Izak Bos
Peter Caligari

Selection Methods in Plant Breeding

2nd Edition

Springer

Selection Methods in Plant Breeding

Selection Methods in Plant Breeding

2nd Edition

by

Izak Bos
University of Wageningen,
The Netherlands

and

Peter Caligari
University of Talca,
Chile

Springer

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-6369-5 (HB)
ISBN 978-1-4020-6370-1 (e-book)

Published by Springer,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

www.springer.com

Cover photo: Bagging of the inflorescence of an oil palm

Printed on acid-free paper

Contents

Preface

Selection procedures used in plant breeding have gradually developed over a very long time span, in fact since settled agriculture was first undertaken. Nowadays these procedures range from very simple mass selection methods, sometimes applied in an ineffective way, to indirect trait selection based on molecular markers. The procedures differ in costs as well as in genetic efficiency. In contrast to the genetic efficiency, costs depend on the local conditions encountered by the breeder. The genetic progress per unit of money invested varies consequently from site to site. This book considers consequently only the genetic efficiency, *i.e.* the rate of progress to be expected when applying a certain selection procedure.

If a breeder has a certain breeding goal in mind, a selection procedure should be chosen. A wise choice requires a wellfounded opinion about the response to be expected from any procedure that might be applied. Such an opinion should preferably be based on the most appropriate model when considering the crop and the trait (or traits) to be improved. Sometimes little knowledge is available about the genetic control of expression of the trait(s). This applies particularly in the case of quantitative variation in the traits. It is, therefore, important to be familiar with methods for the elucidation of the inheritance of the traits of interest. This means, in fact, that the breeder should be able to develop population genetic and quantitative genetic models that describe the observed mode of inheritance as satisfactorily as possible.

The genetic models are generally based, by necessity, on simplifying assumptions. Quite often one assumes:

- a diploid behaviour of the chromosomes;
- an independent segregation of the pairs of homologous chromosomes at meiosis, or, more rigorously, independent segregation of the alleles at the loci controlling the expression of the considered trait;
- independence of these alleles with regard to their effects on the expression of the trait;
- a regular mode of reproduction within plants as well as among plants belonging to the same population; and/or
- the presence of not more than two alleles per segregating locus.

Such simplifying assumptions are made as a compromise between, on the one hand, the complexity of the actual genetic control, and, on the other hand, the desire to keep the model simple. Often such assumptions can be tested and so validated or revoked, but, of course, as the assumptions deviate more from the real situation, decisions made on the basis of the model will be less appropriate.

The decisions concern choices with regard to:

- selection methods, e.g. mass selection versus half sib family selection;
- selection criteria, e.g. grain yield per plant versus yield per ear;
- experimental design, e.g. testing of each of N candidates in a single plot versus testing each of only $\frac{1}{2}N$ candidates in two plots; or
- data adjustment, e.g. moving mean adjustment versus adjustment of observations on the basis of observations from plots containing a standard variety.

In fact such decisions are often made on disputable grounds, such as experience, tradition, or intuition. This explains why breeders who deal in the same region with the same crop work in divergent ways. Indeed, their breeding goals may differ, but these goals themselves are often based on a subjective judgement about the ideotype (ideal type of plant) to be pursued.

In this book, concepts from plant breeding, population genetics, quantitative genetics, probability theory and statistics are integrated. The reason for this is to help provide a basis on which to make selection more professional, in such a way that the chance of being successful is increased. Success can, of course, never be guaranteed because the best theoretical decision will always be made on the basis of incomplete and simplifying assumptions. Nevertheless, the authors believe that a breeder familiar with the contents of this book is in a better position to be successful than a breeder who is not!

Preface to the Second Edition

New and upgraded paragraphs have been added throughout this edition. They have been added because it was felt, when using the first edition as a course book, that many parts could be improved according to a didactical point of view. It was, additionally, felt that – because of the increasing importance of molecular markers – more attention had to be given the use of markers (Section 12.3.2). In connection with this, quantitative genetic theory has, compared to the first edition, been more extensively developed for loci represented by multiple alleles (Sections 8.3.3 and 8.3.4).

It was stimulating to receive suggestions from interested readers. These suggestions have given rise to many improvements. Especially the many and useful suggestions from Ir. Ed G.J. van Paassen, Ir. Joël Schwarz, Dr. Hans-Peter Piepho, Dr. Mohamed Mahdi Sohani and Dr. L.R. Verdooren are acknowledged.

Chapter 1
Introduction

This chapter provides an overview of basic concepts and statistical tools underlying the development of population and quantitative genetics theory. These branches of genetics are of crucial importance with regard to the understanding of equilibria and shifts in (i) the genotypic composition of a population and (ii) the mean and variation exhibited by the population. In order to keep the theory to be developed manageable, two assumptions are made throughout the book, i.e. absence of linkage and absence of epistasis. These assumptions concern traits with quantitative variation.

