1. For design variety 1 the information whether we have a correct selection or not, is stored. The same procedure is repeated for all varieties, thus asssuming that the best variety is assigned to design variety i ($i = 1, 2, \ldots, t$).

- Selection rule *R2*. The same procedure as for selection rule *R1* is used in order to calculate P^* for each design variety. The selection rule is easier : we have a correct selection if $\hat{\tau}_i$ is greater than or equal to the maximum of the other estimated variety parameters minus $\Delta_i s$, when i is assumed to represent the best variety.
- Selection rule *R3* or *R4*. Here we assume that the original varieties are assigned to the design varieties completely at random. Therefore at random one of the design varieties is tagged as being the design variety to which the best variety is assigned. Then we check whether this design variety is selected or not, using selection rule *R3* or *R4* and a series of values of the selection constant.
- Selection rule *R5* or *R6*. In the situation of selection with respect to one or more control varieties, we know which design varieties represent the control varieties. This as contrasted with the situation of selection of the best variety, where we do not know which variety is the best one. As mentioned earlier, a correct selection occurs when all the non-control varieties are selected. Here this is checked using selection rule *R5* or *R6*.

After one simulation round, a new realisation of the variety contrasts is generated and, in case of unknown variance, a new realisation of s is generated. Then again we execute the chosen selection rule and check whether there is a correct selection or not. For selection rules *R3* and *R4* a new design variety is tagged as being the best one. This simulation is repeated m (say) times, and we store the number of correct selections out of the m simulation rounds. This is done for each individual selection constant. Then the minimum probability of correct selection can be estimated by

$$\hat{P}^* = \frac{\text{number of correct selections}}{\text{number of simulations}}.$$

The number of correct selections out of m selections has a binomial distribution with parameters $n = m$ and $p = P^*$. So we can approximate a 95% confidence interval of P^* by

$$[\hat{P}^* - 1.96\sqrt{\hat{P}^*(1-\hat{P}^*)/m} \quad , \quad \hat{P}^* + 1.96\sqrt{\hat{P}^*(1-\hat{P}^*)/m}], \tag{4.32}$$

with 1.96 the 0.975 point of the standard normal distribution. For practical use an approximate lower limit of P^* will be important, because we are not interested in upper limits of P^*. The 95% lower limit will be larger than the lower bound of the 95% confidence interval and therefore better to work with in practice. The approximate 95% confidence lower limit of P^* is calculated as

$$P_{L,0.95} = \hat{P}^* - 1.645\sqrt{\hat{P}^*(1-\hat{P}^*)/m}\,. \tag{4.33}$$

Consider an experimental design of which we know that $\delta_1 = \ldots = \delta_t$ for selection rule $R1$ and $\Delta_1 = \ldots = \Delta_t$ for selection rule $R2$. It is already mentioned that in this situation we can use the selection constant of selection rule $R3$ and $R4$, respectively. However, if we do not want to use the simulation method used for selection rules $R3$ or $R4$, which assumed a completely random assignment of the original varieties to the design varieties, we can use the following method. In section *4.1.1* we have seen that the expected subset size ($E(|S|)$) can be written as

$$E(|S|) = \sum_{i=1}^{t} P(\text{ variety } i \text{ selected}),$$

with S the random number of selected varieties. But for the LFC situation and selection rules $R1$ and $R2$, the minimum probability of correct selection, given that the best variety is assigned to design variety i, is also calculated as the probability that this design variety is selected. For the specific experimental designs under consideration these probabilities are all equal. Hence we can estimate the minimum probability of correct selection as

$$\hat{P}^* = \frac{\hat{E}(|S|)}{t}.$$

The simulation program SELCON offers the possibility to use this method. During the simulation, a frequency table of the subset size is created. In this frequency table f_i indicates the number of times that the subset contained i varieties, with $i = 1, 2, \ldots, t$. Then the expected subset size is estimated by $\hat{E}(|S|) = \Sigma\, i f_i / m$. The variance of S is estimated by $\hat{\text{var}}(|S|) = [\Sigma\, i^2 f_i - (\Sigma\, i f_i)^2/m]/(m-1)$. Hence, the minimum probability of correct selection and the corresponding variance can, for the specific designs under consideration, be estimated by

$$\hat{P}^* = \frac{\hat{E}(|S|)}{t} = \frac{\overline{|S|}}{t}, \tag{4.34}$$

$$\hat{\text{var}}(\hat{P}^*) = \frac{\hat{\text{var}}(|S|)}{t^2 m}. \tag{4.35}$$

A 95% confidence interval for P^* is approximated by

$[\hat{P}^* - 1.96\sqrt{\hat{\text{var}}(\hat{P}^*)} \quad , \quad \hat{P}^* + 1.96\sqrt{\hat{\text{var}}(\hat{P}^*)}]$,

and an approximate 95% lower limit of P^* is $P_{L,0.95} = \hat{P}^* - 1.645\sqrt{\hat{\text{var}}(\hat{P}^*)}$.

The run-time of the simulation program depends on the type of computer used, the experimental design and the chosen selection rule. Using an Olivetti M380/XP4 personal computer, the selection constants for a simple 5x5 lattice design are calculated in 54 minutes for selection rule *R1*, in 10 minutes for selection rule *R2*, in 5 minutes for selection rule *R3* and in 3 minutes for selection rule *R4*. However , using a VAX-8700 (Digital) mainframe computer, it only takes 34 seconds CPU-time to calculate the selection constant for selection rule *R4*. The differences in run-time for the various selection rules indicate to avoid selection rule *R1* when the design is rather extensive.

We want to use the program to find the selection constants for a given P^*. Therefore, the minimum probability of correct selection is estimated for a range of selection constants, which must be given by the user. The relevant information is the estimated P^* and the 95% lower limit of P^*, for each selection constant. Consider a simple 5x5 lattice design. This design is partially balanced with two association classes and therefore the selection constants δ_i corresponding to selection rule *R1* are identical (Driessen, 1992). The simulation program was executed with a range of selection constants ($\delta_i = \delta$) starting from zero and increasing with steps of 0.1. Then the output of the program SELCON, using selection rule *R1* with unknown variance, for the simple 5x5 lattice design becomes:

δ	$\hat{P}^*$	$P_{L,0.95}$
0.0	0.0400	0.0400
0.1	0.0524	0.0520
0.2	0.0676	0.0670
⋮	⋮	⋮
2.1	0.7881	0.7849
2.2	0.8168	0.8137
⋮	⋮	⋮

If the research worker wants the selection constant corresponding with $P^* = 0.80$, then this value is calculated by log-linear interpolation. In our example the selection constant is calculated as

$$\delta = \exp\left\{ \ln(2.1) + \frac{[\ln(2.2) - \ln(2.1)]\,[\ln(0.8) - \ln(0.7849)]}{[\ln(0.8137) - \ln(0.7849)]} \right\} = 2.1523,$$

using the 95% lower limits of P^*. The table of selection constants and estimated minimum probabilities of correct selection, as shown above, needs to be created only once for a particular experimental design. Thus the research worker can make a library of tables corresponding to the experimental designs that he often uses.

4.3.2 Simulation versus numerical integration

We will compare the results of simulation with the results of numerical integration for three experimental designs. Of course only designs can be used for which it is feasible to calculate the minimum probability of correct selection by numerical integration. The first experimental design is a completely randomised design with 5 varieties and three observations per variety. For such a design the four selection (of the best) rules are identical. We will use selection rule *R1* with $\delta_i = \delta$ and selection rule *R3*. The selection constants, calculated by numerical integration for the situation of known variance, can be found in Bechhofer & Dunnett (1988) for $P^* = 0.80, 0.90, 0.95$ and 0.99 and, multiplied by $\sqrt{2}$, in Butler & Butler (1987) for $P^* = 0.50, 0.80, 0.90, 0.95, 0.975, 0.99, 0.995$ and 0.999. The research worker has to specify P^*. With the selected selection constant as input for δ_i or δ in the simulation program SELCON, the minimum probability of correct selection was estimated. For selection rule *R1* we used the method that makes use of the estimated expected subset size, because for this design all the selection constants δ_i are equal. The estimated probability has to be approximately equal to the exact P^*. The results are presented in Table 4.6 (a). The real goal of the simulation program SELCON is to calculate the selection constants corresponding to a particular P^*. In Table 4.6 (b) the approximate selection constants, obtained by interpolation from the simulation results, are given.

Table 4.6. Minimum probability of correct selection (P^*), selection constant (δ), **(a)** estimated P^* ($\hat{P}^*$), 95% confidence interval of P^* (*95%CI*) and 95% lower limit of P^* (*95%LL*), **(b)** estimated selection constant ($\hat{\delta}$), estimated upper limit of the selection constant ($\hat{\delta}_u$); calculated for selection rules *R1* and *R3* for a completely randomised design with 5 varieties and 3 observations per variety, in case of known variance. Number of simulation rounds : 10000.

(a)

P^*	δ	$\hat{P}^*$		*95%CI*		*95%LL*	
		R1	*R3*	*R1*	*R3*	*R1*	*R3*
0.20	0.00000	0.2000	0.2001	0.2000 , 0.2000	0.1923 , 0.2079	0.2000	0.1935
0.50	0.72072	0.4985	0.5004	0.4940 , 0.5030	0.4906 , 0.5102	0.4948	0.4922
0.80	1.45155	0.8000	0.7978	0.7959 , 0.8041	0.7899 , 0.8057	0.7965	0.7912
0.90	1.83827	0.8996	0.8967	0.8965 , 0.9027	0.8907 , 0.9027	0.8970	0.8917
0.95	2.16033	0.9500	0.9488	0.9479 , 0.9522	0.9445 , 0.9531	0.9482	0.9452
0.99	2.77156	0.9912	0.9894	0.9903 , 0.9921	0.9874 , 0.9914	0.9905	0.9877

(b)

P^*	δ	$\hat{\delta}$		$\hat{\delta}u$	
		R1	*R3*	*R1*	*R3*
0.50	0.72072	0.7233	0.7213	0.7318	0.7413
0.80	1.45155	1.4516	1.4588	1.4620	1.4786
0.90	1.83827	1.8413	1.8546	1.8546	1.8758
0.95	2.16033	2.1593	2.1705	2.1745	2.2028
0.99	2.77156	2.7271	2.8043	2.7541	2.8648

When $\delta = 0$ is used, the minimum probability of correct selection becomes $1/t$. The real P^* in Table 4.6 (a) is only once not included in the approximate 95% confidence interval. The approximate 95% lower limits of P^* lie very close to the real values of P^* and are accurate enough for practical use. From Table 4.6 (b) we see that the approximated selection constants have to be rounded to two decimals. To achieve that the estimated selection constants are closer to the true selection constants, the number of simulation rounds has to be increased. For instance, with 100,000 simulations the estimated selection constant corresponding to rule *R3* and $P^* = 0.99$ becomes $\hat{\delta} = 2.7693$.

For the situation of unknown variance, the selection constant δ can be found in Bechhofer & Dunnett (1988) for $P^* = 0.80, 0.90, 0.95$ and 0.99. The number of error degrees of freedom is 10. The results of numerical integration and simulation were also compared for this situation. The results are written in Table 4.7.

Table 4.7. Minimum probability of correct selection (P^*), selection constant (δ), **(a)** estimated P^* ($\hat{P}^*$), 95% confidence interval of P^* (*95%CI*) and 95% lower limit of P^* (*95%LL*), **(b)** estimated selection constant ($\hat{\delta}$), estimated upper limit of the selection constant ($\hat{\delta}_u$); calculated for selection rules *R1* and *R3* for a completely randomised design with 5 varieties and 3 observations per variety, in case of unknown variance. Number of simulation rounds : 10000.

(a)

P^*	δ	$\hat{P}^*$		*95%CI*		*95%LL*	
		R1	*R3*	*R1*	*R3*	*R1*	*R3*
0.20	0.00000	0.2000	0.2001	0.2000 , 0.2000	0.1923 , 0.2079	0.2000	0.1935
0.80	1.54516	0.8001	0.7994	0.7957 , 0.8045	0.7916 , 0.8073	0.7964	0.7928
0.90	2.02419	0.9009	0.8980	0.8975 , 0.9042	0.8921 , 0.9039	0.8981	0.8930
0.95	2.46557	0.9508	0.9493	0.9484 , 0.9533	0.9450 , 0.9536	0.9488	0.9457
0.99	3.45291	0.9903	0.9912	0.9892 , 0.9913	0.9894 , 0.9930	0.9894	0.9897

(b)

P^*	δ	$\hat{\delta}$		$\hat{\delta}_u$	
		R1	*R3*	*R1*	*R3*
0.80	1.54516	1.5455	1.5496	1.5591	1.5752
0.90	2.02419	2.0175	2.0382	2.0178	2.0758
0.95	2.46557	2.4537	2.4732	2.4812	2.5152
0.99	3.45291	3.4298	3.3857	3.4992	3.4687

Table 4.7 shows that also for unknown variance the approximation by computer simulation is satisfactory for practical use. To be on the safe side w.r.t. the minimum probability of correct selection, the upper limit $\hat{\delta}_u$ must be used.

The second design that we will use for the comparison is a completely randomised design with 5 varieties with an unequal number of observations of the varieties. Varieties 1 and 2 have 2 observations, varieties 3 and 4 have 3 observations and variety 5 has 4 observations. Here the design is not variance balanced, and the four selection (of the best) rules are not equal to each other. We will make the comparison with the results of selection rule *R2* in case of unknown variance. The number of error degrees of freedom is 9. First the selection constants were calculated with the use of numerical integration, and the P^*'s corresponding to these selection constants were estimated by simulation. Also the selection constants themselves were estimated by means of simulation. The results are presented in Table 4.8. Notice that $\Delta_1 = \Delta_2$ and $\Delta_3 = \Delta_4$. In this situation we cannot use the method with estimated expected subset size.

Table 4.8. Minimum probability of correct selection (P^*), selection constants (Δ_i), estimated P^*, given that i is the best ($\hat{P}_i^*$), 95% confidence interval of P^*, given that i is the best ($95\%CI_i$), 95% lower limit of P^*, given that i is the best ($95\%LL_i$), estimated selection constant ($\hat{\Delta}_i$) and estimated upper limit of the selection constant ($\hat{\Delta}_{i,u}$); calculated for selection rule $R2$, for an unbalanced completely randomised design with 5 varieties, in case of unknown variance. Number of simulation rounds : 10000.

P^*	Δ_1	$\hat{P}_1^*$	$95\%CI_1$	$95\%LL_1$	$\hat{\Delta}_1$	$\hat{\Delta}_{1,u}$
0.80	1.39422	0.8011	0.7933 , 0.8089	0.7945	1.3883	1.4109
0.90	1.85738	0.9016	0.8958 , 0.9074	0.8967	1.8468	1.8833
0.95	2.28929	0.9519	0.9477 , 0.9561	0.9484	2.2737	2.3182
0.99	3.28609	0.9899	0.9879 , 0.9919	0.9883	3.3000	3.4126
P^*	Δ_2	$\hat{P}_2^*$	$95\%CI_2$	$95\%LL_2$	$\hat{\Delta}_2$	$\hat{\Delta}_{2,u}$
0.80	1.39422	0.7921	0.7841 , 0.8001	0.7854	1.4207	1.4438
0.90	1.85738	0.8991	0.8932 , 0.9050	0.8941	1.8630	1.8931
0.95	2.28929	0.9525	0.9483 , 0.9567	0.9490	2.2544	2.3064
0.99	3.28609	0.9909	0.9890 , 0.9928	0.9894	3.2500	3.3376
P^*	Δ_3	$\hat{P}_3^*$	$95\%CI_3$	$95\%LL_3$	$\hat{\Delta}_3$	$\hat{\Delta}_{3,u}$
0.80	1.35089	0.8033	0.7955 , 0.8111	0.7968	1.3396	1.3619
0.90	1.77503	0.9045	0.8987 , 0.9103	0.8997	1.7526	1.7789
0.95	2.17343	0.9511	0.9469 , 0.9553	0.9476	2.1589	2.1952
0.99	3.09913	0.9899	0.9879 , 0.9919	0.9883	3.1069	3.2126
P^*	Δ_4	$\hat{P}_4^*$	$95\%CI_4$	$95\%LL_4$	$\hat{\Delta}_4$	$\hat{\Delta}_{4,u}$
0.80	1.35089	0.8027	0.7949 , 0.8105	0.7963	1.3410	1.3624
0.90	1.77503	0.9036	0.8978 , 0.9094	0.8987	1.7597	1.7841
0.95	2.17343	0.9512	0.9470 , 0.9554	0.9477	2.1605	2.2029
0.99	3.09913	0.9909	0.9890 , 0.9930	0.9893	3.0500	3.1645
P^*	Δ_5	$\hat{P}_5^*$	$95\%CI_5$	$95\%LL_5$	$\hat{\Delta}_5$	$\hat{\Delta}_{5,u}$
0.80	1.32306	0.8062	0.7985 , 0.8139	0.7997	1.3055	1.3241
0.90	1.72217	0.9024	0.8966 , 0.9082	0.8975	1.7098	1.7360
0.95	2.09827	0.9501	0.9458 , 0.9544	0.9465	2.0989	2.1298
0.99	2.97505	0.9903	0.9884 , 0.9922	0.9887	2.9560	3.0447

The last design that we will use for comparison is a balanced 4x4 lattice design. Such a design has 5 replications, 4 incomplete blocks per replication and 4 varieties per block, with a total of 16 varieties. Because this design is variance balanced, the selection (of the best) rules are identical. We will use selection rule *R1* and *R3* for comparison. As mentioned in *4.1.1*, for a variance-balanced design we can use the selection constants as calculated for the completely randomised design. The selection constants $\delta_i = \delta$ can be found in Bechhofer & Dunnett (1988). The comparison was made for the situation of unknown variance. The number of degrees of freedom for error is 45. The results are presented in Table 4.9.

Table 4.9. Minimum probability of correct selection (P^*), selection constant (δ), (a) estimated P^* ($\hat{P}^*$), 95% confidence interval of P^* (*95%CI*) and 95% lower limit of P^* (*95%LL*), (b) estimated selection constant ($\hat{\delta}$) and estimated upper limit of the selection constant ($\hat{\delta}_u$); calculated for selection rules *R1* and *R3*, for a balanced 4x4 lattice design, in case of unknown variance. Number of simulation rounds : 10000.

(a)

P^*	δ	$\hat{P}^*$		*95%CI*		*95%LL*	
		R1	*R3*	*R1*	*R3*	*R1*	*R3*
0.0625	0.00000	0.0625	0.0668	0.0625 , 0.0625	0.0619 , 0.0717	0.0625	0.0627
0.80	1.93524	0.8014	0.8062	0.7979 , 0.8048	0.7985 , 0.8139	0.7985	0.7997
0.90	2.32298	0.9013	0.9029	0.8988 , 0.9038	0.8971 , 0.9087	0.8992	0.8980
0.95	2.65515	0.9503	0.9508	0.9485 , 0.9520	0.9466 , 0.9550	0.9488	0.9472
0.99	3.31401	0.9897	0.9906	0.9890 , 0.9905	0.9887 , 0.9925	0.9891	0.9890

(b)

P^*	δ	$\hat{\delta}$		$\hat{\delta}_u$	
		R1	*R3*	*R1*	*R3*
0.80	1.93524	1.9307	1.9147	1.9399	1.9383
0.90	2.32298	2.3174	2.3078	2.3277	2.3338
0.95	2.65515	2.6518	2.6486	2.6653	2.6805
0.99	3.31401	3.3211	3.2859	3.3427	3.3922

The comparisons with the results of numerical integration indicate that the method of simulation approximates the minimum probability of correct selection accurately enough. The width of the approximate confidence interval of P^* varies with the number of varieties, $\hat{P}^*$ and the confidence level and also depends on the knowledge about the error variance. Normally the width of the confidence interval is approximated as $2 \times 1.96\sqrt{\hat{P}^*(1-\hat{P}^*)/m}$. The desired width of the confidence interval dictates the number of simulation rounds to be used. For $P^* = 0.90$, the width of the confidence interval is approximately 0.0118 if $m = 10000$ simulation rounds are used. However, the method that makes use of the estimated expected subset size gives a much smaller width of the confidence interval. So, if the selection constants are known to be identical, this method should be preferred.

In order to make statistical selection a tool that can be actually used in practice, it is very important to be able to calculate the selection constants corresponding to a selection rule. It is not necessary that the selection constants are accurate to the fifth decimal, because often the experimental data are also not that accurate. Furthermore, it is most probably irrelevant to a research worker whether $P^* = 0.800$ or $P^* = 0.804$. Therefore the simulation method gives selection constants which are accurate enough to work with in practice. Also we can be somewhat conservative by using the selection constants that correspond to the 95% lower limit of P^*. The availability of selection constants for any experimental design makes it possible to really use statistical selection in practice.

4.3.3 Probability of correct selection and expected subset size

In the previous section the aim was to calculate selection constants, given a particular experimental design and minimum probability of correct selection. With these selection constants we are able to execute the selection rules. To compare the various selection rules, we could be interested in the real probability of correct selection or the expected subset size. In this section we will use simulation methods to calculate these statistics, given the experimental design and the selection constants.

To calculate the probability of correct selection and the expected subset size we have to assume that we know the randomisation procedure of the (randomised) experiment. We assume that the original varieties are assigned completely at random to the design varieties. Hence $P_r(\eta) = 1/t!$ for all $\eta \in \mathrm{H}$. Then the probability of correct selection ($P_R(CS)$) can be written as

$$P_R(CS) = \sum_{\eta \in \mathrm{H}} P(CS \mid \eta) P_r(\eta)$$

$$= \frac{1}{t!} \sum_{\eta \in \mathrm{H}} P(CS \mid \eta).$$

Further we have to know the real values of the differences between the variety parameter of the best variety and the other variety parameters. For selection rule *R1*, in case of known variance, the probability of correct selection can then be written as

$$P_R(CS) = \frac{1}{t!} \sum_{\eta \in \mathrm{H}} P(\hat{\tau}_{\eta((j))} - \hat{\tau}_{\eta((t))} - (\tau_{\eta((j))} - \tau_{\eta((t))}) \leq \delta_{\eta((t))} \nu_{\eta((t))\eta((j))} \sigma + \tau_{(t)} - \tau_{(j)}, \forall j \neq t).$$

This formula and the formulae for the other selection (of the best) rules are developed further in Dourleijn & Driessen (1991). The same problems that arose for calculating P^* arise here for the exact calculation of a single $P(CS \mid \eta)$, using numerical integration. In addition to that $t!$ increases rapidly for increasing t. Therefore we often have to use simulation to approximate the probability of correct selection.

The expected subset size ($E_R(|S|)$) can be written as

$$E_R(|S|) = \sum_{i=1}^{t} P_R(\text{variety } i \text{ selected})$$

$$= \frac{1}{t!} \sum_{i=1}^{t} \sum_{\eta \in H} P(\text{variety } i \text{ selected} \mid \eta).$$

To calculate this expectation, we have to know the true values of the differences between the parameter of variety i and the parameters of the other varieties. For selection rule $R1$, in case of known variance, we can write the expected subset size as

$$E_R(|S|) = \frac{1}{t!} \sum_{i=1}^{t} \sum_{\eta \in H} P(\hat{\tau}_{\eta(j)} - \hat{\tau}_{\eta(i)} - (\tau_{\eta(j)} - \tau_{\eta(i)}) \le \delta_{\eta(i)} v_{\eta(i)\eta(j)} \sigma + \tau_i - \tau_j, \ \forall\, j \ne i).$$

In Dourleijn & Driessen (1991) this formula is further developed, together with those for the other selection (of the best) rules. The exact calculation (by numerical integration) of the expected subset size gives the same problems as calculating the probability of correct selection. Therefore we often have to use simulation.

To approximate the probability of correct selection and the expected subset size, a modification of the Fortran 77 computer program SELCON described in section *4.3.1* was written. For a given selection rule (rule $R1$, $R2$, $R3$ or $R4$) and experimental design we are asked to give the selection constants that correspond with one or more P^* levels. Then we have to give the real values of the differences between the parameter of the best variety and the $t-1$ other variety parameters. Without loss of generality we can assume that $\tau_{(t)} = 0$, and in that case $\tau_{(t-i)} - \tau_{(t)} = \tau_{(t-i)}$, $i = 1, 2, \ldots, t-1$. The true values of the differences are expressed relative to σ, which (without loss of generality) is assumed to be 1. Of the utmost importance is the assignment of the actual varieties to the design varieties. Here we assume that this assignment is completely at random. Then the assignment of the ranked actual varieties to the design varieties is also completely at random.

In section *4.3.1* we described how to create a solution of the reduced normal equations with the choice that $\hat{\tau}_1 = 0$, in case of the Least Favourable Configuration (LFC). In that situation the variety parameters are assumed to be equal, hence the realisations from the $t-1$ dimensional multivariate normal distribution can be taken as realisations of noise only. But in this section we are not dealing with the LFC. Therefore, the expectations of the estimators $\mathbf{A}\hat{\tau}$ should be added to the noise realisations. However, we do not know these expectations, because τ was defined as the vector of design variety parameters. But we do know the randomisation process. Therefore a solution of the reduced normal equations is created as follows: The ranked original varieties (i) are completely at random assigned to the design varieties j, with $i, j \in \{1, 2, \ldots, t\}$. After that we know which design variety represents the best variety. Further we assume that $\hat{\tau}_1 = \tau_1 = \tau_{\eta((i))}$, i being determined by the randomisation procedure. The realisations of the other $\hat{\tau}_i$ $(i = 2, 3, \ldots, t)$ are calculated by adding the values of τ_i to the noise realisations, calculated as in *4.3.1*. The values of τ_i are determined by the randomisation process. This procedure of creating the t estimates of τ_i is repeated every simulation round, with in each round a new randomisation.

In each simulation round we create a new set of realisations of the estimated variety parameters. In case we have the situation of unknown variance, a realisation of s is generated from the $\sqrt{\chi^2_\nu/\nu}$ distribution, as in section *4.3.1*. Then the chosen selection rule is executed, and the subset size is determined. Also we check whether we have a correct selection or not. This simulation is repeated very often, e.g. 10000 times. Then the probability of correct selection is estimated as the number of successful selections divided by the total number of selections. This is the same method as we used for estimating P^* in section *4.3.1*. Also the 95% confidence interval and the 95% lower limit of $P_R(CS)$ can be approximated as described in that section. The approximation of the expected subset size is also mentioned in *4.3.1*. In that section the estimated variance of the subset size is given. With this estimated variance we can approximate the 95% confidence interval of the subset size :

$$[\overline{|S|} - t_{(m-1),0.975}\sqrt{v\hat{a}r(|S|)} \quad , \quad \overline{|S|} + t_{(m-1),0.975}\sqrt{v\hat{a}r(|S|)}].$$

We will compare the results of simulation with the results of numerical integration for only one experimental design. This design is a completely randomised one with three varieties. It is very unbalanced because one of the design varieties has 100 observations and the other two only 2. We assume that the variance is known, and we will compare the two methods for selection rule *R1*. We will use the selection constants that correspond to a minimum probability of correct selection of 0.90. We used 8 different configurations of the ranked variety parameters, namely 4 so-called slippage configurations and 4 equidistant configurations. A slippage configuration with distance parameter q satisfies $\tau_{(1)} = \tau_{(2)} = \ldots = \tau_{(t-1)} = \tau_{(t)} - q$. We will denote such a configuration by SL(q). An equidistant configuration with distance parameter q satisfies $\tau_{(1)} = \tau_{(2)} - q = \tau_{(3)} - 2q = \ldots = \tau_{(t)} - (t-1)q$. Such a configuration is denoted by EQ(q). The results of the comparison are presented in Table 4.10.

Table 4.10. Probability of correct selection ($P_R(CS)$) and expected subset size ($E_R(|S|)$) for selection rule *R1* in case of known variance, calculated by numerical integration. Approximate probability of correct selection ($\hat{P}_R(CS)$) and approximate expected subset size ($\hat{E}_R(|S|)$), the corresponding 95% confidence intervals (95%*CI*) and the corresponding 95% lower limits (95%*LL*), calculated by simulation. Completely randomised design, $t = 3$, $n_1 = 100$ and $n_2 = n_3 = 2$. Results were obtained for various configurations of the ranked variety parameters. $P^* = 0.90$, 10000 simulation rounds.

	$P_R(CS)$	$\hat{P}_R(CS)$	95%*CI*	95%*LL*
SL(1)	0.9946	0.9949	0.9935 , 0.9963	0.9937
SL(2)	0.9999	0.9999	0.9997 , 1.0000	0.9997
SL(3)	1.0000	1.0000	1.0000 , 1.0000	1.0000
SL(4)	1.0000	1.0000	1.0000 , 1.0000	1.0000
EQ(0.25)	0.9599	0.9598	0.9559 , 0.9637	0.9566
EQ(0.50)	0.9835	0.9831	0.9806 , 0.9856	0.9810
EQ(0.75)	0.9930	0.9937	0.9921 , 0.9953	0.9924
EQ(1.00)	0.9971	0.9971	0.9960 , 0.9982	0.9962
	$E_R(\|S\|)$	$\hat{E}_R(\|S\|)$	95%*CI*	95%*LL*
SL(1)	2.1968	2.1957	2.1500 , 2.2414	2.1574
SL(2)	1.3568	1.3471	1.3183 , 1.3759	1.3229
SL(3)	1.0529	1.0528	1.0317 , 1.0739	1.0351
SL(4)	1.0045	1.0048	1.0000 , 1.0245	1.0000
EQ(0.25)	2.6071	2.6078	2.5553 , 2.6603	2.5637
EQ(0.50)	2.3579	2.3547	2.3065 , 2.4029	2.3143
EQ(0.75)	2.0435	2.0373	1.9950 , 2.0796	2.0018
EQ(1.00)	1.7624	1.7595	1.7227 , 1.7963	1.7286

The results calculated by numerical integration are in Table 4.10 always included in the 95% confidence interval. If we want more accurate approximations, we have to increase the number of simulation rounds.

4.4 Subset selection in (combined) variety trials

In the preceding sections of this chapter, it has become clear that statistical subset selection can be a useful tool in the plant breeding practice. However, until now we have used subset selection in the situation of a single experiment with a completely randomised design or an (in)complete block design, and a fixed additive model for the observations. In chapter **3** we have seen that also mixed models play an important role in plant breeding. In addition to that the plant breeder's paramount interest lies in selection on the basis of estimates of variety values in which the information of several sites and/or years is assimilated. In *4.4.1* we will discuss the use of subset selection for all the situations and models mentioned in chapter **3**.

Until now we have assumed that there is only one quantitative character that determines the value of a variety. This character can also be composed of a number of different characters, and is then called a 'selection index'. Some remarks about selection on the basis of multiple characters are given in *4.4.2*.

Finally, *4.4.3* deals with the situation where a number of varieties have to be excluded *a priori* from selection. This can occur when the breeder has decided that specific varieties cannot be accepted, on different grounds than the studied selection character.

4.4.1 Selection based on combined estimates

In the plant breeding practice, variety trials are performed at several sites and sometimes also in two or more years. These variety trials are first analysed separately. Sometimes a trial is even further subdivided into subtrials, and then the subtrials are first analysed, followed by combining the subtrials of one trial. As shown in chapter **3**, local BLUEs of contrasts between variety values can be combined into the BLUEs corresponding to a model for the joint observations of all trials. The ultimate selection will be made on the basis of these BLUEs. Nevertheless, it can also be worthwile to apply subset selection to the results of the separate trials. If the subsets at different sites contain more or less the same varieties, there is little variety × site interaction and it will be safe to select the same varieties for the whole region. On the other hand, if the various subsets have completely different contents, then it will be hazardous to select varieties for the whole region. It then may be reasonable to divide the region into subregions.

For separate trials with observations that can be described by a fixed additive model, subset selection rules have been given in **4.1**, **4.2** and **4.3**. Section **3.3** dealt with the situation where these separate trials are subdivided into subtrials that are only connected by control varieties. There, it was shown that the BLUEs corresponding to a model for the joint observations of the subtrials can be obtained by combination of local estimators from the subtrials. Together with the estimates of the variety parameters we need the value(s) of the selection constant(s) to be able to execute a selection rule. It has been shown in *4.3.1* that with the use of computer simulation the selection constants can be approximated accurately enough to be used for practical application purposes. To approximate the selection constant(s) by simulation, the pseudo-variance/covariance matrix of the estimators of the variety parameters has to be available.

In chapter **3** we discussed several models for the joint observations of a series of experiments. For each model it was shown how to calculate the least squares estimates of contrasts between variety parameters or variety values (for models with fixed interaction terms) with the local BLUEs from the separate trials. Furthermore, the (pseudo-)variance/covariance matrices of the estimators were given for each model. Using these (pseudo-)variance/covariance matrices (divided by σ^2) in the computer program SELCON that approximates the selection constants by simulation, we can determine the correct selection constants. Consequently, we can use the described selection rules for all the models mentioned in chapter **3**. These also include mixed models. However, these are mixed models for which it is assumed that all variance ratios (e.g. σ_L^2/σ^2) are known. If the variance ratios are in fact estimates, we could determine the subset using a lower limit, an upper limit and a few values in between these limits of the variance ratios. If the resulting subsets are quite different, we have to be cautious. At this moment there are no selection rules available that are specifically designed for the situation of several unknown variance components. For that situation it is difficult to determine the selection constants analytically. However, using computer simulation this would indeed be possible.

With most models we can select on the basis of BLUEs of (contrasts between) the variety parameters, or, if preferred, on the BLUEs of the variety values. With the variance/covariance matrix of those estimators the values of v_{ij}, which are included in certain selection rules, can be calculated. If the joint observations of the trials at different sites are described by an interaction model with fixed interaction terms, then we have to base the selection on the variety values.

4.4.2 Selection on the basis of multiple characters

Until now we have assumed that there is only one character on the basis of which the selection is made. However, in the plant breeding practice often numerous characters are taken into consideration while selecting varieties. In chapter 2 it is described that a sugar beet breeder selects his varieties mainly on the observed characters corrected root yield (CRY), sugar content (SC) and the number of bolting plants (BOL), and on the derived characters corrected sugar yield (CSY), white sugar yield (WSY) and white sugar content (WSC). When two or more characters are involved, the definition of the 'best variety' is less self-evident. This definition must be given by the plant breeder, but it appears that it is often very difficult to express the thoughts behind their breeder's eyes in a certain index variable (selection index) that is a function of the observed characters. However, if they manage to do so, the various observed characters can be reduced to a single selection index, and the theory of the previous sections can be used without adjustment. The selection index should be calculated at every plot, after which (contrasts between) variety values for this index can be estimated as described in chapter 3. The derived characters, like WSY, are also selection indices. In chapter 2 we introduced the financial yield (FIN) as a suitable selection index to base the selection on.

The above mentioned procedure is different from a procedure where first the variety values are estimated for each observed character, and next the selection index is calculated with these estimates. The difference is caused by correlations between the observed characters. The correlations can be eliminated and variances standardised by calculating the so-called Mahalanobis distance of each variety. We will not further describe this or any other multivariate method for selection, where it is assumed that the various characters together have a multivariate distribution with a particular (known or estimated) dispersion matrix.

Besides selecting a subset based on a single index, it is likely that breeders also will determine the subsets to be selected for other characters like CRY, WSY, WSC and so on. Let there be k characters. The question now arises whether it is possible to combine these subsets meaningfully. First suppose we take the intersection of the various subsets. This may lead to the problem that this intersection is empty. For instance, since there exists a negative correlation between root yield and sugar content, the corresponding subsets will probably contain different varieties and therefore it is possible that the intersection of these two subsets is empty. Furthermore, after having taken the intersection of subsets that include the best variety with a certain confidence, it is not possible anymore to give a probability statement. This can be seen as follows. With probability of at least P_1^*, say, it can be stated that the best variety for character 1 is included in subset 1, with probability of at least P_2^* the best variety for character 2 is included in subset 2, ..., and with probability P_k^* the best variety for character k is included in subset k. If we now take the intersection of all subsets, it is not possible to give a statement about the probability that the best variety is included. For what is the best variety ? Also, we now cannot state that the best variety for character 1 is included in the intersection with probability at least P_1^*.

Now let subset i be selected in such a way that with probability at least P_i^* all varieties included are good with respect to character i $(i = 1, \ldots, k)$, with 'good' properly defined by a distance measure δ_i^*. The corresponding selection procedure is described in *4.1.1*. Suppose we take the *intersection* of these subsets. We will calculate the minimum probability that the intersection contains varieties that are good with respect to every character. Let the event that all varieties in subset i are good be denoted by E_i. So $P(E_i) \geq P_i^*$ and $P(E_i^c) \leq 1 - P_i^*$, with the superscript c denoting the complement. Boole's inequality says that

$$P\left(\bigcup_{i=1}^{k} E_i^c\right) \leq \sum_{i=1}^{k} P(E_i^c),$$

so the following probability statement can be made :

$$\begin{aligned} P\left(\bigcap_{i=1}^{k} E_i\right) &= 1 - P\left(\bigcup_{i=1}^{k} E_i^c\right) \\ &\geq 1 - \sum_{i=1}^{k} P(E_i^c) \\ &\geq 1 - k + \sum_{i=1}^{k} P_i^* . \end{aligned}$$

For P_i^*'s close to 1, $1-k+\Sigma P_i^*$ is approximately equal to $\prod P_i^*$, which is the correct expression for $P(\cap E_i)$ if the various characters act independently. The equality of these two probabilities also depends on the number of characters k used. Again it is possible that the intersection is empty, but this depends on the definition of 'good'. If the distance measure δ^* is chosen large, so that it would be more appropriate to speak about 'not bad varieties' instead of 'good varieties', then the intersection probably will not be empty.

A selection rule that could be useful for the above approach is a rule defined by Desu (1970). There, the goal is to select a non-empty random sized subset that does not include bad varieties, with probability at least P^*. Variety j is called 'bad' if $\tau_{(t)} - \tau_j > \delta^*$, and δ^* is given by the experimenter. The selection rule is defined for observations from an equi-replicated completely randomised design and known variance σ^2. The rule reads :
Select variety j $(j = 1, \ldots, t)$ if and only if

$$\overline{Y}_j \geq \overline{Y}_{[t]} + \frac{\gamma\sigma}{\sqrt{n}} - \delta^* ,$$

with n the number of observations per variety and γ the same selection constant as in (4.2). The number n has to be chosen large enough such that $(\gamma\sigma)/\sqrt{n} \leq \delta^*$. If subsets are determined for each character, with minimum probability of correct selection P_i^* $(i = 1, \ldots, k)$, then the intersection of these subsets contain varieties that are not bad for every character, with probability at least $1-k+\Sigma P_i^*$.

If we take the *union* of the selection-of-the-best subsets corresponding to the various characters, we can state that the best variety for character 1, as well as the best variety for character 2, ..., as well as the best variety for character k, are included in the union with probability at least $1-k+\Sigma P_i^*$. But the number of varieties present in the union of the subsets is probably very large and it is not inconceivable that all varieties are included. Then this method is not useful in practice. This method can be used if the sizes of the separate subsets are small, for instance if the subsets are selected with the aim to include at least one good variety.

4.4.3 Excluding varieties from selection

Consider the situation where the plant breeder wants to select a subset out of $p < t$ varieties, so $t-p$ varieties are *a priori* excluded from selection. For example, suppose we have an experiment with p new varieties and $t-p$ control varieties.

If we want to select a subset including the best new variety, we do not want a control variety to be present in the subset. We cannot simply remove the control varieties from the selected subset (if they are selected) and still make the inference that the best variety is present in the remaining subset with a certain confidence. In this example we want to exclude certain varieties from selection because they are control varieties. But it is also possible that we want to exclude certain varieties from selection because they have characters for which they will be discarded. These characters, different from the character on which we base the subset selection, can be observed in the same trial or in other trials. We distinguish two situations :

(a) Varieties are excluded on the basis of characters that do not influence the current selection character. We will give an example of this situation. In sugar beet breeding bolting resistance is tested in separate trials. These trials are sown very early on a relatively cold site. Varieties that have no bolting resistance at all have to be discarded. But the results of these trials become available only after the yield trials have been sown. To select a subset including the variety with the largest true yield, we want to exclude the varieties that are not resistant to bolting. In the yield trials, which are sown later, only few plants will bolt. We can asssume that this character has no influence on the yield. Hence we can use all observations to estimate the variety values and the error variance. A second example is the situation of control varieties, described above.