Knowledge of population genetics, quantitative genetics, probability theory and statistics is indispensable for understanding equilibria and shifts with regard to the genotypic composition of a population, its mean value and its variation.

The subject of **population genetics** is the study of equilibria and shifts of allele and genotype frequencies in populations. These equilibria and shifts are determined by five forces:

- Mode of reproduction of the considered crop
  The **mode of reproduction** is of utmost importance with regard to the breeding of any particular crop and the maintenance of already available varieties. This applies both to the natural mode of reproduction of the crop and to enforced modes of reproduction, like those applied when producing a hybrid variety. In plant breeding theory, crops are therefore classified into the following categories: cross-fertilizing crops (Chapter 2), self-fertilizing crops (Chapter 3), crops with both cross- and self-fertilization (Section 3.4) and asexually reproducing crops. In Section 2.1 it is explained that even within a specific population, traits may differ with regard to their mode of reproduction. This is further elaborated in Chapter 4.
- Selection (Chapters 6 and 12)
- Mutation (Section 6.2)
- Immigration of plants or pollen, *i.e.* immigration of alleles (Section 6.2)
- Random variation of allele frequencies (Chapter 7)

A **population** is a group of (potentially) interbreeding plants occurring in a certain area, or a group of plants originating from one or more common ancestors. The former situation refers to cross-fertilizing crops (in which case the term **Mendelian population** is sometimes used), while the latter group concerns, in particular, self-fertilizing crops. In the absence of immigration the population is said to be a **closed population**. Examples of closed populations are

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 1–5.

- A group of plants belonging to a cross-fertilizing crop, grown in an isolated field, *e.g.* maize or rye (both pollinated by wind), or turnips or Brussels sprouts (both pollinated by insects)
- A collection of lines of a self-fertilizing crop, which have a common origin, *e.g.* a single-cross, a three-way cross, a backcross

The subject of **quantitative genetics** concerns the study of the effects of alleles and genotypes and of their interaction with environmental conditions.

Population genetics is usually concerned with the probability distribution of genotypes within a population **(genotypic composition)**, while quantitative genetics considers phenotypic values (and statistical parameters dealing with them, especially mean and variance) for the trait under investigation. In fact population genetics and quantitative genetics are applications of **probability theory** in genetics. An important subject is, consequently, the derivation of probability distributions of genotypes and the derivation of expected genotypic values and of variances of genotypic values. Generally, statistical analyses comprise estimation of parameters and hypothesis testing. In quantitative genetics **statistics** is applied in a number of ways. It begins when considering the experimental design to be used for comparing entries in the breeding programme. Section 11.2 considers the estimation of interesting quantitative genetic parameters, while Chapter 12 deals with the comparison of candidates grown under conditions which vary in a trend.

Considered across the entries constituting a population (plants, clones, lines, families) the expression of an observed trait is a random variable. If the expression is represented by a numerical value the variable is generally termed **phenotypic value**, represented by the symbol $\underline{p}$.

Note 1.1 In this book random variables are underlined.

Two genetic causes for variation in the expression of a trait are distinguished. Variation controlled by so-called **major genes**, *i.e.* alleles that exert a readily traceable effect on the expression of the trait, is called **qualitative variation**. Variation controlled by so-called **polygenes**, *i.e.* alleles whose individual effects on a trait are small in comparison with the total variation, is called **quantitative variation**. In Note 1.2 it is elaborated that this classification does not perfectly coincide with the distinction between **qualitative traits** and **quantitative traits**.

The former paragraph suggests that the term *gene* and *allele* are synonyms. According to Rieger, Michaelis and Green (1991) a **gene** is a continuous region of DNA, corresponding to one (or more) transcription units and consisting of a particular sequence of nucleotides. Alternative forms of a particular gene are referred to as **alleles**. In this respect the two terms 'gene' and 'allele' are sometimes interchanged. Thus the term 'gene frequency' is often used instead of the term 'allele frequency'. The term **locus** refers to the site, alongside a chromosome, of the gene/allele. Since the term 'gene' is often used as a synonym of the term 'locus', we have tried to avoid confusion by preferential

use of the terms 'locus' and 'allele' (as a synonym of the word gene) where possible.

In the case of qualitative variation, the phenotypic value $\underline{p}$ of an entry (plant, line, family) belonging to a genetically heterogeneous population is a **discrete random variable**. The phenotype is then exclusively (or to a largely traceable degree) a function f of the genotype, which is also a random variable $\underline{\mathcal{G}}$. Thus

$$\underline{p} = f(\underline{\mathcal{G}})$$

It is often desired to deduce the genotype from the phenotype. This is possible with greater or lesser correctness, depending for example on the degree of dominance and sometimes also on the effect of the growing conditions on the phenotype. A knowledge of population genetics suffices for an insight into the dynamics of the genotypic composition of a population with regard to a trait with qualitative variation: application of quantitative genetics is then superfluous.

Note 1.2 All traits can show both qualitative and quantitative variation. Culm length in cereals, for instance, is controlled by dwarfing genes with major effects, as well as by polygenes. The commonly used distinction between qualitative traits and quantitative traits is thus, strictly speaking, incorrect. When exclusively considering qualitative variation, *e.g.* with regard to the traits in pea (*Pisum sativum*) studied by Mendel, this book describes the involved trait as a trait showing qualitative variation. On the other hand, with regard to traits where quantitative variation dominates – and which are consequently mainly discussed in terms of this variation – one should realize that they can also show qualitative variation. In this sense the following economically important traits are often considered to be 'quantitative characters':

- Biomass
- Yield with regard to a desired plant product
- Content of a desired chemical compound (oil, starch, sugar, protein, lysine) or an undesired compound
- Resistance, including components of partial resistance, against biotic or abiotic stress factors
- Plant height

In the case of quantitative variation $\underline{p}$ results from the interaction of a **complex genotype**, *i.e.* several to many loci are involved, and the specific growing conditions are important. In this book, by complex genotype we mean the sum of the genetic constitutions of all loci affecting the expression of the considered trait. These loci may comprise loci with **minor genes** (or polygenes), as well as loci with major genes, as well as loci with both. With regard to a trait showing quantitative variation, it is impossible to classify individual plants, belonging to a genetically heterogeneous population, according to their

genotypes. This is due to the number of loci involved and the complicating effect on $\underline{p}$ of (some) variation in the quality of the growing conditions. It is, thus, impossible to determine the number of plants representing a specified complex genotype. (With regard to the expression of qualitative variation this may be possible!). Knowledge of both population genetics and quantitative genetics is therefore required for an insight into the inheritance of a trait with quantitative variation.

The phenotypic value for a quantitative trait is a **continuous random variable** and so one may write

$$\underline{p} = f(\underline{\mathcal{G}}, \underline{e})$$

Thus the phenotypic value is a function f of both the complex genotype (represented by $\underline{\mathcal{G}}$) and the quality of the growing conditions (say **environment**, represented by $\underline{e}$). Even in the case of a genetically homogeneous group of plants (a clone, a pure line, a single-cross hybrid) $\underline{p}$ is a continuous random variable. The genotype is a constant and one should then write

$$\underline{p} = f(\mathcal{G}, \underline{e})$$

Regularly in this book, simplifying assumptions will be made when developing quantitative genetic theory. Especially the following assumptions will often be made:

(i) Absence of linkage of the loci controlling the studied trait(s)
(ii) Absence of epistatic effects of the loci involved in complex genotypes.

These assumptions will now be considered.

Absence of linkage

The assumption of absence of linkage for the loci controlling the trait of interest, *i.e.* the assumption of independent segregation, may be questionable in specific cases, but as a generalisation it can be justified by the following reasoning.

Suppose that each of the n chromosomes in the genome contains M loci affecting the considered trait. This implies presence of n groups of $\binom{M}{2}$ pairs of loci consisting of loci which are more strongly or more weakly linked. The proportion of pairs consisting of linked loci among all pairs of loci amounts then to

$$\frac{n\binom{M}{2}}{\binom{nM}{2}} = \frac{n.M!}{2!(M-2)!} \times \frac{2!(nM-2)!}{(nM)!} = \frac{M-1}{nM-1} = \frac{1-\frac{1}{M}}{n-\frac{1}{M}}$$

For $M = 1$ this proportion is 0; for $M = 2$ it amounts to 0.077 for rye (*Secale cereale*, with $n = 7$) and to 0.024 for wheat (*Triticum aestivum*, with $n = 21$);

for $M = 3$ it amounts to 0.100 for rye and to 0.032 for wheat. For $M \to \infty$ the proportion is $\frac{1}{n}$; *i.e.* 0.142 for rye and 0.048 for wheat.