The selection rule has to be executed with the p not excluded varieties. The variety values or parameters and the error variance are estimated using the observations of all t varieties, and the standard errors of the contrast estimators are calculated using the complete design with t varieties. To calculate the selection constant first the pseudo-variance/covariance matrix (divided by σ^2) of the estimators of the variety parameters is calculated using the complete design. Next the rows and columns corresponding to the excluded varieties are deleted and the remaining $p \times p$ matrix is the pseudo-variance/covariance matrix of the remaining p estimators. This matrix is then used to calculate the selection constant. The possibility to exclude certain varieties is implemented in the selection constant simulation program SELCON, described in *4.3.1*.

(b) Varieties are excluded on the basis of characters that do influence the current selection character. Consider the following example. In an experiment t varieties are grown in an experiment with a block design. However, during the growing season some varieties practically die because of some disease. Because these varieties are very susceptible to this disease, their yield observations are very much influenced. All these observations will be close to zero. The variances of these observations will not be equal to those of the observations of the remaining p varieties, which are resistant to the disease. Hence for these observations models with equal error variances will be wrong. In an additive model an observation is the sum of a general level, a variety term, a block term and an error term. We make the assumption that there is no variety × block interaction, meaning that the differences between varieties are the same in the various blocks; hence an increase in the fertility of a block leads towards larger observations of all the varieties in that block. When there is some interaction, this term is confounded with the error term if we use an additive model. In the above example the susceptibility of the varieties to the disease will give rise to variety × block interaction. This term will be added to the error term in an additive model and therefore the residuals will become large. At the examination of the residuals these observations will be considered to be outliers.

Outliers are often removed from the data set. However, if we delete the observations of the varieties that we want to exclude, the experimental design becomes more unbalanced and perhaps disconnected. The standard errors of the contrast estimators will increase, which results in a larger selected subset size. Therefore, we suggest the following approach. Analyse the complete design with t varieties and examine the residuals. If there is no reason to delete certain observations from the data set, we can proceed as in (a). However, if some observations absolutely disagree with the additive model we have to delete them. Then the variety parameters and the error variance are estimated with the use of the remaining observations and also the pseudo-variance/covariance matrix of $\hat{\tau}$ is calculated using the altered experimental design. After deletion (if still necessary) of the rows and columns that correspond to the $t-p$ excluded varieties, the selection constant(s) for p varieties can be calculated with the use of this variance/covariance matrix. Next the selection rule can be executed using this (these) selection constant(s), the estimated variety parameters and standard errors of the contrast estimators.

4.5 Modifications of subset selection procedures

With the proposed estimation procedures and selection rules statistical subset selection can be used in the plant breeding practice. However, the selected subsets are often disappointingly large. This is not only due to large standard errors of estimators, but is also caused by the stringent probability requirement. The probability of correct selection has to be larger than or equal to P^*, for *every* configuration of the variety parameters. For that reason the selection constants are calculated for the Least Favourable Configuration (LFC) of variety parameters, because in that case the probability of correct selection is equal to P^*. For subset selection, the LFC is a configuration where all variety parameters are equal. This configuration is not realistic from the practical point of view. The real configuration will be different from the LFC and therefore the probability of correct selection will be larger than P^*. If a breeder wants to select varieties with the probability of correct selection approximately equal to P^*, then he will frequently select too many varieties. The number of selected varieties can be reduced, however at some expenses. In *4.5.1* an approach is proposed where we have to make the extra assumption that the variety parameters are a random sample from a Normal population. In *4.5.2* an approach is proposed where information about the ranked variety parameter contrasts is used to select the ultimate subset.

4.5.1 Assuming a superpopulation of variety parameters

The probability of correct selection depends on the selection rule, the selection constant, the experimental design, the randomisation procedure and the configuration of the variety parameters. The true configuration of the variety parameters is of course never known, hence the probability of correct selection cannot be calculated. The absolute minimum of this unconditional probability is the probability of correct selection calculated for the LFC of the variety parameters. This P^* is a special case of a conditional probability of correct selection, i.e. the probability of correct selection given a particular configuration. The LFC is purely theoretical in the plant breeding practice, and therefore P^* is not very relevant to the plant breeder. He is more interested in a probability of correct selection that is closer to reality.

In some situations we can make the assumption that the variety parameters constitute a random sample from a population. This population of all possible variety parameters is denoted by 'superpopulation', because every variety itself also represents a population. For example, if two potato varieties are crossed, it is generally assumed that the variety parameters for yield of the genotypes in the offspring of this cross follow a Normal distribution. This because yield is a polygenic character and can be thought of as being the result of many other characters. We now assume that the variety parameters are a random sample from a Normal distribution with variance σ_g^2 (the genetic variance). Without loss of generality the expectation can be assumed to be zero. The superpopulation assumption is reasonable in the first selection stage, where no previous selection has occurred. The plant breeder, who is working for years with the same crop, has a fairly good idea about the ratio σ_g^2/σ_e^2, with σ_e^2 the error variance. This ratio determines the heritability $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$, which is the proportion of the total variance that is attributable to the effects of genes.

For the analysis of the experiment the variety parameters are considered fixed, because we are interested in those specific parameters and not in the population. For a given configuration of variety parameters, drawn from a $N(0,\sigma_g^2)$ distribution, we can calculate the conditional probability of correct selection, using a particular selection rule (see *4.3.3*). However, using the superpopulation assumption, we are more interested in the unconditional probability of correct selection. The aim can be to control the expectation of this probability, denoted by $E[P(CS)]$. With the use of computer simulation we can estimate this $E[P(CS)]$ for a given selection rule. The expectation and the variance of the unconditional probability of correct selection can be estimated as

$$\begin{aligned} &\hat{E}[P(CS)] = \hat{E}[\hat{P}(CS) \mid conf.]\,, \\ &\hat{\mathrm{var}}[P(CS)] = \hat{E}\{\hat{\mathrm{var}}[\hat{P}(CS) \mid conf.]\} + \hat{\mathrm{var}}[\hat{P}(CS) \mid conf.]\,, \end{aligned} \tag{4.36}$$

where $[\hat{P}(CS) \mid conf.]$ denotes the estimate of the conditional probability of correct selection, given the configuration according to the sample drawn from the superpopulation. The variance can be estimated by running a simulation program a number of times, each time estimating the conditional probability of correct selection for a different configuration (but with all configurations being samples from a Normal superpopulation) and estimating the variance of the estimator of the conditional probability. The average of the variance estimates gives the first right hand term in (4.36) and the estimated variance of the estimated probabilities gives the second right hand term in (4.36). In Table 4.11 the values of $\hat{E}[P(CS)] - 1.645\sqrt{\hat{\text{var}}[P(CS)]}$, which is the approximate 95% confidence lower limit of the unconditional probability of correct selection, are given for the situation of a completely randomised design with $t = 5$ or $t = 25$ varieties. Table 4.11 shows that the variance of the unconditional probability of correct selection decreases with increasing t and/or $\hat{E}[P(CS)]$ and decreasing h^2. If we choose $\hat{E}[P(CS)] = 0.95$, the probability of correct selection will only occasionally be smaller than 0.90.

Table 4.11. Approximate 95% confidence lower limits of the unconditional probability of correct selection, for various values of the estimated expected unconditional probability of correct selection ($\hat{E}[P(CS)]$), heritability (h^2) and number of varieties (t) in a completely randomised design.

h^2	t	$\hat{E}[P(CS)]=0.80$	$\hat{E}[P(CS)]=0.90$	$\hat{E}[P(CS)]=0.95$	$\hat{E}[P(CS)]=0.99$
0.001	5	0.785	0.891	0.944	0.988
	25	0.788	0.892	0.945	0.988
0.01	5	0.757	0.875	0.934	0.987
	25	0.769	0.881	0.939	0.987
0.1	5	0.675	0.824	0.907	0.981
	25	0.721	0.850	0.924	0.984
0.2	5	0.628	0.797	0.892	0.977
	25	0.695	0.835	0.916	0.982
0.3	5	0.591	0.775	0.880	0.976
	25	0.673	0.822	0.907	0.981
0.4	5	0.558	0.756	0.869	0.974
	25	0.653	0.809	0.899	0.980
0.5	5	0.527	0.740	0.855	0.970
	25	0.629	0.796	0.889	0.979

The present selection rules are not designed to deal with two unknown variance components (σ_g^2 and σ_e^2). In order to calculate the selection constant we will therefore assume that the heritability is known. Sometimes the experienced plant breeder knows the magnitude of the heritability. Otherwise he can approximate a 99% lower bound for the heritability from previous experiments and use this lower bound as the known heritability.

The simulation program to calculate the selection constant(s), as described in *4.3.1*, was extended to estimate the expectation of the probability of correct selection in the case of a superpopulation assumption. In this simulation program the error variance is 1, hence the variance of the superpopulation is equal to $\sigma_g^2 = h^2/(1-h^2)$. The variety parameters are drawn from a Normal distribution with zero expectation and variance σ_g^2. New realisations (a random sample from the superpopulation) are generated every simulation round again. Knowing the realisations, we also know which variety is the best. After generating realisations of $\hat{\tau}_i - \hat{\tau}_1$ $(i = 2, \ldots, t)$, the selection rule can be executed and we can proceed as in the original simulation program. The expectation of the probability of correct selection is approximated, for a range of selection constants, as the number of correct selections divided by the number of simulation rounds. The selection constant corresponding to a particular $\hat{E}[P(CS)]$ can be found by log-linear interpolation. Suppose the selection-of-the-best subset rule is used with a selection constant corresponding to $\hat{E}[P(CS)] = 0.95$. Then, given the assumption that we have a Normal superpopulation of variety parameters with a particular heritability, we can state that the expectation of the probability that the best variety is included in the subset is 0.95.

The Normal superpopulation is often a good assumption if we are dealing with varieties in the first selection stage. For later selection stages the Normal superpopulation assumption is rather artificial, because selection has flawed the Normality of the population. We will demonstrate the aforesaid approach with a small example.

Example

Consider a completely randomised design with 100 varieties and 2 observations per variety. Suppose we make use of selection rule *R3* (see **4.2**). The variety parameters are assumed to be a random sample from a Normal distribution. We can approximate the expected probability of correct selection for a range of selection constants, using simulation techniques. This can be done for various heritability values. In Figure 4.1 the results are presented for $h^2 = 0, 0.01, 0.1, 0.3$ and 0.5. For $h^2 = 0$ the results are equal to P^*, because with $h^2 = 0$ there is no genetic variation and all variety parameters will be equal (i.e. the LFC).

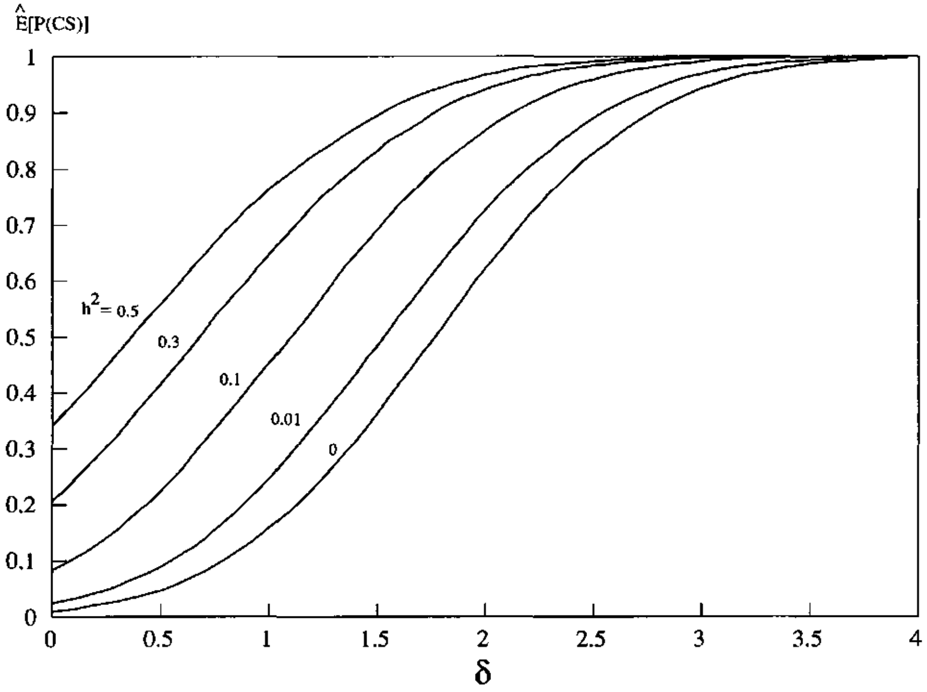

Figure 4.1. The estimated expectation of the probability of correct selection ($\hat{E}[P(CS)]$) in relation to the selection constant (δ), calculated for a completely randomised design with 100 treatments and 2 observations per treatment. The results are calculated for different heritabilities (h^2) corresponding to the superpopulation assumption.

From Figure 4.1 we see that the selection constant can be reduced substantially, if we make use of the superpopulation approach. This results in a smaller selected subset. Although the probability statements are not as rigid as the statements corresponding with the conventional subset rules, the practical usefulness of the superpopulation approach is evident.

4.5.2 Using simultaneous lower bounds of ranked variety parameter contrasts

If we insert simultaneous confidence lower bounds of the ranked variety parameter contrasts $\tau_{(t)} - \tau_{(i)}$ $(i = 1, \ldots, t-1)$ in the expression of the probability of correct selection, we get a confidence lower bound for this probability (see *4.1.1*). For randomised experiments this probability can be approximated by simulation, as described in *4.3.3*. Lam (1989) derived lower confidence bounds for $\tau_{(t)} - \tau_{(i)}$ $(i = 1, \ldots, t-1)$, for experiments with an equi-replicated completely randomised design. Driessen (1991) extended these results for (in)complete block designs. If we want to calculate simultaneous $(1-\alpha) \times 100\%$ confidence lower bounds for $\tau_{(t)} - \tau_{(i)}$, we first have to calculate separate constants for each variety, which we have previously called 'confidence constants' and denoted by Δ_i^c. The confidence constants must be chosen in such a way that the following relation holds :

$$P(\hat{\tau}_j - \tau_j - \Delta_i^c v_{ij} s \le \hat{\tau}_i - \tau_i \le \hat{\tau}_j - \tau_j + \Delta_i^c v_{ij} s \quad \forall j \ne i) = 1 - \alpha$$

$$P\left(-\Delta_i^c \frac{s}{\sigma} \le \frac{\hat{\tau}_i - \hat{\tau}_j - (\tau_i - \tau_j)}{v_{ij}\sigma} \le \Delta_i^c \frac{s}{\sigma} \quad \forall j \ne i \right) = 1 - \alpha$$

$$P\left(-\Delta_i^c \mathbf{1}_{t-1} \le \frac{\mathbf{T}_i}{s/\sigma} \le \Delta_i^c \mathbf{1}_{t-1} \right) = 1 - \alpha, \tag{4.37}$$

with $\mathbf{T}_i = \left(\frac{\hat{\tau}_i - \hat{\tau}_1 - (\tau_i - \tau_1)}{v_{i1}\sigma}, \ldots, \frac{\hat{\tau}_i - \hat{\tau}_{i-1} - (\tau_i - \tau_{i-1})}{v_{i(i-1)}\sigma}, \frac{\hat{\tau}_i - \hat{\tau}_{i+1} - (\tau_i - \tau_{i+1})}{v_{i(i+1)}\sigma}, \ldots, \frac{\hat{\tau}_i - \hat{\tau}_t - (\tau_i - \tau_t)}{v_{it}\sigma} \right)'$.

$\mathbf{T}_i/(s/\sigma)$ has a standard $(t-1)$ multivariate Student distribution with a particular correlation matrix, depending on the experimental design. $\mathbf{T}_i/(s/\sigma)$ is symmetric around $\mathbf{0}_{(t-1)}$. If the experimental design is variance-balanced, then the confidence constants are equal to each other ($\Delta_i^c = \Delta^c$). To calculate the confidence constants we can use numerical integration or simulation. However, we also can use the results of the calculation of the selection constants for the selection rules, because the following relation holds :

$$1-\alpha = P\left(-\Delta_i^c \mathbf{1}_{t-1} \le \frac{\mathbf{T}_i}{s/\sigma} \le \Delta_i^c \mathbf{1}_{t-1}\right)$$

$$\ge P\left(\frac{\mathbf{T}_i}{s/\sigma} \le \Delta_i^c \mathbf{1}_{t-1}\right) - \left\{1 - P\left(\frac{\mathbf{T}_i}{s/\sigma} \ge -\Delta_i^c \mathbf{1}_{t-1}\right)\right\}$$

$$= P\left(\frac{\mathbf{T}_i}{s/\sigma} \le \Delta_i^c \mathbf{1}_{t-1}\right) + P\left(\frac{\mathbf{T}_i}{s/\sigma} \le \Delta_i^c \mathbf{1}_{t-1}\right) - 1 ,$$

hence

$$P\left(\frac{\mathbf{T}_i}{s/\sigma} \le \Delta_i^c \mathbf{1}_{t-1}\right) \le 1 - \frac{\alpha}{2} . \qquad (4.38)$$

The left hand side of (4.38), equated to P^*, gives the equations to calculate the selection constants for a subset selection rule with separate selection constants for each variety, with $\Delta_i^c = \delta_i$ (see *4.3.1*). If we use as confidence constants the selection constants calculated for $P^* = 1 - \alpha/2$, then the confidence of the lower bounds will at least be $(1-\alpha)$.

For two completely randomised designs with 4 and 21 varieties, respectively, the values of δ corresponding to $P^* = 0.90$ and 0.95 and those of Δ^c corresponding to $1 - \alpha = 2P^* - 1 = 0.80$ and 0.90 were tabulated in Table 4.12 (a), using the tables of Bechhofer & Dunnett (1988). This was done for the situation of 10 and infinite degrees of freedom for error. We see that for the higher confidence level the difference between the two constants is smaller than for the lower confidence level. This difference also gets smaller when the number of degrees of freedom for error becomes larger. Furthermore the number of varieties influences the difference between the two constants. For 10 degrees of freedom the larger number leads towards larger differences, for infinite degrees of freedom the larger number of varieties results in smaller differences. In general, however, we can say that the differences between the two constants are only small, especially if we are interested in a confidence level of 0.90 and higher. This can also be seen in Table 4.12 (b), where the true confidence level that would be obtained if we had used the selection constant corresponding to $P^* = 1 - \alpha/2$ as the confidence constant, is calculated. This is done for a completely randomised design with 4 varieties and 10 degrees of freedom for error. From this table we see that the true confidence level is only notably higher for relatively small values of $1-\alpha$. So in practice we can use the selection constant that corresponds to $P^* = 1-\alpha/2$ as confidence constant to calculate $(1-\alpha)\times 100\%$ confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ and $P(CS)$.

Table 4.12. (a) Selection constant (δ) and confidence constant (Δ^c) for two values of minimum probability of correct selection (P^*) and confidence level ($1-\alpha$), with $P^* = 1-\alpha/2$. The entries are calculated for two completely randomised designs with $t = 4$ and 21 varieties, respectively, and two levels for the error degrees of freedom (df). (b) True confidence level ($1-\alpha\,(\Delta^c = \delta)$) if the selection constant corresponding to $P^* = 1-\alpha/2$ is used as confidence constant; calculated for a completely randomised design with 4 varieties and 10 degrees of freedom for error.

(a)	$t=4$		$t=21$	
$P^*, 1-\alpha$	δ, Δ^c for 10 df	δ, Δ^c for ∞ df	δ, Δ^c for 10 df	δ, Δ^c for ∞ df
0.90, 0.80	1.89856, 1.89367	1.73352, 1.73306	2.65124, 2.63786	2.34699, 2.34689
0.95, 0.90	2.33756, 2.33534	2.06208, 2.06204	3.11715, 3.10965	2.64492, 2.64491

(b)	P^*	δ	$1-\alpha\,(\Delta^c = \delta)$
	0.80	1.41977	0.6078
	0.90	1.89856	0.8015
	0.95	2.33756	0.9004
	0.99	3.31424	0.9800

After the calculation of the confidence constants Δ_i^c, we can calculate the $(1-\alpha)\times 100\%$ confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ $(i \neq t)$. How this can be done is described in *4.1.1*.