One may suppose that loci located on the same chromosome, but on different sides of the centromere, behave as unlinked loci. If each of the n chromosomes contains $m(=\frac{1}{2}M)$ relevant loci on each of the two arms then there are $2n$ groups of $\binom{m}{2}$ pairs consisting of linked loci. Thus considered, the proportion of pairs consisting of linked loci amounts to

$$\frac{2n\binom{m}{2}}{\binom{2nm}{2}} = \frac{2n.m!}{2!(m-2)!} \times \frac{2!(2nm-2)!}{(2nm)!} = \frac{1-\frac{1}{m}}{2n-\frac{1}{m}}$$

For $m = 1$ this proportion is 0; for $m = 2$ it amounts to 0.037 for rye and to 0.012 for wheat; for $m = 3$ it amounts to 0.049 for rye and to 0.016 for wheat. For $m \to \infty$ the proportion is $\frac{1}{2n}$; *i.e.* 0.071 for rye and 0.024 for wheat.

For the case of an even distribution across all chromosomes of the polygenic loci affecting the considered trait it is concluded that the proportion of pairs of linked loci tends to be low. (In an autotetraploid crop the chromosome number amounts to $2n = 4x$. The reader might like to consider what this implies for the above expressions.)

Absence of epistasis

Absence of **epistasis** is another assumption that will be made regularly in this book, notably in Sections 8.3.2 and 10.1. It implies **additivity** of the effects of the single-locus genotypes for the loci affecting the level of expression for the considered trait. The genotypic value of some complex genotype consists then of the sum of the genotypic value of the complex genotype with regard to all non-segregating loci, here represented by m, as well as the sum of the contributions due to the genotypes for each of the K segregating polygenic loci B_1-$b_1, \ldots, B_K$-b_K. Thus

$$\mathcal{G}_{B_1\text{-}b_1,\ldots,B_K\text{-}b_K} = m + \mathcal{G}'_{B_1\text{-}b_1} + \ldots + \mathcal{G}'_{B_K\text{-}b_K} \tag{1.1}$$

where $\mathcal{G}'$ is defined as the contribution to the genotypic value, relative to the population mean genotypic value, due to the genotype for the considered locus (Section 8.3.3). The assumption implies the absence of **inter-locus interaction**, *i.e.* the absence of **epistasis** (in other words: absence of **non-allelic interaction**). It says that the effect of some genotype for some locus $B_i - b_i$ in comparison to another genotype for this same locus does not depend at all on the complex genotype determined by all other relevant loci.

In this book, in order to clarify or substantiate the main text, theoretical examples and results of actual experiments are presented. Notes provide short additional information and appendices longer, more complex supplementary information or mathematical derivations.

Chapter 2
Population Genetic Effects of Cross-fertilization

Cross-fertilization produces populations consisting of a mixture of plants with a homozygous or heterozygous (complex) genotype. In addition, the effects of a special form of cross-fertilization, i.e. panmixis, are considered. It is shown that continued panmixis leads sooner or later to a genotypic composition which is completely determined by the allele frequencies. The allele frequencies do not change in course of the generations but the haplotypic and genotypic composition may change considerably. This process is described for diploid and autotetraploid crops.

2.1 Introduction

There are several mechanisms promoting cross-pollination and, consequently, cross-fertilization. The most important ones are

- **Dioecy**, *i.e.* male and female gametes are produced by different plants.

Asparagus	*Asparagus officinalis* L.
Spinach	*Spinacia oleracea* L.
Papaya	*Carica papaya* L.
Pistachio	*Pistacia vera* L.
Date palm	*Phoenix dactylifera* L.

- **Monoecy**, *i.e.* male and female gametes are produced by separate flowers occurring on the same plant.

Banana	*Musa* spp.
Oil palm	*Elaeis guineensis* Jacq.
Fig	*Ficus carica* L.
Coconut	*Cocos nucifera* L.
Maize	*Zea mays* L.
Cucumber	*Cucumis sativus* L.

In musk melon (*Cucumis melo* L.) most varieties show **andromonoecy**, *i.e.* the plants produce both staminate flowers and bisexual flowers, whereas other varieties are monoecious.

- **Protandry**, *i.e.* the pollen is released before receptiveness of the stigmata.

Leek	*Allium porrum* L.
Onion	*Allium cepa* L.

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 7–32.

Carrot	*Daucus carota* L.
Sisal	*Agave sisalana* Perr.

- **Protogyny**, *i.e.* the stigmata are receptive before the pollen is released.

Tea	*Camellia sinensis* (L.) O. Kuntze
Avocado	*Persea americana* Miller
Walnut	*Juglans nigra* L.
Pearl millet	*Pennisetum typhoides* L. C. Rich.

- **Self-incompatibility**, *i.e.* a physiological barrier preventing normal pollen grains fertilizing eggs produced by the same plant.

Cacao	*Theobroma cacao* L.
Citrus	*Citrus* spp.
Tea	*Camellia sinensis* L. O. Kuntze
Robusta coffee	*Coffea canephora* Pierre ex Froehner
Sugar beets	*Beta vulgaris* L.
Cabbage, kale	*Brassica oleracea* spp.
Rye	*Secale cereale* L.
Many grass species,	*e.g.* perennial ryegrass (*Lolium perenne* L.)

- **Flower morphology**

Fig	*Ficus carica* L.
Primrose	*Primula veris* L.
Common buckwheat	*Fagopyrum esculentum* Moench.
and probably in the Bird of Paradise flower	*Strelitzia reginae* Banks

Effects with regard to the haplotypic and genotypic composition of a population due to (continued) reproduction by means of **panmixis** will now be derived for a so-called **panmictic population**. Panmictic reproduction occurs if each of the next five conditions apply:

(i) Random mating
(ii) Absence of random variation of allele frequencies
(iii) Absence of selection
(iv) Absence of mutation
(v) Absence of immigration of plants or pollen

In the remainder of this section the first two features of panmixis are more closely considered.

Random mating
Random mating is defined as follows: in the case of random mating the fusion of gametes, produced by the population as a whole, is at random with regard to the considered trait. It does not matter whether the mating occurs by means of crosses between pairs of plants combined at random, or by means of open pollination.

Open pollination in a population of a cross-fertilizing (**allogamous**) crop may imply random mating. This depends on the trait being considered. One should thus be careful when considering the mating system. This is illustrated in Example 2.1.

Example 2.1 Two types of rye plants can be distinguished with regard to their epidermis: plants with and plants without a waxy layer. It seems justifiable to assume random mating with regard to this trait. With regard to time of flowering, however, the assumption of random mating may be incorrect. Early flowering plants will predominantly mate *inter se* and hardly ever with late flowering plants. Likewise late flowering plants will tend to mate with late flowering plants and hardly ever with early flowering ones. With regard to this trait, so-called **assortative mating** (see Section 4.1) occurs.

One should, however, realize that the ears of an individual rye plant are produced successively. The assortative mating with regard to flowering date may thus be far from perfect. Also, with regard to traits controlled by loci linked to the locus (or loci) controlling incompatibility, *e.g.* in rye or in meadow fescue (*Festuca pratensis*), perfect random mating will therefore probably not occur.

Selection may interfere with the mating system. Plants that are resistant to an agent (*e.g.* disease or chemical) will mate *inter se* (because susceptible plants are eliminated). Then assortative mating occurs due to selection.

Crossing of neighbouring plants implies random mating if the plants reached their positions at random; crossing of contiguous inflorescences belonging to the same plant (**geitonogamy**) is, of course, a form of selfing.

Random mating does not exclude a fortuitous relationship of mating plants. Such relationships will occur more often with a smaller population size. If a population consists, generation after generation, of a small number of plants, it is inevitable that related plants will mate, even when the population is maintained by random mating. Indeed, mating of related plants yields an increase in the frequency of homozygous plants, but in this situation the increase in the frequency of homozygous plants is also due to another cause: fixation occurs because of non-negligible random variation of allele frequencies. Both causes of the increase in homozygosity are due to the small population size (and not to the mode of reproduction).

This ambiguous situation, so far considered for a single population, occurs particularly when numerous small **subpopulations** form together a large **superpopulation**. In each subpopulation random mating, associated with non-negligible random variation of the allele frequencies, may occur, whereas in the superpopulation as a whole inbreeding occurs. Example 2.2 provides an illustration.

Example 2.2 A large population of a self-fertilizing crop, *e.g.* an F_2 or an F_3 population, consists of numerous subpopulations each consisting of a single plant. Because the gametes fuse at random with regard to any trait, one may state that random mating occurs within each subpopulation. At the level of the superpopulation, however, selfing occurs.

Selfing is impossible in dioecious crops, *e.g.* spinach (*Spinacia oleracea*). Inbreeding by means of continued sister × brother crossing may then be applied. This full sib mating at the level of the superpopulation may imply random mating within subpopulations consisting of full sib families (see Section 3.1).

Seen from the level of the superpopulation, inbreeding occurs if related plants mate preferentially. This may imply the presence of subpopulations, reproducing by means of random mating. If very large, the superpopulation will retain all alleles. The increasing homozygosity rests on gene fixation in the subpopulations. If, however, only a single full sib family produces offspring by means of open pollination, implying crossing of related plants, then the population as a whole (in this case just a single full sib family) is still said to be maintained by random mating.

Absence of random variation of allele frequencies
The second characteristic of panmixis is absence of random variation of allele frequencies from one generation to the next. This requires an infinite **effective size** of the population, originating from an infinitely large sample of gametes produced by the present generation. Panmixis thus implies a **deterministic model**. In populations consisting of a limited number of plants, the allele frequencies vary randomly from one generation to the next. Models describing such populations are **stochastic models** (Chapter 7).