The probability of correct selection is a function of the ranked variety parameter contrasts $\tau_{(t)} - \tau_{(i)}$. To calculate the minimum probability of correct selection these contrasts were set to zero. But with the lower bounds of these contrasts available, we can replace $\tau_{(t)} - \tau_{(i)}$ by their lower bounds. This lower bound configuration is closer to reality than the LFC, and therefore the probability of correct selection, calculated with this configuration, will be closer to the real probability of correct selection. This probability is a $(1-\alpha)\times 100\%$ confidence lower bound of the real probability of correct selection, and is denoted by $P(CS)_L$.

The lower bounds of $\tau_{(t)} - \tau_{(i)}$, as such forming the configuration of variety parameters, have to be expressed relatively to σ. However, often we assume that σ is unknown. We then can use a $Q \times 100\%$ confidence upper bound for σ, namely $s/\sqrt{\chi^2_{\nu,1-Q}/\nu}$, with ν the number of degrees of freedom for error. Then the total confidence level will become $1-\alpha+Q-1=Q-\alpha$. Then we can say that with confidence $(Q-\alpha)$ the probability of correct selection is at least $P(CS)_L$.

We now propose the following method of selection. First give a minimum probability of correct selection, say P_1^*. Given the experimental design and the selection rule used, we can determine the selection constant(s). Then the subset can be selected for which we can state that the probability of correct selection is at least P_1^*. Also, simultaneous $(1-\alpha)\times 100\%$ confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ $(i \neq t)$ and a $(Q-\alpha)\times 100\%$ confidence lower bound of the probability of correct selection can be calculated. If $P(CS)_L$ is (too) large, then we can select less varieties than indicated by the selection rule. This means that the initial value of the minimum probability of correct selection was too high.

Then we choose a different minimum probability of correct selection P_2^*, with $P_2^* < P_1^*$, calculate the selection constant(s) and determine the subset. The probability of correct selection now is at least P_2^*. Again, the $(Q-\alpha)\times 100\%$ confidence lower limit of the probability of correct selection can be calculated, using the new selection constants. Now the subset size and the lower bound for the probability of correct selection are smaller, as we wanted, but the minimum probability of correct selection is also smaller. The probability for the confidence statement $\{P(CS) \geq P(CS)_L\}$ is equal to $Q-\alpha$, and the probability that $\{P_2^* \leq P(CS) < P(CS)_L\}$ is equal to $1-Q+\alpha$, because P_2^* is a guaranteed minimum.

The above cycle is continued until a satisfactory combination of P_l^* and $P(CS)_L$ is found.

Example

Twenty-one sugar beet hybrids were grown in an experiment with a complete block design with 4 replications. The white sugar yield (WSY) is one of the observed characters. The least squares WSY variety values for variety 1 ,..., variety 21 are 10.43, 11.88, 12.31, 11.94, 10.16, 12.70, 10.72, 12.85, 9.42, 10.23, 11.88, 10.54, 11.93, 12.13, 9.60, 10.39, 12.49, 11.98, 11.90, 11.86, 9.87 ton/ha, respectively. The estimated error variance, based on 60 degrees of freedom, is

$s^2 = 0.27\,(\text{ton/ha})^2$, hence $s = 0.52$ ton/ha. From the 21 varieties a subset is selected. We want the best variety to be included in the selected subset, and we start with $P_1^* = 0.95$. The selection rule we will use is rule *R3*. The selection constant for the complete block design with 4 replications and 21 varieties is equal to $\delta = 2.72$. With $P_1^* = 0.95$, 12 varieties (varieties 8,6,17,3,14,18,4,13,19,11,2 and 20) have to be selected.

Additionally, simultaneous 90% confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ can be calculated. As confidence constant the selection constant that corresponds to $P_1^* = 0.95$ was used. So $\Delta^c = 2.72$. Then the lower bounds are : $L_1 = 1.44, L_2 = 1.26, L_3 = 0.70, L_4 = 0.63, L_5 = 0.47, L_6 = 0.43, L_7 = 0.32, L_8 = 0.14$ and $L_9 = \ldots = L_{20} = 0$. With these lower bounds also a confidence lower bound for the $P(CS)$ can be determined. Because σ is not known but is estimated, we first determine a 99% upper bound for it. The number of degrees of freedom is 60 and $\chi^2_{60,0.01} = 37.485$, hence an upper bound for σ is $\sigma_u = 0.52/\sqrt{37.485/60} = 0.66$. So the simultaneous lower bounds of $\tau_{(t)} - \tau_{(i)}$ can be written as : $L_1 = 2.18\ \sigma_u, L_2 = 1.91\ \sigma_u, L_3 = 1.06\ \sigma_u, L_4 = 0.95\ \sigma_u,\quad L_5 = 0.71\ \sigma_u,\quad L_6 = 0.65\ \sigma_u,\quad L_7 = 0.48\ \sigma_u,\quad L_8 = 0.21\ \sigma_u$ and $L_9 = \ldots = L_{20} = 0$. Now a 99-10=89% confidence lower bound of $P(CS)$ can be determined by simulation (see *4.3.3*), assuming that the variance is known and using the standardised L_i. This lower bound was equal to $P(CS)_L = 0.97$.

This lower bound of the probability of correct selection is rather large, so we could also try another, smaller, P^*. With $P_2^* = 0.90$ only 6 varieties have to be selected (varieties 8,6,17,3,14,18). The 90 % confidence lower bounds of $\tau_{(t)} - \tau_{(i)}$ remain unchanged, but the 89% confidence lower bound of $P(CS)$ does change, because the selection constant now has dropped to $\delta = 2.39$. With simulation we found that $P(CS)_L = 0.93$. This combination of P^* and $P(CS)_L$ may be satisfactory to the breeder, also in view of the corresponding subset size.

The success of this approach depends heavily on the real configuration of the variety parameters and the magnitude of the error variance. If the lower bounds for $\tau_{(t)} - \tau_{(i)}$ are only small and most of them zero, then the lower bound $P(CS)_L$ will be close to P^*. Then the extra effort does not pay off. However, lower bounds for $\tau_{(t)} - \tau_{(i)}$ and $P(CS)$ are always informative.

REFERENCES

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., C.W. Dunnett & M. Sobel (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika* 41, 170-176.

Bechhofer, R.E. & C.W. Dunnett (1988). Percentage points of multivariate Student t distributions. In : *Selected tables in mathematical statistics*, vol. 11. Providence, American Mathematical Society, Rhode Island.

Bose, R.C., W.H. Clatworthy & S.S. Shrikhande (1954). Tables of partially balanced designs with two associate classes. *Techn. Bull. No.* 107. North Carolina Agric. Exp. Station.

Bulmer, M.G. (1985). *The mathematical theory of quantitative genetics*. Clarendon Press, Oxford.

Butler, K.L. & D.G. Butler (1987). *Tables for selecting the best population*. Queensland Biometrical Bulletin 2, Department of Primary Industries, Queensland Government, Brisbane.

Cochran, W.G. (1951). Improvement by means of selection. *Proc. 2nd Berkeley symp. on Math. Statist. and Prob.*, 449-470.

Cochran, W.G. & G.M. Cox (1957). *Experimental designs*. 2nd edition. Wiley & Sons, New York.

Curnow, R.N. (1961). Optimal programmes for varietal selection. *J. R. Stat. Soc. B* 23, 282-318.

Desu, M.M. (1970). A selection problem. *Ann. Math. Statist.* 41, 1596-1603.

Dourleijn, C.J. & S.G.A.J. Driessen (1991). Subset selection procedures in randomized designs. *Technical Note* 91-02, Dept. of Mathematics, Agricultural University Wageningen, The Netherlands.

Driessen, S.G.A.J. (1991). Multiple comparisons with and selection of the best treatment in (incomplete) block designs. *Communications in Statistics - Theory & Methods* 20(1), 179-217.

Driessen, S.G.A.J. (1992). Statistical selection : multiple comparison approach. PhD. thesis, Eindhoven University of Technology.

Driessen, S.G.A.J., P. van der Laan & B. van Putten (1990). Robustness of the probability of correct selection against deviations from the assumption of a common known variance, *Biometrical Journal* 32, 131-142.

Dunnett, C.W. & M. Sobel (1954). A bivariate generalization of Student's t distribution, with tables for special cases. *Biometrika* 41, 153-169.

Falconer, D.S. (1986). *Introduction to quantitative genetics*. 2nd edition. Longman Scientific & Technical, Harlow.

Finney, D.J. (1958). Statistical problems in plant selection. *Bull. Inst. Int. Stat.* 36, 242-268.

Finney, D.J. (1960). *An introduction to the theory of experimental design.* The University of Chicago Press, Chicago.

Gibbons, J.D., I. Olkin & M. Sobel (1977). *Selecting and ordering populations, a new statistical methodology.* Wiley & Sons, New York.

Gupta, S.S. (1956). *On a decision rule for a problem in ranking means.* Ph.D. dissertation (and Mimeograph Series No. 150). Institute of Statistics, University of North Carolina, Chapel Hill.

Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics* 7, 225-245.

Gupta, S.S. & S. Panchapakesan (1979). *Multiple decision procedures. Theory and methodology of selecting and ranking populations.* Wiley & Sons, New York.

Gupta, S.S. & M. Sobel (1958). On selecting a subset which contains all populations better than a standard. *Ann. Math. Statist.* 29, 235-244.

Hsu, J.C. (1981). Simultaneous confidence intervals for all distances from the 'best'. *Ann. Statist.* 9, 1026-1034.

Hsu, J.C. (1984). Constrained simultaneous confidence intervals for multiple comparisons with the best, *Ann. Statist.* 12, 1136-1144.

John, P.W.M. (1971). *Statistical design and analysis of experiments.* The Macmillan Company, New York.

Laan, P. van der & B. van Putten (1990). Robustness of the normal means selection procedure with common known variance against logistic deviations and the use of the logistic approximation for the normal distribution. *Publication de l'Institute de Statistique de l'Université de Paris* 35, 79-92.

Laan, P. van der & L.R. Verdooren (1989). Selection of populations : An overview and some recent results. *Biometrical Journal* 31, 383-420.

Lam, K. (1989). The multiple comparison of ranked parameters. *Communications in Statistics - Theory & Methods* 18 (4), 1217-1237.

Young, J.C. (1972). An investigation of procedures for multiple-stage selection for a variate subject to errors of measurement. *Biometrika* 59, 323-334.

CHAPTER 5

Executing subset selection rules

In this chapter we will describe a case study of the use of subset selection in the plant breeding practice. In the previous chapters various obstacles on the way to practical application of statistical selection procedures have been removed. The case study, which is described in section **5.2**, concerns selection of sugar beet varieties that have been grown in trials at several sites. However, it is not feasible to execute the selection rules manually. Therefore, in section **5.1** computer software that shoulders this task is described first.

5.1 Computer software

If we want to use subset selection in the plant breeding practice, a computer program which executes the selection rule of our choice is indispensable. This computer program should also be able to deal with unbalanced incomplete block designs, because most of the plant breeding trials have this type of design. For the t independent samples situation (completely randomised designs) the computer package RANKSEL (Edwards, 1985) and the RSMCB procedure of SAS (Statistical Analysis System, version 5.*) (Aubuchon, Gupta & Hsu, 1985) can be used. The author has written a computer program in Fortran 77, called SUBSET, that uses the output of the selection constant simulation program SELCON (described in **4.3**). With the simulation program SELCON and the selection program SUBSET four types of selection-of-the-best subset selection rules (*R1* - *R4*, as described in **4.2**) can be executed for all possible designs. Also, the two rules corresponding to selection w.r.t. the average of control varieties (*R5* and *R6*, see **4.3**) can be executed. We will describe this selection program and demonstrate it with a small example.

Running the program we first have to indicate whether we want to select the best variety or select with respect to the average of control varieties. If we have chosen the first option, then we are asked whether we are interested in selecting good varieties too. If the answer is 'yes' then we have to give the distance parameter δ^*, in units of the estimated standard deviation. Next we have to choose the selection rule we want to use. If we have chosen one of the selection rules that uses standard errors of the contrast estimators, we have to supply the program with an input file which contains either the incidence scheme of the experimental design or the pseudo-variance/covariance matrix (divided by σ^2) of the parameter estimators. With the use of this information the standard errors of the contrast estimators are calculated. For selection with respect to control varieties, the number of controls has to be given. Furthermore, we have to supply the program with two input files: one containing a table with a range of selection constants and the corresponding 95% confidence lower limits of P^* (this is the output of the selection constant simulation program SELCON), and the other with the estimates of the standard deviation and the (contrasts between) variety parameters.

If the selection goal is to select the smallest subset that contains the best variety, or to select the smallest subset that contains at least one good variety, or to select the smallest subset that includes all varieties sufficiently better than the average of the control varieties, then the program calculates for each variety the smallest P^* leading to selection of that variety. The results are written to an output file. From the output we can easily read which subset size $|S|$ corresponds with which P^* and which varieties have to be selected. If the selection goal is to select the largest subset containing only good varieties, then the program calculates for each variety the largest P^* leading to selection of that variety. The P^* corresponding to a particular selection constant is approximated using log-linear interpolation. The output of the selection program can be best explained by an example.

Example

We will use a small example, also described by Driessen (1991). The experiment has a simple 2x2 lattice design. The estimates of the four treatment parameters are -1.875, -0.125, -0.625 and 2.625. We will assume that the treatments are different varieties. The error variance is assumed to be known and equal to 1. When σ is assumed to be known we replace s by σ in the selection rules and calculate the selection constant for the situation of known variance. A table of selection constants and corresponding 95% confidence lower limits of P^* for rule *R3* was created with the selection constant simulation program SELCON. For the simple 2x2 lattice design the selection constant δ, calculated for rule *R3*, is identical to the selection constants δ_i corresponding to rule *R1*. Suppose we are interested in selection of the best and selection of good varieties. The distance parameter δ^* is taken to be equal to 1. Using selection rule *R3*, we get the following output of the selection program:

```
Design : 2x2 simple lattice
Number of varieties : 4
Selection of the best
Selection rule : (3)

Explanation :
A subset size |S| is obtained for P* in the interval [P*l, P*u>.
Hence a variety R, with estimate EST, is selected for a P*
larger
than or equal to P*min.

Selection of the smallest subset including the best variety.

|S|  [P*l, P*u>                R    EST         P*min
------------------------------------------------------------
1    [ 0.0000 ,  0.9664>  4     2.6250    0.0000
2    [ 0.9664 ,  0.9885>  2    -0.1250    0.9664
3    [ 0.9885 ,  0.9977>  3    -0.6250    0.9885
4    [ 0.9977 ,  1.0000>  1    -1.8750    0.9977

Selection of the smallest subset including a good variety. A
good variety is at most 1.0000 units worse than the best variety.
This is 1.00 times the estimated standard deviation.

|S|  [P*l, P*u>                R    EST         P*min
------------------------------------------------------------
1    [ 0.0000 ,  0.9966>  4     2.6250    0.0000
2    [ 0.9966 ,  0.9992>  2    -0.1250    0.9966
3    [ 0.9992 , +0.9997>  3    -0.6250    0.9992
4    [+0.9977 ,  1.0000>  1    -1.8750    +.9997

Selection of the largest subset containing good varieties only.
A good variety is at most 1.0000 units worse than the best
variety. This is 1.00 times the estimated standard deviation.

Explanation :
A subset size |S| is obtained for P* in the interval [P*l, P*u>.
Here variety R, with corresponding character, is selected if
P* is smaller than or equal to P*max.

|S|  [P*l, P*u>                R    EST         P*max
------------------------------------------------------------
4                         1    -1.8750    0.0000
.                         3    -0.6250    0.0000
.                         2    -0.1250    0.0000
1    [ 0.0000 ,  0.9966>  4     2.6250    0.9966
```

The notation +0.9997 denotes a value between 0.9997 and 1.0000. The table with selection constants and corresponding 95 % lower limits of P^* stopped at a P^* value of 0.9997. Suppose we would like to select with $P^* = 0.99$. For obtaining the smallest subset including the best variety with confidence 0.99, we would have to select variety 4 (with parameter estimate 2.6250), variety 2 (with parameter estimate -0.1250) and variety 3 (with parameter estimate -0.6250). Thus the subset size would be 3. For obtaining the smallest subset including a good variety we only would have to select variety 4. However, notice that the definition of 'good' is very wide. To select the largest subset containing good varieties only, we would have to select variety 4. In this example, 1 is the largest subset size possible. Notice that with $P^* > 0.9966$ the subset comprising good varieties only is empty; even the variety with the largest estimated variety parameter cannot be selected as being 'good'.

The use of the above described selection program facilitates the use of subset selection and gives better insight in the consequences of selecting a prespecified number of varieties. For example, if one decides to select only one variety (variety 4 with parameter estimate 2.6250), then the statement that the best variety is included in this subset (with size 1) can be made with confidence $P^* = 0.9664$.

5.2 A case study

In this section we will apply statistical selection procedures, more specifically subset selection procedures, to a dataset generously supplied by the research division of The Royal Vanderhave Group, situated in Rilland, the Netherlands. In doing so, we will come across several topics discussed in the previous chapters. It is not feasible to study all treated topics in this case study, because then it would become too extensive. In *5.2.1* the experiment is described, as are the selection aim and the observed characters. In *5.2.2* contrasts between variety values are estimated, paying attention to estimation in concatenated trials and combined estimators. Section *5.2.3* deals with subset selection of the best variety, both at separate sites and using 'mean performance' estimates. Other selection goals are discussed in *5.2.4* and *5.2.5*, namely selection of at least one good variety and selection with respect to the control varieties, respectively. In *5.2.6* the extra assumption of the existance of a superpopulation of variety parameters is made, and the varieties selected with a rule that satisfies the $\hat{E}[P(CS)]$-condition (see **4.5**). Finally, the case study is evaluated in *5.2.7*.

5.2.1 Description of the experiment

We consider variety trials of sugar beet varieties. There are 110 new varieties, of which only a limited amount of seed is available. There are 6 experimental fields, namely fields situated near Rilland (The Netherlands), in the Flevopolder (The Netherlands), near Ingeleben (Germany), near Hevesen (Germany), near Rosière (France) and near Avelin (France). However, because of shortage of seed or lack of space at the individual sites, the varieties cannot be grown at all sites. To be able to compare the results of the various sites, the same three control varieties are added to the group of varieties at each site. These control varieties have variety numbers 1, 2 and 3.

At the individual sites the trials are of the, in **3.3** defined, concatenated type. The trial at a site can be broken down into smaller subtrials with different new varieties, and these subtrials are only connected by the three control varieties. So varieties 1,2 and 3 are grown in all subtrials. The subtrials have either a 5×5 lattice design with 3 replications (denoted by L3), or a 5×5 lattice design with 4 replications (L4), or a randomised complete block design with 4 blocks containing 25 varieties (C4). Consequently, a subtrial contains 22 new varieties and 3 control varieties. The group of new varieties in one subtrial is called a series, and this series is grown at three or four sites. The dataset contains 5 series (5×22=110 new varieties), spread over 6 sites. The experimental scheme can be summarised as

	The Netherlands		Germany		France	
Site :	Rilland	Flevopolder	Ingeleben	Hevesen	Rosière	Avelin
Series : 1		L4	L4	L4	L4	
2	C4	L4	L4	L4		
3	L3	L3	L3	L3		
4		L3	L3		L3	L3
5	L3	L3				L3

We assume that the aim is to select varieties that on the average are superior at the chosen sites. For that reason several characters have been observed. We restrict ourselves to 3 characters. Two basic characters are corrected root yield (CRY) and sugar content (SC). A derived character is white sugar yield (WSY). We assume that these characters can be considered Normally distributed. The latter character can be seen as a selection index. WSY is defined in chapter **2**. To achieve the selection aim we will use statistical selection.

5.2.2 Estimation of contrasts between variety values

The first part of statistical selection is the estimation of (contrasts between) variety values. Also the determination of the pseudo-variance/covariance matrix (divided by σ^2) of the variety parameter estimators and the estimation of the error variance are important to later approximate the selection constant(s) and to execute the selection rules, respectively. We will give the BLUEs of the contrasts between the new variety parameters and the average of the control variety parameters, calculated at the individual sites. Furthermore we will give the combined estimates, also called the 'mean performance' contrast estimates of the varieties.