2.2 Diploid Chromosome Behaviour and Panmixis

2.2.1 One Locus with Two Alleles

The majority of situations considered in this book involve a locus represented by not more than two alleles. This is certainly the case in diploid species in the following populations:

- Populations tracing back to a cross between two pure lines, say, a single cross
- Populations obtained by (repeated) backcrossing (if, indeed, both the donor and the recipient have a homozygous genotype)

It is possibly the case in populations tracing back to a three-way cross or a double cross. It is improbable in other populations, like populations of

cross-fertilizing crops, populations tracing back to a complex cross, landraces, multiline varieties.

To keep (polygenic) models simple, it will often be assumed that each of the considered loci is represented by only two alleles. Quite often this simplification will violate reality. The situation of multiple allelic loci is explicitly considered in Sections 2.2.2 and 8.3.3.

If the expression for the trait of interest is controlled by a locus with two alleles A and a (say locus A-a) then the probability distribution of the genotypes occurring in the considered population is often described by

	Genotype		
	aa	Aa	AA
Probability	f_0	f_1	f_2

One may represent the probability distribution (in this book mostly the term **genotypic composition** will be used) by the row vector (f_0, f_1, f_2). The symbol f_j represents the probability that a random plant contains j A-alleles in its genotype for locus A-a, where j may be equal to 0, 1 or 2. It has become custom to use the word **genotype frequency** to indicate the probability of a certain genotype and for that reason the symbol f is used.

The plants of the described population produce gametes which have either haplotype a or haplotype A. (Throughout this book the term **haplotype** is used to indicate the genotype of a gamete.) The probability distribution of the haplotypes of the gametes produced by the population is described by

	Haplotype	
	a	A
Probability	g_0	g_1

The symbol g_j represents the probability that a random gamete contains j A-alleles in its haplotype for locus A-a, where j may be equal to 0 or 1. The row vector (g_0, g_1) describes, in a condensed way, the **haplotypic composition** of the gametes. The habit to use the symbol q instead g_0 and the symbol p instead of g_1 is followed in this book whenever a single locus is considered. The term **allele frequency** will be used to indicate the probability of the considered allele.

So far it has been assumed that the allele frequencies are known and hereafter the theory is further developed without considering the question of how one arrives at such knowledge. In fact allele frequencies are often unknown. When one would like to estimate them one might do that in the following way. Assume that a random sample of N plants is comprised of the following numbers of plants of the various genotypes:

	Genotype		
	aa	Aa	AA
Number of plants	n_0	n_1	n_2

For any value for N the frequencies q and p of alleles a and A may then be estimated as

$$q = \frac{2n_0 + n_1}{2N} \text{ and } p = \frac{n_1 + 2n_2}{2N}$$

Throughout the book the expressions 'the probability that a random plant has genotype Aa', or 'the probability of genotype Aa', or 'the frequency of genotype Aa' are used as equivalents. This applies likewise for the expressions 'the probability that a gamete has haplotype A', or 'the probability of A'.

Fusion of a random female gamete with a random male gamete yields a genotype specified by $\underline{j}$, the number of A alleles in the genotype. (The number of a alleles in the genotype amounts – of course – to $2 - j$.) The probability that a plant with genotype aa results from the fusion is in fact equal to the probability of the event that $\underline{j}$ assumes the value 0. The quantity $\underline{j}$ assumes thus a certain value (0 or 1 or 2) with a certain probability. This means that $\underline{j}$ is a random variable.

The probability distribution for $\underline{j}$, *i.e.* for the genotype frequencies, is given by the binomial probability distribution:

$$P(\underline{j} = j) = \binom{2}{j} p^j q^{2-j}$$

Fusion of two random gametes therefore yields

- With probability q^2 a plant with genotype aa
- With probability $2pq$ a plant with genotype Aa
- With probability p^2 a plant with genotype AA

The probabilities for the multinomial probability distribution of plants with these genotypes may be represented in a condensed form by the row vector $(q^2, 2pq, p^2)$. This notation represents also the genotypic composition to be expected for the population obtained after panmixis in a population with gene frequencies (q, p). In the case of panmixis there is a direct relationship between the **gene frequencies** in a certain generation and the genotypic composition of the next generation (see Fig. 2.1). Thus if the genotype frequencies f_0, f_1 and f_2 of a certain population are equal to, respectively, $q^2, 2pq$ and p^2, the considered population has the so-called **Hardy–Weinberg (genotypic) composition**. The actual genotypic composition is then equal to the composition expected after panmixis. With continued panmixis, populations of later generations will continue to have the Hardy–Weinberg composition. Therefore such composition may be indicated as the **Hardy–Weinberg equilibrium**. The names of Hardy (1908) and Weinberg (1908) are associated with this genotypic composition, but it was in fact derived by Castle in 1903 (Keeler, 1968).

With two alleles per locus the maximum frequency of plants with the Aa genotype in a population originating from panmixis is $\frac{1}{2}$ for $p = q = \frac{1}{2}$ (Fig. 2.1). This occurs in F_2 populations of self-fertilizing crops. The F_2 originates from selfing of individual plants of the F_1, but because each plant of the

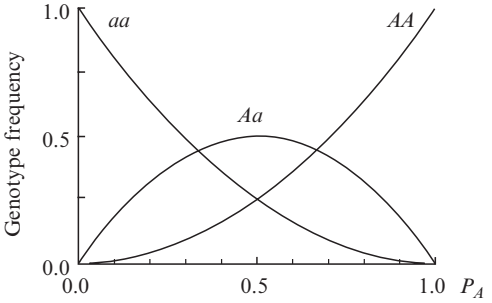


Fig. 2.1 The frequency of plants with genotype *aa, Aa* or *AA* in the population obtained by panmixis in a population with gene frequency P_A

F_1 has the same genotype, panmixis within each plant coincides with panmixis of the F_1 as a whole. (The F_1 itself may be due to bulk crossing of two pure lines; the proportion of heterozygous plants amounts then to 1.)

The Hardy–Weinberg genotypic composition constitutes the basis for the development of population genetic theory for cross-fertilizing crops. It is obtained by an infinitely large number of pairwise fusions of random eggs with random pollen, as well as by an infinitely large number of crosses involving pairs of random plants. One may also say that it is expected to occur both after pairwise fusions of random eggs and pollen, and when crossing plants at random.

In a number of situations two populations are crossed as bulks. One may call this **bulk crossing**. One population contributes the female gametes (containing the eggs) and the other population the male gametes (the pollen, containing generative nuclei in the pollen tubes). In such a case, crosses within each of the involved populations do not occur. A possibly unexpected case of bulk crossing is described in Note 2.1.

Note 2.1 Selection among plants after pollen distribution, *e.g.* selection with regard to the colour of the fruits (if fruit colour is maternally determined), implies a special form of bulk crossing: the rejected plants are then excluded as effective producers of eggs (these plants will not be harvested), whereas all plants (could) have been effective as producers of pollen. The results, to be derived hereafter, in the main text, for a bulk cross of two populations with different allele frequencies, are applied in Section 6.3.5.

A bulk cross is of particular interest if the haplotypic composition of the eggs differs from the haplotypic composition of the pollen. Thus if population I, with allele frequencies (q_1, p_1), contributes the eggs and population II, with allele frequencies (q_2, p_2), the pollen, then the expected genotypic composition of the obtained hybrid population, in row vector notation, is

$$(q_1q_2, p_1q_2 + p_2q_1, p_1p_2) \tag{2.1}$$

This hybrid population does not result from panmixis. The frequency of allele A is

$$\begin{aligned} p &= \tfrac{1}{2}(p_1q_2 + p_2q_1) + p_1p_2 = \tfrac{1}{2}p_1q_2 + \tfrac{1}{2}p_1p_2 + \tfrac{1}{2}p_2q_1 + \tfrac{1}{2}p_1p_2 \\ &= \tfrac{1}{2}p_1(q_2 + p_2) + \tfrac{1}{2}p_2(q_1 + p_1) = \tfrac{1}{2}(p_1 + p_2) \end{aligned} \tag{2.2}$$

as

$$q_2 + p_2 = q_1 + p_1 = 1$$

N.B. Further equations based on $p + q = 1$ are elaborated in Note 2.2.

Note 2.2 When deriving Equation (2.2) the equation $p + q = 1$ was used. On the basis of the latter equation several other equations, applied throughout this book, can be derived:

$$q^2 + 2pq + p^2 = 1 \tag{2.3}$$

$$p - q = 2p - 1 = 1 - 2q \tag{2.4}$$

$$(p - q)^2 = (p^2 - 2pq + q^2) = 1 - 4pq \tag{2.5}$$

$$p^2 - q^2 = (p + q)(p - q) = p - q = f_2 - f_0 \tag{2.6}$$

$$p - q + 2pq = p^2 - q^2 + 2pq = p^2 + 2pq - q^2 = 1 - 2q^2 \tag{2.7}$$

and

$$\begin{aligned} p^4 + p^3q + pq^3 + q^4 - (p - q)^2 &= p^3 + q^3 - p^2 + 2pq - q^2 \\ &= p^2(p - 1) + q^2(q - 1) + 2pq \\ &= -p^2q - pq^2 + 2pq \\ &= -pq(p + q - 2) = 2pq \end{aligned} \tag{2.8}$$