At a single site we will use the fixed additive model (3.1) to describe the observations. One could also decide to use the mixed additive model (3.7) with fixed replication terms and random blocks within replication terms. We decided to use the fixed model because we have no knowledge about the ratio σ_B^2/σ^2, with σ_B^2 the variance component between blocks within replications and σ^2 the error variance component. Assume $\sigma_B^2 \gg \sigma^2$. If one still wants to use the mixed model, then there is nothing for it but to estimate this ratio from the current experiment, e.g. using the REML procedure. Because in this case study the number of varieties, subtrials and blocks is not very large, it is feasible to analyse these subtrials as one trial in one step, e.g. with the SAS computer package. However, we can also use the theory developed in **3.3** and perform the estimation procedure in two steps. Then, first the variety parameters are estimated as good as possible at the separate subtrials and next these subtrial estimates are combined into the best variety parameter estimates at the trial level. In *3.3.3* it has been proven that the local estimator of a contrast between a new variety parameter and the average of the control variety parameters is already the BLUE if the new variety is included in the subtrial with the C4 design. The estimates w.r.t. CRY, SC and WSY are presented in Tables 5.1, 5.2 and 5.3, respectively. If we look at Table 5.1, we notice that most of the CRY contrast estimates are negative. The three control varieties are varieties that are currently on the market and therefore are very good. However, they do not have a very high sugar content, because in Table 5.2 we see that most SC contrast estimates have a positive sign. But the average control variety value estimates for WSY are larger than most of the variety value estimates of the new varieties, as seen in Table 5.3.

The results at the separate sites can be extended by the coefficient of variation ($c.v.$), the root of the error mean square (s) and the degrees of freedom for error (df_e). They are equal to :

		Rill.	Flev.	Inge.	Heve.	Rosi.	Avel.
CRY	$c.v.$	0.05	0.03	0.04	0.04	0.04	0.06
	s	31.12	21.33	23.48	19.03	22.36	34.22
	df_e	148	228	190	152	94	74
SC	$c.v.$	0.02	0.01	0.02	0.02	0.02	0.01
	s	40.51	24.90	27.58	28.61	28.78	20.57
	df_e	148	228	190	152	94	74
WSY	$c.v.$	0.05	0.03	0.04	0.04	0.04	0.06
	s	46.17	36.97	35.87	33.34	31.95	51.17
	df_e	148	228	190	152	94	74

where Rill. means Rilland, Flev. means the Flevopolder, Inge. means Ingeleben, Heve. means Hevesen, Rosi. means Rosière and Avel. means Avelin.

After the analyses at the individual sites, a model for the joint observations of the various sites is considered. We will use the fixed additive model (3.8). In the current case study the analysis can be done in one step, e.g. with the SAS package. If experimenters cannot make use of powerful statistical packages, or if the number of sites is very large, the BLUEs of contrasts between variety values can be calculated in two steps. Then the local BLUEs are combined into the 'mean performance' BLUEs. The corresponding theory is already elaborated in *3.2.2*, where also examples are given. This two step method will be useful in practice if software is developed that makes the two step method easy to perform. The local BLUEs and the pseudo-variance/covariance matrix of the variety parameter estimators become available in the first step, because breeders also want to study the local estimates and execute a selection rule with these estimates (the variance/covariance matrix is necessary to calculate the selection constants). Therefore all ingredients are there to calculate the combined estimators (which are BLUEs), and it is not necessary to calculate these estimators with the use of all observations in one step.

The outcomes of the 'mean performance' BLUEs of the contrasts between the new variety parameters and the average of the control variety parameters are given in Table 5.4. Not given here is a pseudo-variance/covariance matrix of the variety parameter estimators. However, this matrix was determined for later use

Table 5.1. Outcomes of the BLUEs of the contrasts between new variety (NV) parameters and the average of the control variety parameters, for the character CRY (100 kg/ha) at Rilland (Rill.), the Flevopolder (Flev.), Ingeleben (Inge.), Hevesen (Heve.), Rosière (Rosi.) and Avelin (Avel.).

NV	Rill.	Flev.	Inge.	Heve.	Rosi.	Avel.
4		-49.5	-76.5	-45.1	-50.2	
5		-21.5	-40.0	-19.0	-33.1	
6		-63.5	-90.4	-65.2	-51.0	
7		-41.2	-48.4	-59.6	-52.4	
8		-21.3	-22.4	0.4	-29.5	
9		10.2	9.6	5.0	-6.2	
10		-2.4	-50.1	-14.4	-40.3	
11		30.8	7.0	3.3	1.8	
12		-14.2	-64.3	-34.5	-47.2	
13		-49.7	-74.0	-55.7	-42.0	
14		-55.5	-53.3	-46.7	-58.5	
15		-40.2	-49.7	-48.2	-6.4	
16		7.4	-58.5	-30.7	-46.6	
17		-61.1	-30.3	-7.2	-28.2	
18		-20.9	-36.8	-18.7	-27.1	
19		-45.0	-62.9	-33.4	-52.4	
20		-22.4	-53.3	-10.5	-34.1	
21		-55.8	-52.8	-14.3	-51.1	
22		-66.2	-79.5	-50.2	-32.4	
23		-22.1	-64.1	-19.4	-43.7	
24		-9.9	-17.3	-2.9	-29.3	
25		-67.0	-74.0	-67.7	-66.7	
26	-18.9	-8.9	-29.4	-0.1		
27	-4.2	-43.1	-33.4	-26.4		
28	-42.4	-62.8	-62.8	-40.2		
29	-18.9	-52.1	-73.9	-34.8		
30	-14.9	-63.9	-44.4	-38.7		
31	-41.9	-30.6	-34.7	-20.3		
32	-34.4	-72.7	-87.0	-18.2		
33	-27.4	-65.0	-28.3	-27.4		
34	-37.4	-49.0	-82.2	-17.6		
35	-54.7	-11.3	-29.6	-4.6		
36	-44.4	-30.1	-56.8	-46.9		
37	-83.7	-84.7	-70.9	-3.2		
38	-105.4	-180.7	-79.2	-65.4		
39	-97.2	-92.4	-69.5	-44.7		
40	-70.9	-100.8	-94.5	-50.5		
41	-98.7	-60.3	-67.7	-26.1		
42	-53.9	-47.2	-49.8	-31.8		
43	-109.4	-72.8	-73.4	-4.8		
44	-51.7	-35.8	-25.0	5.5		
45	-82.2	-70.0	-74.5	-28.0		
46	-83.4	-47.0	-36.7	-36.5		
47	-53.7	-59.2	-79.5	-32.6		
48	-54.4	-0.7	-4.7	-40.0		
49	-79.1	-60.5	-28.3	-57.0		
50	-12.4	-4.5	5.7	1.5		
51	-58.5	-33.2	-4.1	-5.4		
52	1.2	-33.1	-11.7	-21.9		
53	-23.1	-66.4	-62.1	-45.2		
54	15.3	-28.4	-29.3	-29.1		
55	-5.4	-75.9	-58.3	-71.9		
56	-51.5	-30.2	-44.4	-33.0		
57	-61.3	-90.3	-55.4	-49.2		
58	-30.4	-88.0	-84.6	-55.9		
59	-54.9	-65.2	-29.0	-73.7		
60	-66.1	-58.5	-49.2	-60.0		
61	-48.5	-34.9	-73.8	-18.5		
62	-43.7	-26.4	-12.1	-1.1		
63	-30.1	-25.9	-64.0	-21.9		
64	-57.0	-23.9	-41.9	-21.9		
65	-32.9	-70.3	-3.7	-30.7		
66	-14.6	-33.6	-54.2	-35.1		
67	-63.1	-59.9	-85.2	-42.9		
68	-53.4	-76.3	-75.5	-61.0		
69	-27.5	-47.9	-36.3	-57.7		
70		29.4	-59.7		-1.4	108.3
71		16.1	-46.8		7.2	50.2
72		37.5	-57.2		31.7	49.1
73		82.2	-108.8		76.8	58.2
74		-7.7	-86.8		22.1	-2.8
75		2.2	-75.3		43.9	42.2
76		-20.3	-13.4		47.6	93.6
77		72.8	-123.0		21.6	48.4
78		50.3	-22.7		-10.9	5.7
79		49.1	-43.4		96.5	75.9
80		-35.5	-87.1		10.2	-11.3
81		40.6	-63.2		22.8	20.2
82		44.9	-64.1		21.3	12.4
83		44.1	-14.0		24.6	57.8
84		112.1	-22.2		92.2	63.8
85		38.4	28.5		47.5	45.4
86		49.9	-73.4		28.9	87.5
87		16.3	-54.4		47.9	31.1
88		28.9	-56.6		-4.1	32.1
89		20.8	-84.4		21.1	10.7
90		6.9	-30.3		46.9	59.0
91		75.2	-4.1		67.0	72.7
92	-147.2	-163.5				-100.2
93	-47.8	-72.9				-42.1
94	-68.0	-77.5				-63.7
95	-87.9	-82.7				-57.3
96	-28.7	-85.8				-59.6
97	-108.1	-115.6				-73.3
98	-47.6	-39.6				-16.7
99	-74.3	-106.2				-90.8
100	-27.6	-59.4				-28.8
101	-25.3	-40.5				-38.6
102	-56.1	-59.7				-20.4
103	-89.4	-75.6				-24.1
104	-103.8	-84.4				-58.0
105	-56.1	-34.8				-39.4
106	-61.2	-45.5				-40.1
107	-76.1	-78.7				-68.9
108	-73.8	-103.0				-62.9
109	-28.4	-44.1				-28.9
110	-43.7	-10.1				-10.2
111	-35.7	-6.3				13.8
112	-76.4	-28.8				-7.3
113	-42.8	-17.7				-3.9

Table 5.2. Outcomes of the BLUEs of the contrasts between new variety (NV) parameters and the average of the control variety parameters, for the character SC (1/100 %) at Rilland (Rill.), the Flevopolder (Flev.), Ingeleben (Inge.), Hevesen (Heve.), Rosière (Rosi.) and Avelin (Avel.).

NV	Rill.	Flev.	Inge.	Heve.	Rosi.	Avel.
4		69.5	42.9	46.6	43.3	
5		20.3	28.6	31.1	-26.5	
6		25.5	55.4	42.4	40.7	
7		65.7	71.6	28.7	109.7	
8		-10.8	15.4	-7.3	21.3	
9		7.9	43.7	-6.5	-7.4	
10		-10.5	40.7	-24.3	23.8	
11		-4.5	43.1	11.0	10.7	
12		25.9	61.9	-21.2	34.4	
13		28.9	61.4	30.2	56.7	
14		83.6	152.0	22.7	51.7	
15		24.4	48.8	17.2	38.3	
16		26.4	65.2	9.3	59.1	
17		26.9	29.8	-13.2	-15.1	
18		18.5	61.2	44.6	29.5	
19		30.7	49.3	-14.4	-21.4	
20		56.8	51.5	-7.7	51.7	
21		74.9	97.1	5.4	67.6	
22		8.6	54.9	-12.2	15.9	
23		24.9	48.2	39.3	60.3	
24		23.1	64.8	51.5	41.7	
25		25.7	67.8	-19.1	26.4	
26	25.3	-8.9	23.3	26.9		
27	-12.0	6.7	41.7	-11.8		
28	65.5	36.0	92.5	28.1		
29	34.8	45.6	95.2	20.1		
30	35.0	27.3	76.5	10.7		
31	-1.5	22.2	54.4	24.7		
32	8.8	7.0	46.1	32.0		
33	44.5	82.9	78.6	33.0		
34	73.0	83.0	142.1	86.5		
35	4.3	-3.0	26.4	-57.4		
36	68.0	59.2	111.8	7.7		
37	91.0	128.5	152.5	51.0		
38	71.5	-5.7	107.2	-0.1		
39	119.8	93.9	91.5	39.2		
40	43.3	40.4	98.2	60.9		
41	76.0	42.0	86.4	13.2		
42	54.5	51.5	92.9	65.5		
43	73.5	35.7	97.0	31.3		
44	62.3	76.0	82.6	22.9		
45	73.8	41.2	109.9	47.4		
46	44.3	46.0	79.7	19.5		
47	5.0	29.3	93.3	45.4		
48	89.0	35.7	50.0	16.7		
49	-3.3	39.2	23.5	-19.8		
50	29.8	30.4	4.5	15.3		
51	37.6	24.7	11.6	32.7		
52	53.8	38.5	12.4	-1.9		
53	128.2	26.5	81.1	33.3		
54	73.9	4.1	2.0	49.5		
55	64.6	32.6	43.2	19.3		
56	101.4	83.1	34.9	20.5		
57	89.3	12.0	46.6	0.8		
58	30.2	27.6	54.7	34.9		
59	49.9	34.7	21.4	-7.2		
60	113.2	59.3	57.3	41.3		
61	143.6	98.6	110.3	74.9		
62	51.5	39.9	93.1	8.5		
63	68.6	51.6	59.3	50.0		
64	118.8	79.2	59.5	40.6		
65	107.6	76.0	59.6	9.2		
66	73.9	59.2	78.7	83.7		
67	146.9	81.2	47.7	88.6		
68	143.6	123.9	76.6	36.0		
69	96.4	80.7	104.3	30.2		
70		5.2	37.4		11.2	14.0
71		18.2	-24.9		10.1	13.5
72		42.7	-34.1		24.9	19.0
73		-29.0	-81.1		-18.3	19.2
74		-47.0	-107.1		-59.0	-33.3
75		18.7	38.7		58.8	21.0
76		43.6	-2.2		50.2	70.2
77		-22.5	-65.6		-8.7	-3.3
78		21.5	2.9		68.4	22.1
79		-0.6	-22.1		6.3	20.3
80		56.1	19.7		44.8	36.6
81		22.0	-7.5		48.3	32.8
82		67.4	35.5		85.1	78.7
83		11.8	-46.9		12.0	-6.8
84		-7.4	-64.6		-25.2	-6.5
85		29.5	-35.5		24.1	43.4
86		47.0	8.3		10.6	39.3
87		-2.7	-44.8		4.7	25.5
88		25.8	-6.7		11.7	25.2
89		21.0	-11.9		23.0	17.2
90		17.0	-11.5		0.5	26.1
91		23.3	-70.3		-20.2	-8.5
92	151.5	99.9				97.3
93	95.7	55.1				66.9
94	85.1	55.3				20.9
95	124.8	129.5				94.3
96	56.2	65.6				-1.1
97	62.1	51.8				27.8
98	62.4	26.4				-33.7
99	82.3	93.4				89.9
100	159.2	104.7				93.5
101	89.6	52.5				84.3
102	61.5	74.0				56.8
103	136.6	108.4				66.0
104	78.7	43.1				-0.2
105	148.8	120.8				104.1
106	131.6	97.3				97.3
107	103.5	87.0				74.1
108	34.9	50.4				-34.1
109	20.7	33.0				-19.2
110	61.1	55.5				26.4
111	32.2	27.8				14.0
112	79.1	72.9				65.6
113	71.3	82.4				77.4

Table 5.3. Outcomes of the BLUEs of the contrasts between new variety (NV) parameters and the average of the control variety parameters, for the character WSY (10 kg/ha) at Rilland (Rill.), the Flevopolder (Flev.), Ingeleben (Inge.), Hevesen (Heve.), Rosière (Rosi.) and Avelin (Avel.).

NV	Rill.	Flev.	Inge.	Heve.	Rosi.	Avel.
4		-31.2	-73.1	-49.4	-46.7	
5		-7.2	-23.2	-8.9	-55.1	
6		-84.2	-85.7	-82.4	-39.9	
7		-23.7	-23.2	-80.5	-7.8	
8		-35.6	-11.3	-1.2	-24.1	
9		14.7	42.1	-3.0	-20.8	
10		-9.6	-33.8	-33.0	-40.0	
11		57.9	48.1	15.6	14.0	
12		-0.2	-44.6	-64.9	-42.0	
13		-58.0	-63.6	-78.0	-22.7	
14		-25.1	22.6	-63.6	-48.4	
15		-43.7	-36.5	-66.3	10.5	
16		45.5	-26.0	-38.5	-21.8	
17		-78.4	-15.8	-10.4	-46.4	
18		-8.3	-2.4	-1.8	-8.1	
19		-49.3	-51.1	-58.6	-82.1	
20		13.7	-31.5	-12.3	-9.9	
21		-35.7	-14.5	-21.4	-30.2	
22		-102.7	-74.1	-89.6	-36.6	
23		-23.9	-68.5	-12.6	-25.5	
24		7.1	23.6	27.7	-12.8	
25		-99.6	-66.8	-116.3	-76.1	
26	-20.2	-22.9	-30.0	9.8		
27	-15.9	-64.2	-18.0	-43.6		
28	-14.9	-72.8	-29.4	-46.3		
29	2.3	-36.9	-33.2	-39.8		
30	-6.4	-82.1	-10.1	-53.3		
31	-77.2	-40.2	-23.4	-21.0		
32	-52.9	-115.1	-96.3	-14.7		
33	-18.9	-42.4	9.7	-27.3		
34	6.6	-1.8	-19.0	22.9		
35	-95.4	-27.2	-35.4	-50.3		
36	-11.2	10.5	0.9	-60.8		
37	-70.2	-47.5	-3.3	23.0		
38	-125.7	-310.7	-47.0	-100.2		
39	-75.7	-79.6	-31.3	-48.6		
40	-92.2	-142.3	-81.4	-50.5		
41	-106.7	-64.8	-41.1	-32.8		
42	-42.7	-37.1	-15.7	-20.9		
43	-129.2	-86.8	-40.9	14.7		
44	-43.9	3.5	25.3	20.5		
45	-78.9	-77.2	-36.7	-21.4		
46	-99.2	-37.4	6.5	-41.4		
47	-83.4	-72.9	-50.9	-25.7		
48	-24.8	29.8	29.0	-55.7		
49	-121.9	-67.0	-22.4	-100.4		
50	14.9	20.9	16.3	14.0		
51	-64.5	-32.7	1.0	5.4		
52	30.2	-28.2	-9.3	-34.8		
53	57.0	-79.8	-37.7	-55.5		
54	75.9	-44.5	-42.4	-20.6		
55	33.0	-93.2	-60.9	-105.6		
56	-9.8	17.5	-42.4	-41.4		
57	-28.1	-127.7	-48.0	-75.6		
58	-18.8	-121.9	-90.1	-72.1		
59	-56.0	-87.9	-27.4	-120.8		
60	-17.1	-47.5	-27.8	-75.0		
61	25.2	20.1	-46.4	11.4		
62	-20.6	-3.6	54.2	5.2		
63	8.8	0.9	-54.6	-8.8		
64	-12.1	21.7	-20.1	-13.9		
65	9.8	-67.8	26.4	-51.6		
66	28.8	-7.1	-30.6	-15.5		
67	12.6	-28.8	-94.0	-25.6		
68	18.1	-33.6	-69.1	-79.8		
69	36.6	-8.9	18.7	-74.7		
70		46.1	-64.1		6.3	160.9
71		36.6	-75.2		14.9	77.6
72		96.6	-98.7		53.4	73.3
73		101.1	-209.3		72.9	80.6
74		-63.4	-208.0		-23.4	-33.6
75		27.6	-86.4		90.8	71.9
76		4.9	-22.9		87.4	172.6
77		78.3	-227.9		17.0	57.5
78		107.6	-10.8		33.1	31.1
79		76.8	-78.2		126.0	116.9
80		-20.9	-115.8		31.3	-1.1
81		89.0	-91.2		61.6	45.4
82		135.2	-68.9		80.1	61.7
83		91.4	-40.5		44.0	79.1
84		174.2	-67.7		99.5	85.5
85		98.4	25.6		77.3	93.4
86		108.9	-106.7		41.7	145.2
87		18.0	-112.1		49.8	52.3
88		71.1	-75.9		0.7	59.2
89		47.1	-138.4		32.8	18.6
90		24.3	-46.9		61.6	101.7
91		133.6	-57.0		66.9	89.4
92	-149.6	-180.1				-87.9
93	-5.9	-56.7				-12.5
94	-50.2	-65.5				-69.8
95	-59.1	-23.1				-24.2
96	-2.8	-68.4				-79.1
97	-117.8	-134.1				-74.4
98	-30.2	-32.4				-33.0
99	-58.0	-85.8				-73.2
100	53.5	-11.2				5.9
101	14.6	-15.3				-8.3
102	-48.1	-31.6				1.3
103	-62.9	-32.1				-1.1
104	-119.9	-99.8				-83.5
105	3.0	54.9				3.3
106	-23.8	6.6				-1.6
107	-53.2	-48.5				-54.2
108	-92.2	-118.7				-111.1
109	-30.3	-31.8				-47.9
110	-41.1	26.2				-6.2
111	-47.5	10.7				22.7
112	-60.9	35.0				33.5
113	-6.4	69.5				61.5

Table 5.4. Outcomes of the 'mean performance' BLUEs of the contrasts between new variety (NV) parameters and the average of the control variety parameters, for the characters CRY (100 kg/ha), SC (1/100 %) and WSY (10 kg/ha).

NV	CRY	SC	WSY
4	-53.9	51.2	-47.4
5	-28.4	14.5	-23.0
6	-66.2	42.2	-70.3
7	-50.8	69.1	-34.2
8	-17.2	4.7	-16.8
9	4.2	9.2	7.7
10	-26.9	7.9	-28.8
11	10.7	14.8	34.5
12	-40.2	26.3	-37.6
13	-54.4	44.6	-53.9
14	-53.2	77.7	-27.9
15	-36.1	32.6	-33.9
16	-32.8	40.6	-10.7
17	-31.0	7.5	-36.5
18	-25.6	38.8	-4.6
19	-49.0	11.2	-61.0
20	-28.4	38.4	-7.2
21	-43.3	61.4	-25.1
22	-56.4	17.5	-74.1
23	-37.3	43.7	-32.0
24	-14.3	46.2	12.6
25	-69.7	24.6	-90.2
26	-15.2	16.8	-17.4
27	-26.0	5.9	-34.5
28	-52.0	56.7	-39.8
29	-44.4	48.7	-26.2
30	-39.6	37.4	-36.6
31	-33.2	24.5	-43.0
32	-53.3	23.0	-71.0
33	-36.4	60.9	-18.2
34	-44.6	95.1	4.9
35	-27.5	-6.9	-56.5
36	-43.8	63.1	-13.0
37	-61.5	103.7	-27.0
38	-107.6	44.4	-145.3
39	-75.8	86.7	-58.3
40	-78.0	59.3	-90.8
41	-65.5	56.4	-64.0
42	-44.9	63.7	-28.9
43	-66.8	61.0	-62.7
44	-27.3	61.7	0.5
45	-62.9	68.9	-51.3
46	-51.7	45.7	-45.3
47	-54.0	41.7	-55.7
48	-22.8	47.2	-2.5
49	-56.0	11.0	-77.1
50	-3.0	19.9	15.6
51	-24.7	26.3	-22.1
52	-14.8	25.4	-8.4
53	-47.0	67.6	-25.5
54	-19.3	33.3	-9.4
55	-53.3	39.7	-57.7
56	-38.2	60.2	-16.4
57	-62.8	38.0	-67.4
58	-64.5	37.5	-74.9
59	-55.9	23.6	-74.2
60	-57.7	67.9	-40.7
61	-42.6	108.3	5.6
62	-21.1	48.8	9.0
63	-34.4	56.3	-12.4
64	-36.6	73.9	-7.7
65	-34.3	65.0	-19.6
66	-34.3	74.9	-5.2
67	-62.1	90.0	-33.6
68	-65.2	93.7	-40.2
69	-42.0	77.8	-6.7
70	21.7	17.9	41.7
71	6.4	2.3	12.1
72	16.3	12.0	32.2
73	30.6	-27.1	16.5
74	-16.3	-62.8	-79.5
75	4.9	35.8	29.5
76	20.4	40.1	51.1
77	4.9	-24.6	-18.2
78	3.8	28.8	38.5
79	48.8	-1.5	64.4
80	-28.7	40.2	-22.8
81	2.7	24.0	22.3
82	5.1	67.9	54.9
83	24.8	-7.5	39.1
84	62.0	-27.0	72.8
85	38.2	15.0	70.9
86	28.4	25.7	54.5
87	8.3	-5.6	-0.7
88	1.9	16.2	17.7
89	-9.1	12.3	-11.7
90	21.9	9.8	37.7
91	54.9	-22.5	59.1
92	-135.7	116.8	-136.4
93	-55.6	72.4	-26.9
94	-68.7	53.9	-59.4
95	-77.2	116.2	-36.8
96	-59.7	41.3	-52.5
97	-98.7	47.6	-108.6
98	-33.0	17.1	-29.8
99	-90.9	86.0	-74.0
100	-36.5	119.1	18.9
101	-33.6	77.7	-0.6
102	-47.6	64.5	-29.0
103	-63.6	103.9	-31.9
104	-80.7	39.7	-99.7
105	-42.5	124.8	22.8
106	-50.5	106.7	-9.3
107	-73.9	90.3	-49.3
108	-81.3	18.0	-109.9
109	-34.1	13.0	-35.5
110	-21.3	48.7	-6.4
111	-10.8	24.9	-6.8
112	-37.5	69.5	0.6
113	-19.8	77.6	44.7

when the selection constant has to be approximated. The coefficient of variation, the root of the error mean square and the degrees of freedom for error are equal to :

	CRY	SC	WSY
$c.v.$	0.05	0.02	0.05
s	28.22	31.39	45.30
df_e	1204	1204	1204

5.2.3 Selection of the best variety

Now that the first part of statistical selection, i.e. the estimation of (contrasts between) variety values, is accomplished, we can continue with the second part. This part consists of the execution of selection rules. In this section we will concentrate on the selection of a minimum sized subset containing the best variety. This can be done for each of the three characters, bearing in mind that WSY is a selection index derived from (among other characters) CRY and SC. The control varieties should be excluded from selection, so that they cannot enter the selected subset.

We will make use of selection rule *R3* (see *4.2.2*). This selection rule contains a single selection constant δ that has to be approximated by simulation, as described in *4.3.1*. We need the incidence matrix of the trial or a pseudo-variance/covariance matrix (with $\sigma^2 = 1$) of the variety parameter estimators. If the trial is extensive (in the number of blocks), then the latter option is easier to work with. In chapter **3** it is shown how to calculate the pseudo-variance/covariance matrix of combined experiments. The matrix rows and columns corresponding to the control varieties were deleted, in order to calculate a selection constant for the situation where the control varieties are excluded from selection. With the selection constant simulation program SELCON described in *4.3.1* the selection constants were approximated (using 10000 iterations). We will give the values corresponding to $P^* = 0.70$, $P^* = 0.80$ and $P^* = 0.90$ for the separate sites and the combined trial (denoted by 'Overall').

	Rill.	Flev.	Inge.	Heve.	Rosi.	Avel.	Overall
$\delta\ (P^* = 0.70)$	2.10	2.21	2.14	2.10	1.96	1.96	2.22
$\delta\ (P^* = 0.80)$	2.34	2.46	2.39	2.34	2.22	2.23	2.47
$\delta\ (P^* = 0.90)$	2.72	2.82	2.75	2.71	2.60	2.61	2.82

The execution of the selection rule took place with the selection computer program SUBSET described in **5.1**. The output from this program is translated (reduced) into Tables 5.5 and 5.6. In Table 5.5 the selection results are presented for the separate sites and characters. We notice that the percentage selected varieties can differ enormously from site to site. This is probably mainly due to the fact that different sites have (partly) different varieties. In Table 5.6 the selection results on the basis of the 'mean performance' estimates are presented for each character. There are a few varieties with a very good CRY, and therefore we only have to select 4 % of the total number of new varieties to make the statement that with probability at least 0.80 the best variety w.r.t. CRY is included in the selected subset. For SC and WSY more varieties have to be selected to make the analogous probability statements.

Many of the varieties selected on the basis of 'mean performance' estimates are either also selected on *all* individual sites at which they were grown (CRY 84, 91; SC 34, 37, 61, 92, 95, 99, 100, 103, 105, 106, 107; WSY 85,113; $P^* = 0.80$) or selected at *most* of the relevant individual sites (CRY 85; SC 39,67,68; WSY 50, 72, 76, 82, 83, 84, 91; $P^* = 0.80$). On the other hand, it is possible that a variety is not selected at any of the individual sites but still is selected on the basis of 'mean performance' estimates (WSY 70). The fact that varieties are selected at all or most sites means that the varieties are consistant in their performance. Notice that w.r.t. SC the selected varities are very consistent.

In Figure 5.1 a scatter plot is given of the CRY and SC 'mean performance' contrast estimates of the 110 new varieties. Varieties that are selected (with $P^* = 0.80$) if the selection is focussed on the character CRY are indicated by a '1'. Varieties included in the subset if the selection is focussed on SC are indicated by a '2'. Varieties that are selected if the selection is focussed on WSY are indicated by a '*'. It can be seen in Figure 5.1 that the varieties selected on the basis of their CRY are also selected on the basis of their WSY. Only two of the varieties selected on the basis of SC are also selected on the basis of WSY. The two subsets corresponding to CRY and SC are disjunct. This is to be expected because these two characters are negatively correlated, as can be seen in the scatter plot. As selection index WSY seems an important criterium to base the selection on.

Table 5.5. Selected varieties with the selection-of-the-best rule *R3*, for $P^* = 0.70$, $P^* = 0.80$ and $P^* = 0.90$. The selection is performed at the separate sites and is based on CRY, SC and WSY, successively. Between brackets the subset size is expressed relatively to the number of new varieties at the particular site.

CRY, Rilland.

$P^* = 0.70$: 27,28,29,30,31,32,33,34,36,50,52,53,54,55,58,63,65,66,69,93,96,98,100,101,109, 110,111,113 (42 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 44,61,62,102,105 (50 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 26,35,42,47,48,51,56,59,64,67,68,94,106 (70 %).

CRY, Flevopolder.

$P^* = 0.70$: 73,77,84,91 (4 %).
$P^* = 0.80$: as for $P^* = 0.70$.
$P^* = 0.90$: as for $P^* = 0.70$.

CRY, Ingeleben.

$P^* = 0.70$: 9,11,24,48,50,51,52,62,65,76,83,85,91 (15 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 8,44,49,59,84 (20 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 17,26,27,33,35,54,69,78,90 (31 %).

CRY, Hevesen.

$P^* = 0.70$: 8,9,10,11,17,18,20,21,23,24,26,31,32,34,35,37,43,44,50,51,52,54,61,62,63,64,65 (41 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 27,33,41,45,56,66 (50 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 4,12,16,19,42,47,48,67 (62 %).

CRY, Rosière.

$P^* = 0.70$: 73,79,84,91 (9 %).
$P^* = 0.80$: as for $P^* = 0.70$.
$P^* = 0.90$: as for $P^* = 0.70$, + 76,85,87,90 (18 %).

CRY, Avelin.

$P^* = 0.70$: 71,72,73,76,77,79,83,84,86,90,91 (25 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 75,85 (30 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 87,88 (34 %).

SC, Rilland.

$P^* = 0.70$: 37,39,48,53,54,56,57,60,61,64,65,66,67,68,69,92,93,94,95,99,100,101,103,105,106, 107 (39 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 26,34,38,41,43,45,55,63,104,112,113 (56 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 28,36,44,52,59,62,97,98,102,110 (71 %).

SC, Flevopolder.

$P^* = 0.70$: 14,37,39,56,61,64,65,67,68,69,92,95,99,100,103,105,106,107,113 (17 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 21,33,34 (20 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 5,44,82,96,102,112 (25 %).

SC, Ingeleben.

$P^* = 0.70$: 14,34,36,37,38,45,61,69 (9 %).
$P^* = 0.80$: as for $P^* = 0.70$, + 62 (10 %).
$P^* = 0.90$: as for $P^* = 0.80$, + 21,29,40,43,53,66 (17 %).

(Table 5.5 continued)

SC, Hevesen.

$P^* = 0.70$: 4,5,6,18,23,24,34,37,40,42,45,47,54,58,60,61,63,64,66,67,68 (32 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 7,13,39,51,53,69 (41 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 14,26,28,32,33,43,48,55,56 (55 %).

SC, Rosière.

$P^* = 0.70$: 7,21,75,78,82 (11 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 16,23,76 (18 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 13,14,80,81 (27 %).

SC, Avelin.

$P^* = 0.70$: 76,82,92,93,95,99,100,101,105,106,107,113 (27 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 103,112 (32 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 102 (34 %).

WSY, Rilland.

$P^* = 0.70$: 29,30,34,36,50,52,53,54,55,56,61,63,64,65,66,67,68,69,93,96,100,101,105,106,113 (38 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 27,28,33,48,58,60,62,98,109 (52 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 57,94,102,107,110,111 (61 %).

WSY, Flevopolder.

$P^* = 0.70$: 78,82,84,85,86,91 (5 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 72,73,83 (8 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 77,81,113 (11 %).

WSY, Ingeleben.

$P^* = 0.70$: 4,8,9,11,14,18,24,27,30,33,34,36,37,42,44,46,48,50,51,52,62,65,69,76,78,85 (30 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 17,21,31,49,64 (35 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 7,16,26,28,29,35,39,45,59,60,66,83 (49 %).

WSY, Hevesen.

$P^* = 0.70$: 8,9,11,17,18,20,21,23,24,26,31,32,33,34,37,42,43,44,45,47,50,51,52,54,61,62,63,
64,66,67 (45 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 4,41,56 (50 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 10,16,27,28,29,46,48,53,65 (64 %).

WSY, Rosière.

$P^* = 0.70$: 73,75,76,79,82,84,85 (16 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 81,91 (20 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 72,90 (25 %).

WSY, Avelin.

$P^* = 0.70$: 71,73,76,79,84,85,86,90,91 (20 %).

$P^* = 0.80$: as for $P^* = 0.70$, + 72,75,83,113 (30 %).

$P^* = 0.90$: as for $P^* = 0.80$, + 77,82,87,88,112 (41 %).

Table 5.6. Selected varieties with the selection-of-the-best rule *R3*, for $P^* = 0.70$, $P^* = 0.80$ and $P^* = 0.90$. The selection is performed on the basis of the 'mean performance' estimates of CRY, SC and WSY, successively. Between brackets the subset size is expressed relatively to the 110 new varieties.

CRY	$P^* = 0.70$:	79,84,85,91 (4 %).
	$P^* = 0.80$:	as for $P^* = 0.70$.
	$P^* = 0.90$:	as for $P^* = 0.70$, + 73,86 (5 %).
SC	$P^* = 0.70$:	34,37,61,67,68,92,95,100,103,105,106,107 (11 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 39,99 (13 %).
	$P^* = 0.90$:	as for $P^* = 0.80$, + 69 (14 %).
WSY	$P^* = 0.70$:	11,70,72,75,76,78,79,82,83,84,85,86,90,91,100,105,113 (15 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 50 (16 %).
	$P^* = 0.90$:	as for $P^* = 0.80$, + 24,62,73,81,88 (21 %).

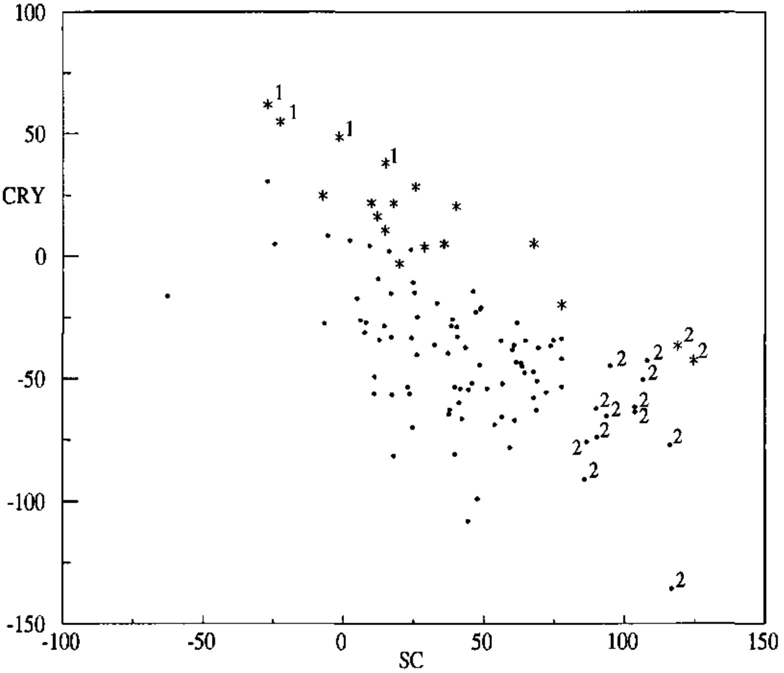

Figure 5.1. Selected varieties in case of subset selection of the best variety, based on the 'mean performance' estimates of CRY, SC and WSY separately, with $P^* = 0.80$. The scatter points are the 'mean performance' estimates of contrasts between new variety parameters and the average of the control variety parameters, for CRY (100 kg/ha) and SC (1/100 %). The varieties selected on the basis of CRY are indicated by a '1', the varieties selected on the basis of SC are indicated by a '2' and the varieties selected on the basis of WSY are plotted as a '*'.

5.2.4 Selection of at least one good variety

Selection of the *best* variety is a very stringent goal. Often the plant breeder is willing to relax this aim. He might consider the selection to be successful if at least one good variety is selected, with probability at least P^*. This less stringent selection goal should lead to smaller selected subsets. The definition of 'good' is descibed in *4.1.1*. The distance measure δ^* that gives the maximum distance between the parameter value of the best variety and the parameter value of a good variety has to be given by the experimenter. With his knowledge about the crop he should be able to give a meaningful value of δ^*. In this case study we take δ^* equal to two-tenth of the standard deviation estimates corresponding to a particular character.

We have used the selection rule (4.11) to make the selection. This selection rule makes use of the same selection constants as the selection-of-the-best rule *R3*. The selection rule was executed with the computer program SUBSET described in **5.1**, with control varieties 1, 2 and 3 excluded from selection. The selection was based on the 'mean performance' estimates of the contrasts bewteen the parameters of the new varieties and the average of the parameters of the control varieties. The selection results are given in Table 5.7, based on CRY, SC and WSY, successively.

Table 5.7. Selected varieties with the selection-of-at-least-one-good rule (4.11), for $P^* = 0.70$, $P^* = 0.80$ and $P^* = 0.90$. The distance measure δ^* is set to $0.2 \times s$. The selection is performed on the basis of the 'mean performance' estimates of CRY, SC and WSY, successively. Between brackets the subset size is expressed relatively to the 110 new varieties.

CRY	$P^* = 0.70$:	79,84,91 (3 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 85 (4 %).
	$P^* = 0.90$:	as for $P^* = 0.80$.
SC	$P^* = 0.70$:	37,61,68,92,95,100,103,105,106 (8 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 34,67,107 (11 %).
	$P^* = 0.90$:	as for $P^* = 0.80$, + 39,99 (13 %).
WSY	$P^* = 0.70$:	11,70,76,78,79,82,83,84,85,86,90,91,113 (12 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 72,105 (14 %).
	$P^* = 0.90$:	as for $P^* = 0.80$, + 75,100 (15 %).

From 5.7 we notice that the subset sizes have decreased by 0-6 %. The larger δ^* is chosen to be, the smaller the subset size will be.

5.2.5 Selection of all varieties better than the control varieties

A different selection goal could be to select a minimum sized subset that includes *all* varieties better than the average of the three control varieties. 'Better' is defined in *4.1.1*, and the distance measure δ^* has to be chosen by the experimenter. We have set δ^* to half the size of the standard deviation estimate s. Selection rule *R5*, which is used here, has been executed with the help of the computer program SUBSET described in **5.1.** The results, based on the 'mean performance' estimates of the contrasts between new varieties and the average of the control varieties, are presented in Table 5.8.

Table 5.8. Selected varieties with the selection-w.r.t.-controls rule *R5*, for $P^* = 0.70$, $P^* = 0.80$ and $P^* = 0.90$. The selection is performed on the basis of the 'mean performance' estimates of CRY, SC and WSY, successively. Between brackets the subset size is expressed relatively to the 110 new varieties.

CRY	$P^* = 0.70$:	9,11,50,70,71,72,73,75,76,77,78,79,81,82,83,84,85,86,87,88, 90,91,111,89 (22 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 52 (23 %).
	$P^* = 0.90$:	as for $P^* = 0.80$, + 74,110,113 (25 %).
SC	$P^* = 0.70$:	all except 73,74,77,84,91 (95 %).
	$P^* = 0.80$:	as for $P^* = 0.70$.
	$P^* = 0.90$:	as for $P^* = 0.70$.
WSY	$P^* = 0.70$:	9,11,16,18,20,24,34,36,44,48,50,51,52,54,56,61,62,63,64,65, 66,69,70,71,72,73,75,76,77,78,79,81,82,83,84,85,86,87,88,89, 90,91,93,100,101,105,106,110,111,112,113 (46 %).
	$P^* = 0.80$:	as for $P^* = 0.70$, + 8,26,80,98,102,103 (52 %).
	$P^* = 0.90$:	as for $P^* = 0.80$, + 33,53,109 (55 %).

The subset sizes are very large, especially for SC and WSY. If we look at Figure 5.1 we see that most of the varieties have SC contrast estimates that are positive. So, there are many varieties better than the average of the controls and the selected subset becomes quite extensive. The smaller δ^*, the larger the selected subset.

5.2.6 Assuming a superpopulation of the variety parameters

Sometimes it is reasonable to assume that the variety parameters are a random sample from a population of parameters. We now assume that the variety parameters have been drawn from a Normal population. Furthermore, we assume that the heritability h^2 is equal to 0.2. With these assumptions we can calculate a new set of selection constants, with the use of the simulation program SELCON described in *4.3.1*. Now a particular selection constant is not related to a certain P^*, but with a certain $\hat{E}[P(CS)]$. In *4.5.1* we have seen that we have to use values of $\hat{E}[P(CS)]$ of about 0.95. We will give the selection constants, corresponding to the combined trial (Overall), for $\hat{E}[P(CS)] = 0.90$, 0.95 and 0.99.

	Overall
δ ($\hat{E}[P(CS)] = 0.90$)	1.54
δ ($\hat{E}[P(CS)] = 0.95$)	1.85
δ ($\hat{E}[P(CS)] = 0.99$)	2.52

The selection-of-the-best rule *R3* was executed with the developed computer program SUBSET. The results are given in Table 5.9. For e.g. $\hat{E}[P(CS)] = 0.95$, we can make the statement that, given the assumption that there is a Normal superpopulation of variety parameters with heritability 0.2, the expectation of the probability that the best variety is included in the selected subset is 0.95. In Table 5.9 we see that this selection procedure leads to smaller subsets than selection of the best without the superpopulation assumption. The subset sizes are reduces by 0-8 %.

Table 5.9. Selected varieties with the selection-of-the-best rule *R3* and the assumption that the variety parameters represent a random sample from a Normal population, with heritability equal to 0.2; for $\hat{E}[P(CS)] = 0.90$, $\hat{E}[P(CS)] = 0.95$ and $\hat{E}[P(CS)] = 0.99$. The selection is performed on the basis of the 'mean performance' estimates of CRY, SC and WSY, successively. Between brackets the subset size is expressed relatively to the 110 new varieties.

CRY	$\hat{E}[P(CS)] = 0.90:$	79,84,91 (3 %).
	$\hat{E}[P(CS)] = 0.95:$	as for $\hat{E}[P(CS)] = 0.90$, + 85 (4%).
	$\hat{E}[P(CS)] = 0.99:$	as for $\hat{E}[P(CS)] = 0.95$.
SC	$\hat{E}[P(CS)] = 0.90:$	37,61,92,95,100,103,105,106 (7 %).
	$\hat{E}[P(CS)] = 0.95:$	as for $\hat{E}[P(CS)] = 0.90$, + 34,68 (9 %).
	$\hat{E}[P(CS)] = 0.99:$	as for $\hat{E}[P(CS)] = 0.95$, + 39,67,99,107 (13 %).
WSY	$\hat{E}[P(CS)] = 0.90:$	76,79,82,84,85,86,91,113 (7 %).
	$\hat{E}[P(CS)] = 0.95:$	as for $\hat{E}[P(CS)] = 0.90$, + 11,70,78,83,90 (12 %).
	$\hat{E}[P(CS)] = 0.99:$	as for $\hat{E}[P(CS)] = 0.95$, + 50,72,75,100,105 (16 %).

5.2.7 Evaluation of the case study

This case study shows that it is possible to successfully use statistical selection procedures in the plant breeding practice. The estimation part gave no difficulties.

Estimation in two steps is a clear procedure, but to use it on a routine basis requires suitable computer software. The calculation of the (pseudo-)variance/covariance matrix (divided by σ^2) of the variety parameter estimators is necessary to calculate the selection constants with the computer program SELCON. Selection rule *R3* is very convenient, because with this rule we do not have to calculate a separate selection constant for each variety. The actual selection can be performed with the use of the computer program SUBSET.

The fact that the sites do not contain the same varieties makes it difficult to compare the selected subsets at the various sites. Selection on the basis of the 'mean performance' estimates results in subsets with an acceptable size. The selected sugar beet varieties seem to be reasonable consistent, because they are often selected at all sites (where they were grown). The subset sizes can be reduced by altering the selection goal (selection of at least one good variety) or by making an extra assumption (superpopulation of the variety parameters). Whether the superpopulation assumption is acceptable depends on the selection phase. In an early phase such an assumption seems plausible. The least favourable configuration is of the utmost theoretical importance, but in practice it is very unlikely to occur. The superpopulation assumption is an attempt to get closer to reality. However, the $\hat{E}[P(CS)]$-statement is not as sharp as the P^*-statement.

Presentation of the estimates and the selection results in a scatter plot like Figure 5.1 is very enlightening. The experimenter can take in everything at a glance.

Selection of the smallest subset that includeds all varieties better than the average of the control varieties is not very successful. The subset sizes are very large, only the varieties that are really bad are discarded.

REFERENCES

Aubuchon, J.C., S.S. Gupta & J.C. Hsu (1985). Interfacing PROC RSMCB with the SAS System. *Technical Report* 313, Columbus, OH: Department of Statistics, Ohio State University.

Driessen, S.G.A.J. (1991) Multiple comparisons with and selection of the best treatment in (incomplete) block designs. *Communications in Statistics - Theory & Methods* 20(1), 179-217.

Edwards, H.P. (1985) RANKSEL - An interactive computer package of ranking and selection procedures. In : E.J. Dudewicz (Ed.) *The frontiers of modern statistical inference procedures*. American Sciences Press, 169-179.

CHAPTER 6

Discussion and conclusions

In this final chapter we will strike a balance w.r.t. the use of statistical selection procedures in the plant breeding practice. In **6.1** we discuss the previous chapters, followed by the final conclusions in **6.2**.

6.1 Discussion

Statistical selection deals with the selection of varieties with the aim that eventually (some of) these varieties are marketed. Here, 'varieties' can be interpreted liberally and stand for any reproducable genotypes. These can be pure lines of a self-pollinated crop, hybrids of such lines or clones of asexually propagated crops (Mayo, 1987). This type of selection is different from selection *within* a population of genotypes, with the aim to increase the level of the total population. In the latter type of selection also recombination of genes by mating is involved. This type of selection is described in many textbooks and articles about quantitative genetics (e.g. Falconer, 1986). Statistical selection procedures are concerned with selection *between* populations, and this type of selection is performed by plant breeding companies and also by the official variety testing authorities. For both, it is of crucial importance that the correct selection decisions are made. If a plant breeding company does not come up with varieties that are better than the existing ones, then the obvious result is that there is no income. It is decided by the official variety testing authorities whether new varieties are included in the official recommended list of agricultural crop varieties. The decision whether a new variety is or is not included in the official recommended variety list has large financial consequences for the owner of this variety. Therefore, the decisions taken by the official variety testing authorities should be statistically well founded.

The method of statistical selection is twofold. First, the varieties are grown in experiments at a number of sites and sometimes in two or more years. The observations from these experiments are described by a model, and the relevant model parameters are estimated. The selection is based on these estimates, so in order to select as good as possible the best linear unbiased estimators (BLUEs) have to be used to estimate the model parameters. Second, the actual selection is made with the use of a statistical selection rule. With such a selection rule, the probability of correct selection $P(CS)$ is controlled. The probability statement that accompanies a selection made with a statistical selection rule, is a quantitative measure about the uncertainty associated with that selection. Using statistical selection, the plant breeder gets acquainted with the uncertainty under which he is working.

The (contrasts between) variety values are often first estimated at the separate sites. Frequently, the fixed additive model is used to describe the observations at a single site. Then the estimation procedure is relatively simple. However, with today's computer facilities available, there are no reasons not to use the mixed additive model when this seems more appropriate. If the variance ratio σ_B^2/σ^2 becomes very large, the estimates obtained with the mixed additive model are equal to the estimates obtained with the fixed additive model. For the 'nice' designs, such as the complete block designs, both models lead towards the same estimates, regardless the variance ratio. However, these 'nice' designs are almost never found in practice, because different designs are used or because missing observations ruin the 'niceness' of the design. A possible problem of using estimators corresponding to the mixed additive model is giving a value for σ_B^2/σ^2. The knowledge about this variance ratio should be based on historical data. If one is not very sure about the value for this ratio, the estimates could be calculated for a series of ratio values. If the ranking of the varieties on the basis of their estimates is very different for different values of the variance ratio, one should be very cautious.

For the plant breeding practice, where often resolvable designs like lattice designs and alpha designs are used, a mixed additive model with fixed replication terms and random block within replication terms seems appropriate to describe the observations at a single site.

In literature, much attention has been given to the design and analysis of single experiments. However, in the plant breeding practice these experiments are not the most important ones. Of greater importance are the series of experiments. Until now, only limited attention has been given to the design and analysis of series of experiments. This is not right, because the ultimate selection is based on 'mean performance' estimates of the (contrasts between) variety values. With 'mean performance' estimates the BLUEs corresponding to a model for the joint observations of all experiments are meant. It is convenient if these estimates can be calculated in two steps. First, the estimates are calculated at the separate experiments, and next these estimates are combined into the 'mean performance' estimates. This estimation procedure is convenient if the estimation results of various research groups at different sites (countries) have to be combined. Each research group can perform the first estimation stage and thus reduce the amount of information before combining.

In **3.2** the estimation in case of a series of experiments has been elaborated for several models of the joint observations. If this model contains fixed interaction terms, then different parameters are estimated at the different sites, and the 'mean performance' estimates are weighted averages of the separate estimates. In all other cases the same parameters are being estimated at the various sites, and the 'mean performance' estimates are in general multivariately weighted averages of the local estimates. In case the model has additional random terms besides the error terms, the corresponding variance ratios of the variance of these random terms w.r.t. the error variance has to be given. There have been studies to estimate variance components in variety trials, e.g. in the United Kingdom (Talbot, 1984) and in Germany (Kienzl, 1975). Plant breeders should learn from their experiments (in the past) which variance ratios are appropriate for their situation. If one finds it difficult to pinpoint the values of the variance ratios, the 'mean performance' estimates can be calculated for a series of variance ratio values. With a fast computer this should not take too much time.

To be able to combine the local estimates in a univariate way, the experiments should be designed appropriately. In *3.2.2* we have seen that in case of a fixed additive model the **C** matrices of the individual experiments have to be proportional to each other, in order to calculate the 'mean performance' estimates as univariately weighted averages of the local estimates. However, missing values in an experiment result in an altered **C** matrix. Therefore, in that case multivariate weighting may become necessary to obtain the BLUEs.

Although concatenated trials are not very sophisticated from the statistical point of view, they are very useful from the practical point of view. For this very reason they are also used in sugar beet breeding, described in chapter 2. A subtrial, which is the only one to include a certain new variety, can be compared with a site at which a local variety is tested. The estimation of the BLUEs can be performed with the theory developed for a series of experiments. For certain experimental designs the local estimator is already the BLUE.

The estimation procedures proposed in chapter 3 make the calculation of the BLUEs feasible, especially because it are two step procedures. However, to perform these estimation procedures on a routine basis, useful computer software is necessary.

From the two basic approaches in statistical selection, subset selection seems to be the more useful one for the plant breeding practice. The other approach, Indifference Zone selection, will not be feasible when the number of tested varieties is large. However, Indifference Zone selection can perhaps successfully be used after reduction of the number of varieties by subset selection. More research on this topic seems necessary.

The random subset size is a property of subset selection procedures that is not appreciated by the plant breeder. However, this is the price he has to pay for not designing the experiments good enough to be able to select a fixed small number of varieties. Due to shortage of seed, money or experimental field the number of replications can only be limited. The randomness of the subset size reflects the idea that the conditions for selection are not always the same, but that in one season it is easier to select then in another season. This depends on the actual configuration of the variety parameters, but also on the standard errors of the BLUEs of the contrasts between variety values. Since these two can vary every season, it seems reasonable that the selection percentage is different from season to season.

At first sight, selection of the best variety seems to be the ideal of the plant breeder. However, we can ask the question whether this selection goal is not too exacting. It is probably more than the breeder wants. The negative result of such a strict selection goal is a large selected subset. Furthermore, it is easy to define 'best' for a single character, but for several characters simultaneously this becomes difficult. An index of characters would make things relatively easy, but the formulation of such an index requires much thought. Combining subsets selected on the basis of different characters does not seem a very useful alternative.

Selection of a minimum sized subset that includes at least one good variety seems to be a more realistic selection goal than selection of the best variety. If the best variety parameter is only a very small distance apart from the best but-one variety parameter, then a breeder does not care whether he selects the best or the almost best variety. The distance measure (δ^*) has to be given by the plant breeder, so he has to decide which distance between variety parameters is of practical importance. The use of this selection goal results in a smaller subset size than selection of the best variety. Therefore, this goal may be more convenient for practical use than selection of the best variety.

In the case study, selection of a subset that includes all varieties better than the average of the control varieties was not very useful. Due to the fact that many varieties were better than the controls, a very large subset size was obtained. Probably, the selection goal is not really suited for the plant breeding practice. A more useful selection goal from the practical point of view would be the following: select the smallest subset that includes at least one variety that is better than the average of the control varieties, with probability at least equal to P^*. However, this question has not been solved theoretically, and probably is impossible to solve.

The original subset selection procedures, which assumed the use of a completely randomised design or a randomised complete block design, were not very useful for the plant breeding practice, for the very reason that nowadays the mentioned designs are not often used in that field. A major step towards practical use of subset selection rules was made by the extension of the known theory to incomplete block designs by Driessen (1991). However, in general the subset selection rules he used contained separate selection constants for each variety. Although these selection constants are equal to each other for a lot of experimental designs, this is in general not convenient in practical use. The subset selection rules described in **4.2** that contain a single selection constant remove this barrier to practical use. To calculate the selection constant associated with these rules, computer simulation is indispensible. Our selection constant simulation program SELCON can be used to approximate the selection constants. With modern powerful and fast computers the use of simulation methods on a routine basis has become feasible. Therefore, this technique can also be used at plant breeding companies. The selection constants can also be calculated for the series of experiments and the concatenated trials described in chapter **3**. Thus, the subset selection rules can be used for experiments that are of practical importance.

If the additional assumption can be made that the variety parameters are a random sample from a superpopulation, we can use a subset selection procedure that makes use of this assumption. This selection procedure does not guarantee a minimum probability of correct selection, but aims at a subset that is associated with a certain estimated expected probability of correct selection. We hope that this expected probability of correct selection, using the superpopulation assumption, is closer to the real probability of correct selection than the minimum probability of correct selection P^*. We have seen that the true probability of correct selection is often much higher than P^*. This is to be expected, because the minimum probability of correct selection corresponds to the least favourable configuration, which is a configuration that probably is also least likely to occur. An approximate probability of correct selection is probably of more value to the plant breeder than an exact absolute minimum of this probability. Therefore, using a superpopulation assumption seems very useful.

The other proposed modification of subset selection procedures, namely using simultaneous lower bounds of ranked variety parameter contrasts, does not seem very useful for application on a routine basis.

The case study was the acid test for application of subset selection procedures in the plant breeding practice. To execute the proposed selection rules computer software is indispensible. To execute the rules with data that come from ordinary trials with incomplete block designs, series of experiments or concatenated trials, our computer program SUBSET can be used successfully. Using subset selection procedures, the breeder gets insight in the probability of correct selection and the number of varieties that have to be selected if this probability has to be larger than a prespecified value. This way the plant breeder knows what he is doing and the risks he is taking if he decides to select less varieties than indicated by the selection rule. For the final decision is up to the breeder...

6.2 Conclusions

We will first recapitulate the conclusions that were explicitly made in the sections of the previous chapters. Next we will give some general conclusions.

1) Differences between variety values are sufficient for statistical selection. In case the model used for the observations of a series of experiments includes fixed interaction terms, these differences are not equal to the differences between variety parameters. In all other models described in this thesis they are.

2) Calculation of the best linear unbiased estimates (BLUEs) for contrasts between variety values, using a model for the observations of a series of experiments, can be done in two steps of reduced size. First, the BLUEs are calculated at the individual experiments, next they are combined into the 'mean performance' BLUEs. Also the pseudo-variance/covariance matrix $\dot{\mathbf{D}}[\hat{\tau}]$ and the estimated variance s^2 can be calculated in steps.

3) The BLUEs for contrasts between variety values, calculated for a series of experiments, are in general multivariately weighted averages of the BLUEs calculated at the individual experiments. An exception is the situation where a model with fixed interaction terms is used. Then the 'mean performance' BLUEs are univariately weighted averages of the 'local' BLUEs.

4) The multivariate weight matrices $\mathbf{W}_k$, used to calculate the BLUEs for the contrasts $\mathbf{p}'\tau$ in a series of experiments, reduce to univariate weights w_k if $\mathbf{p}$ is a common eigenvector of all $\mathbf{W}'_k$ matrices with corresponding eigenvalue w_k. For the fixed additive model this is the case if the $\mathbf{C}$ matrices of the separate experiments are proportional to each other. This is e.g. the case if the experiments have variance-balanced designs.

5) In concatenated trials the interest lies in the estimation of the contrasts between the new variety parameters and the average of the control variety parameters. The BLUEs of these contrasts can be calculated in two steps. First, the desired contrasts are estimated at the separate subtrials. Next, the 'local' BLUEs from the substrials are combined into the BLUEs corresponding to the model for the joint observations of all subtrials.

6) The 'local' BLUEs for the contrasts between new variety parameters and the average of the control variety parameters in a concatenated trial are sometimes already equal to the BLUEs corresponding to the model for the joint observations of all subtrials. The exact conditions are described in *3.3.3*. If a subtrial has a variance-balanced design, then the above mentioned equality is true for the new varieties included in that subtrial.

7) Subset selection has to be preferred to selecting a fixed number of varieties, following the Finney approach (Finney, 1958). The latter approach aims at maximising the expected gain of selection. This aim, however, is questionable in variety testing. The probability that the best varieties are lost is not controlled in the Finney approach.

8) From the described subset selection rules, the rules that include the standard errors of the variety contrast estimators appear to be better than the rules that do not. Furthermore, for general use with all kinds of designs, the rules that contain a single selection constant are very useful and easy to work with. A practical recommendation then could be to use the selection rules that have a single selection constant and also make use of the standard errors of the variety contrast estimators.

9) Approximating selection constants by computer simulation is accurately enough for practical use. The accuracy of the results depend of course on the number of iterations used. We recommend a minimum of 10000 simulation rounds.

10) The additional assumption that the variety parameters are a random sample from a specified (Normal) superpopulation makes modification of subset rules possible. This modification results in a less rigid probability statement that can be very useful in practice.

11) Simulation can also be used to estimate the probability of correct selection, if the experiment is randomised. With lower bounds of the ranked variety parameters a confidence lower limit of the probability of correct selection can be calculated. This lower limit can be useful to the breeder in making his selection decisions.

12) Good computer software is indispensible to : a) calculate the estimates of the contrasts between variety values, the variance/covariance matrix and the variance estimate; b) calculate the selection constants; c) execute the selection rules. Software for b) and c) are our programmes SELCON and SUBSET, respectively.

General conclusions are :

13) Subset selection can play an important role in variety testing, both at the plant breeding companies and at the official variety testing authorities. The breeders should have more knowledge about the probability of correct selection.

14) Simulation is indispensible to calculate parameters that are necessary for the practical application of statistical selection procedures in experiments with incomplete block designs. With fast computers available nowadays, simulation methods should become part of the toolbox of the modern plant breeder.

15) With the proposed estimation procedures, subset selection rules and methods to calculate the selection constants, subset selection can successfully be used in the plant breeding practice. This answers the original objective of this research.

REFERENCES

Driessen, S.G.A.J. (1991). Multiple comparisons with and selection of the best treatment in (incomplete) block designs. *Communications in Statistics - Theory & Methods* 20(1), 179-217.

Falconer, D.S. (1986). *Introduction to quantitative genetics*. 2nd edition. Longman Scientific & Technical, Harlow.

Finney, D.J. (1958). Statistical problems in plant selection. *Bull. Inst. Int. Stat.* 36, 242-268.

Kienzl, H. (1975). Biometrische Characterisierung von Standorteinflüssen auf das Sortenversuchswesen; dargestellt an den Bayerischen Landessortenversuchen bei Winterweizen und Sommergerste in den Jahren 1963-1972. *Bayerisches Landwirtschaftliches Jahrbuch* Bnd 52 Heft 1, 18-86.

Mayo, O. (1987). *The theory of plant breeding*. 2nd. edition. Clarendon Press, Oxford.

Talbot, M. (1984). Yield variability of crop varieties in the U.K. *Journal of Agricultural Science*, Cambridge, 102, 315-321.

Summary

The ultimate goal of plant breeding is the development of new varieties. An important phase in the development process is testing and selecting potential new varieties. The varieties are tested by means of experiments at various sites, (sometimes) in several years. The observations from the experiments are usually modelled with a linear model. The best linear unbiased estimators (BLUEs) of certain estimable combinations of parameters (named variety values) in such a model are used to compare varieties mutually and with control varieties, and to make a selection. Selection means making decisions whether or not to discard particular varieties. The plant breeder certainly knows that these decisions are subject to uncertainty, but until now he did not have a quantitative measure for this uncertainty.

In this thesis we advocate the use of statistical theory in the selection phase of the breeding process. Here, statistical selection is split up into two components : 1) Estimating contrasts between variety values as good as possible, 2) Application of statistical selection procedures on these best estimates.

The results of this thesis are not restricted to a particular crop or breeding programme. However, we have studied the sugar beet breeding practice and use data from this field to illustrate our findings. In the sugar beet breeding programmes there are several stages where "varieties" are tested and selected. The plant breeder is interested in the specific varieties included in the variety trials and not in some population of varieties. Therefore, the variety terms in the models for the observations are chosen fixed.

Usually, an experiment at a particular site has incomplete blocks to take account of heterogeneity of the soil. With many varieties to be tested, the design of such an experiment is almost never balanced. Experiments are laid out at various sites (and sometimes in several years), but not every variety is tested at all sites. This results in a variety × site scheme that is not completely filled.

The selection decisions are based on several characters of the crop. This makes selection complicated, because all characters can seldomly be reduced into one selection index on which the decisions can be based. In case of sugar beets it is proposed to use the financial yield as selection index.

As said, we consider estimation of contrasts between variety values to be the first part of statistical selection. The value of a certain variety is defined in this thesis as a weighted average of the expectations of the observations corresponding to this variety. The corresponding linear combinations of model parameters can be estimated best using the least squares method. Although a breeder will base the selection decisions on variety values that include information from various sites (so-called mean performance values),

he is also interested in the variety values at the separate sites. For the observations at a single site, either a fixed additive model or a mixed additive model with random block terms is used.

The definition of the mean performance variety value depends on the model used. This model describes the joint observations of the series of experiments performed at different sites. The ranking of the varieties based on the estimates of their variety value can be different for different models. Therefore, the model has to be chosen with care. The model choice concerns questions as to whether model terms have to be considered fixed or random, and whether an additive or an interaction model has to be used.

A procedure to obtain the BLUEs of the mean performance variety values, without analysing the joint observations of all experiments, is proposed in this thesis. First, the experiments at the various sites are analysed individually and contrasts between variety values at these individual sites are estimated. Next, the BLUEs of the mean performance variety values can be calculated for models without fixed variety × site interaction terms as a multivariately weighted average of BLUEs calculated at the individual sites. For models that include fixed interaction terms the BLUEs must be calculated as a univariately weighted average of the 'local' BLUEs, with the weights given by the breeder.

For the situation of a fixed additive model for the joint observations of all sites, it is shown that for some experimental designs the multivariate weights reduce to univariate weights. This is e.g. the case when the **C** matrices (from the reduced normal equations) of the individual sites are proportional to each other. Regardless of the model or design used (as far as investigated), we can say that contrasts between (mean performance) variety values can be estimated (with BLUEs) in two steps of reduced size. Also the variance/covariance matrix of the estimators and the usual estimate of the error variance can be calculated in such a way.

The same principles can be applied to a so-called concatenated trial. Such a trial, located at one site, is subdivided into subtrials that include new varieties not grown in any other subtrial and control varieties grown in all subtrials. Here, meaningful contrasts to estimate are the differences between parameters of a new variety and the average of the parameters of the control varieties. The BLUEs of these contrasts can be calculated by combining local BLUEs from the subtrials. In certain cases the local BLUEs are already the 'overall' BLUEs.

The use of statistical selection procedures is considered to be the second component of statistical selection. In this thesis we pay much attention to subset selection rules. Using subset selection, the breeder selects a random sized subset of varieties. The subset size is chosen as small as possible, but large enough to guarantee that the probability of correct selection (i.e. the probability that the desired variety is included in the selected subset) is at least P^*, with P^* a predefined value. The desired variety can be the best variety (where 'best' must be defined) or a good variety (where 'good' must be defined), or maybe the desired varieties are all varieties better than a control variety. The subset is selected by means of a specified selection rule, which includes estimates of differences between variety values, the estimated variances of the corresponding estimators and so-called selection constants. The selection constants are associated with the experimental design used.

Often used in practice is the selection of a predetermined number of varieties. However, we have shown that this way the probability of correct selection cannot be controlled. This could mean that the desired variety is lost too often.

For unbalanced incomplete block designs selection constants had to be calculated for each variety. For practical use this is very inconvenient, because this type of designs is often used. Therefore, we have developed selection rules that only need a single selection constant, regardless the experimental design used. Such rules can only be used if the experiment is randomised, which means that the design has to be randomised and the actual varieties have to be assigned to the design varieties (numbers) by means of a defined randomisation process.

The calculation of the selection constants by numerical integration is only feasible for the situation of variance-balanced designs. In the other cases we can use computer simulation to approximate the desired selection constants. Our computer program SELCON performs this simulation, making it possible to calculate the selection constants for every experimental design. It appears that the simulation results are accurately enough to be used for practical purposes. Simulation can also successfully be used to calculate the probability of correct selection and the expected subset size, given the configuration of variety parameters. This can e.g. be used to compare different selection rules.

Using the variance/covariance matrix of the BLUEs of contrasts between mean performance variety values the selection constants can be approximated by simulation. So, the subset selection rules can also be used for the combined results of a series of experiments. The information that certain varieties are *a priori* excluded from selection affects the value of the selection constants and is therefore taken into account in the simulation program.

Two modifications of the subset selection procedures are proposed in order to make these procedures more useful in practice. In the first proposed modification it is assumed that the variety parameters represent a sample from a (Normal) superpopulation. This extra assumption leads to smaller selection constants and thus smaller subsets. However, now the expected probability of correct selection is controlled and not the minimum probability of correct selection, as is the case for the ordinary subset procedures. Subset selection procedures with the superpopulation assumption seem very useful for the plant breeding practice. The second proposed modification, namely the use of simultaneous lower bounds of the ranked variety parameter contrasts in order to calculate a confidence lower bound of the probability of correct selection, seems less practical.

To be able to execute the subset selection rules on a routine basis, software is needed. Therefore, we wrote the program SUBSET. This program executes subset selection rules using the selection constants as calculated by the program SELCON. In the output of SUBSET the breeder is informed about the uncertainties corresponding with certain selection decisions, and this enables him to make a well-considered selection. The developed theories and computer programs were successfully tested in a case study.

We finally reach the conclusion that statistical selection procedures, especially subset selection procedures, can successfully be used in the plant breeding practice.

Samenvatting

Over statistische selectie in de plantenveredeling

In de plantenveredeling wordt er uiteindelijk naar gestreefd nieuwe rassen te ontwikkelen. Een belangrijke fase in dit ontwikkelingsproces is het beproeven en selecteren van potentiële nieuwe rassen. De nieuwe rassen worden beproefd in experimenten, uitgevoerd op diverse locaties en (soms) in een aantal jaren. De waarnemingen uit deze proeven worden gewoonlijk gemodelleerd met behulp van een lineair model. De nauwkeurigste (i.e. beste) lineaire zuivere schatters (BLZSs) van bepaalde schatbare combinaties van parameters (raswaarden genoemd) in zo'n model worden gebruikt om de rassen onderling en met controle rassen te vergelijken, en om een selectie te maken. Selectie betekent het nemen van beslissingen omtrent het wel of niet weggooien van bepaalde rassen. De plantenveredelaar weet heel goed dat deze beslissingen behept zijn met onzekerheid, maar tot nu toe had hij geen kwantitatieve maat voor deze onzekerheid.

In dit proefschrift wordt een lans gebroken voor het gebruik van statistische theorie in de selectiefase van het veredelingsproces. Statistische selectie wordt daarbij opgesplitst in twee delen : 1) Het zo goed mogelijk schatten van contrasten tussen raswaarden, 2) Het toepassen van statistische selectieprocedures op deze schattingen.

De resultaten van dit proefschrift zijn niet slechts geldig voor één specifiek gewas of kweekprogramma. We hebben echter alleen de praktijk van de suikerbietenveredeling bestudeerd en gebruiken gegevens uit die praktijk om onze resultaten te illustreren. In het kweekprogramma van suikerbieten zijn verschillende fasen aanwezig waar "rassen" beproefd en geselecteerd worden. Daarbij is de veredelaar geïnteresseerd in de beproefde rassen zelf, en niet in een zekere achterliggende populatie van rassen. Daarom worden de rastermen in de gebruikte modellen als niet-stochastisch beschouwd.

In het algemeen zijn in een experiment op een bepaalde locatie incomplete blokken opgenomen om rekening te houden met de heterogeniteit van de bodem. Het grote aantal rassen dat beproefd moet worden zorgt ervoor dat het proefontwerp zelden gebalanceerd is. De rassenproeven worden uitgevoerd op verscheidene locaties (en soms in meerdere jaren), maar niet elk ras wordt op dezelfde locaties beproefd. Dit leidt tot een ras × locatie schema dat niet volledig gevuld is.

De selectiebeslissingen worden gebaseerd op verscheidene eigenschappen van het gewas. Dit maakt selectie gecompliceerd, want zelden kunnen alle eigenschappen gereduceerd worden tot één selectie-index op basis waarvan de beslissingen genomen kunnen worden. Er wordt voor suikerbieten voorgesteld om de financiële opbrengst als selectie-index te gebruiken.

De raswaarde van een bepaald ras wordt in dit proefschrift gedefinieerd als een gewogen gemiddelde van de verwachtingen van de waarnemingen behorende bij dit ras. De corresponderende lineaire combinaties van de modelparameters kunnen het beste geschat worden m.b.v. de kleinste kwadraten methode. De veredelaar zal de selectiebeslissingen baseren op raswaarden waarin informatie van meerdere locaties verwerkt is (zogenaamde gemiddelde raswaarden), maar hij heeft ook interesse in de raswaarden op de afzonderlijke locaties. Om de waarnemingen op één locatie te beschrijven wordt ofwel een vast additief model ofwel een gemengd additief model met stochastische bloktermen gebruikt.

De definitie van een gemiddelde raswaarde hangt af van het gekozen model, dat alle waarnemingen uit de experimenten op de verschillende locaties beschrijft. De rangschikking van de rassen, gebaseerd op hun raswaarden, kan verschillend zijn voor verschillende modellen. Daarom moet het model zorgvuldig gekozen worden. Daarbij moet besloten worden of modeltermen vast of stochastisch zijn en of een additief dan wel een interactie model gebruikt wordt.

In dit proefschrift wordt een procedure voorgesteld om de BLZSs van de gemiddelde raswaarden te verkrijgen, zonder dat alle waarnemingen van de experimenten te zamen geanalyseerd moeten worden. Eerst worden de experimenten op de diverse locaties afzonderlijk geanalyseerd en worden contrasten tussen raswaarden op deze locaties geschat. Vervolgens kunnen de BLZSs van de gemiddelde raswaarden voor modellen zonder vaste ras × locatie interactietermen berekend worden als een multivariaat gewogen gemiddelde van BLZSs die op de afzonderlijke locaties berekend werden. Als het model vaste ras × locatie interactietermen bevat, dan moeten de BLZSs berekend worden als een univariaat gewogen gemiddelde van de 'plaatselijke' BLZSs, waarbij de gewichten gegeven moeten worden door de kweker.

In geval van een vast additief model voor de gezamenlijke waarnemingen van alle locaties is aangetoond dat voor sommige proefschema's de multivariate gewichten reduceren tot univariate gewichten. Dit is bijvoorbeeld zo wanneer de **C** matrices (afkomstig van de gereduceerde normaalvergelijkingen) van de afzonderlijke locaties proportioneel zijn t.o.v. elkaar. Onafhankelijk van het gebruikte model of proefschema (voor zover onderzocht) kunnen we zeggen dat contrasten tussen (gemiddelde) raswaarden in twee stappen van beperkte omvang geschat kunnen worden (met BLZSs). Ook de variantie/covariantie matrix van de schatters en de gebruikelijke schatting van de restvariantie kunnen op zo'n manier berekend worden.

Dezelfde principes kunnen toegepast worden op een zogenaamde aaneengeschakelde proef. Zo'n proef, die uitgevoerd wordt op één locatie, is onderverdeeld in deelproeven die nieuwe rassen bevatten welke niet opgenomen zijn in enig andere deelproef en controle rassen die in elke deelproef voorkomen. Zinvolle te schatten contrasten zijn hier verschillen tussen parameters van nieuwe rassen en het gemiddelde van de parameters behorende bij de controle rassen. De BLZSs van deze contrasten kunnen berekend worden door plaatselijke BLZSs uit de deelproeven te combineren.

Als tweede component van statistische selectie werd het gebruik van statistische selectieprocedures genoemd. In dit proefschrift schenken we veel aandacht aan 'subset'-selectieregels. Bij 'subset'-selectie selecteert de veredelaar een 'subset' van rassen. De 'subset'-omvang is stochastisch : zo klein mogelijk, maar groot genoeg om te garanderen dat de kans op correcte selectie (dit is de kans dat het gewenste ras in de 'subset' aanwezig is) minstens P^* is, met P^* vooraf gegeven. Het gewenste ras kan het beste ras zijn (waarbij 'beste' gedefinieerd moet worden), of een goed ras (waarbij 'goed' gedefinieerd moet worden). Of misschien zijn de gewenste rassen alle rassen die beter zijn dan een controle ras. De 'subset' wordt geselecteerd m.b.v. een zekere selectieregel, waarin schattingen van verschillen tussen raswaarden, de geschatte varianties van de corresponderende schatters en de zogenaamde selectieconstanten verwerkt zijn. De waarden van de selectieconstanten worden bepaald door het gebruikte proefontwerp.

Vaak wordt in de praktijk een vooraf vastgesteld aantal rassen geselecteerd. Er is echter aangetoond dat de kans op correcte selectie op deze manier niet onder controle gehouden kan worden, hetgeen kan betekenen dat het gewenste ras te vaak verloren gaat.

Voor ongebalanceerde proefschema's moesten de selectieconstanten voor elk ras afzonderlijk berekend worden. Dit is erg onplezierig voor de praktijk, omdat daar veel van dit type proefschema's gebruik gemaakt wordt. Daarom hebben we selectieregels ontwikkeld die slechts één enkele selectieconstante vragen, ongeacht het proefontwerp. Zulke selectieregels kunnen alleen gebruikt worden bij gewarde experimenten. Zo'n experiment heeft een gewarde proefopzet en de werkelijke rassen zijn toegewezen aan de ontwerprassen (getallen) d.m.v. een gedefinieerde warringsprocedure.

Het berekenen van de selectieconstanten m.b.v. numerieke integratie is alleen haalbaar voor variantie-gebalanceerde proefontwerpen. In de overige gevallen kunnen we computersimulatie gebruiken om de selectieconstanten te benaderen. Met ons computerprogramma SELCON kunnen de selectieconstanten voor alle mogelijke proefontwerpen m.b.v simulatie berekend worden. Het blijkt dat de simulatieresultaten nauwkeurig genoeg zijn om in de praktijk gebruikt te kunnen worden. Simulatie is ook een bevredigende methode om de kans op correcte selectie en de verwachte 'subset'grootte te berekenen (bij een gegeven configuratie van rasparameters). Deze grootheden kunnen bijvoorbeeld gebruikt worden om verschillende selectieregels onderling te vergelijken.

Gebruikmakend van de variantie/covariantie matrix van de BLZSs van contrasten tussen gemiddelde raswaarden kunnen de selectieconstanten benaderd worden m.b.v. simulatie. Dus kunnen de 'subset'-selectieregels ook gebruikt worden voor de gecombineerde resultaten van een reeks experimenten. *A priori* informatie dat bepaalde rassen uitgesloten moeten worden van selectie beïnvloedt de grootte van de selectieconstante. Daarom wordt daar rekening mee gehouden in het simulatieprogramma.

Om 'subset'-selectieprocedures meer bruikbaar te maken in de veredelingspraktijk worden twee aanpassingen van deze procedures voorgesteld. In het eerste voorstel wordt ervan uitgegaan dat de rasparameters een steekproef uit een (Normale) superpopulatie vormen. Deze extra veronderstelling leidt tot kleinere selectieconstanten en dus tot kleinere 'subset'-groottes. Hierbij wordt echter de verwachte kans op correcte selectie gecontroleerd en niet de minimum kans op correcte selectie, hetgeen het geval is bij de gebruikelijke 'subset'-procedures. 'Subset'-selectieprocedures met een superpopulatie veronderstelling lijken erg zinvol voor de veredelingspraktijk. Het tweede voorstel tot modificatie van 'subset'-selectie blijkt minder praktisch. Dit voorstel betreft het gebruik van simultane ondergrenzen voor de gerangschikte rasparametercontrasten teneinde een betrouwbaarheidsondergrens voor de kans op correcte selectie te berekenen.

Om 'subset'-selectieregels routinematig uit te kunnen voeren is computerprogrammatuur nodig. Daartoe werd het programma SUBSET geschreven. Dit programma voert 'subset'-selectieregels uit, gebruikmakend van selectieconstanten die door het programma SELCON berekend werden. In de uitvoer van SUBSET wordt de veredelaar geïnformeerd omtrent de onzekerheden behorende bij bepaalde selectiebeslissingen. Dit geeft hem de mogelijkheid om een weloverwogen selectie te maken. De ontwikkelde theorieën en computerprogramma's werden met succes beproefd in een praktijkstudie.

Tenslotte kan geconcludeerd worden dat statistische selectieprocedures, in het bijzonder 'subset'-selectieprocedures, met succes gebruikt kunnen worden in de praktijk van de plantenveredeling.

Curriculum vitae

Cornelis Johannes Dourleijn werd op 24 januari 1965 geboren te West-Souburg (thans gemeente Vlissingen). In 1983 behaalde hij het diploma Ongedeeld VWO aan de Chr. Scholengemeenschap Walcheren te Middelburg. In datzelfde jaar ging hij Plantenveredeling studeren aan de Landbouwuniversiteit (toen nog Landbouwhogeschool geheten) te Wageningen. In september 1988 studeerde hij met lof af, met als afstudeervakken Plantenveredeling, Wiskundige Statistiek en Akkerbouw. Direct daarna begon hij als assistent in opleiding (AIO) bij de sectie Wiskundige en Toegepaste Statistiek van de vakgroep Wiskunde van de Landbouwuniversiteit, met als onderzoeksthema de toepasbaarheid van statistische rangschikkings- en selectieprocedures. Het in deze periode verrichte onderzoek is beschreven in dit proefschrift.

Sinds september 1992 is Johan werkzaam als Universitair Docent bij de vakgroep Wiskunde van de Landbouwuniversiteit Wageningen.