Panmictic reproduction of this hybrid population produces offspring with the Hardy–Weinberg genotypic composition. The hybrid population contains, compared to the offspring population, an excess of heterozygous plants. The excess is calculated as the difference in the frequencies of heterozygous plants:

$$\begin{aligned} (p_1q_2 + p_2q_1) - 2pq &= (p_1q_2 + p_2q_1) - 2[\tfrac{1}{2}(p_1 + p_2)\tfrac{1}{2}(q_1 + q_2) \\ &= \tfrac{1}{2}(p_1q_2 + p_2q_1 - p_1q_1 - p_2q_2) \\ &= \tfrac{1}{2}(p_1 - p_2)(q_2 - q_1) = \tfrac{1}{2}(p_1 - p_2)^2 \end{aligned} \tag{2.9}$$

This square is positive, unless $p_1 = p_2$. Thus the hybrid does indeed contain an excess of heterozygous plants. Example 2.3 illustrates that the superiority of hybrid varieties might (partly) be due to this excess. This is further elaborated in Section 9.4.1. Example 2.4 pays attention to the case of both inter- and intra-mating of two populations.

Example 2.3 It is attractive to maximize the frequency of hybrid plants whenever they have a superior genotypic value. This is applied when producing single-cross hybrid varieties by means of a bulk cross between two well-combining pure lines. If $p_1 = 1$ (thus $q_1 = 0$) in one parental line and $p_2 = 0$ (thus $q_2 = 1$) in the other, the excess of the frequency of heterozygous plants will be at its maximum, because $\frac{1}{2}(p_1 - p_2)^2$ attains then its maximum value, *i.e.* $\frac{1}{2}$. The genotypic composition of the single-cross hybrid is (0, 1, 0). Equation (2.2) implies that panmictic reproduction of this hybrid yields a population with the Hardy-Weinberg genotypic composition ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$). The excess of heterozygous plants in the hybrid population is thus indeed $\frac{1}{2}$. (Panmictic reproduction of a hybrid population tends to yield a population with a reduced expected genotypic value; see Section 9.4.1).

The excess of heterozygous plants is low when one applies bulk crossing of similar populations. At $p_1 = 0.6$ and $p_2 = 0.7$, for example, the hybrid population has the genotypic composition (0.12; 0.46; 0.42), with $p = 0.65$. The corresponding Hardy–Weinberg genotypic composition is then (0.1225; 0.4550; 0.4225) and the excess of heterozygous plants is only 0.005.

As early as 1908 open-pollinating maize populations were crossed in the USA with the aim of producing superior hybrid populations. This had already been suggested in 1880 by Beal. Shull (1909) was the first to suggest the production of single-cross hybrid varieties by crossing pure lines.

Example 2.4 Two populations of a cross-fertilizing crop, *e.g.* perennial rye grass, are mixed. The mixture consists of a portion, P, of population I material and a portion, $1 - P$, of population II material. In the mixture both mating between and within the populations occur. When assuming

- simultaneous flowering,
- simultaneous ripening,
- equal fertility of the plants of both populations and
- random mating

the proportion of hybrid seed is $2P(1 - P)$; see Foster (1971). For $P = \frac{1}{2}$ this proportion is maximal, *i.e.* $\frac{1}{2}$.

2.2.2 One Locus with more than Two Alleles

Multiple allelism does not occur in the populations considered so far. However, multiple allelism is known to occur in self- and cross-fertilizing crops (see Example 2.5). It may further be expected in three-way-cross hybrids, and their offspring, as well as in mixtures of pure lines (landraces or multiline varieties).

Example 2.5 The intensity of the anthocyanin colouration in lettuce (*Lactuca sativa*), a self-fertilizing crop, is controlled by at least three alleles.

The colour and location of the white leaf spots of white clover (*Trifolium repens*), a cross-fertilizing crop, are controlled by a multiple allelic locus. The expression for these traits appears to be controlled by a locus with at least 11 alleles. Another locus, with at least four alleles, controls the red leaf spots (Julén, 1959). (White clover is an autotetraploid crop with a gametophytic incompatibility system and a diploid chromosome behaviour; $2n = 4x = 32$).

The frequencies (f) of the genotypes A_iA_j (with $i \leq j; j = 1, \ldots, n$) for the multiple allelic locus A_1-A_2-$\ldots$-A_n attain their equilibrium values following a single round of panmictic reproduction. The genotypic composition is then:

	Genotype		
	$A_1A_1\ldots$	$A_iA_j\ldots$	A_nA_n
f	${p_1}^2$	$2p_ip_j$	${p_n}^2$

The proportion of homozygous plants is minimal for $p_j = \frac{1}{n}$ (for $j = 1, \ldots, n$) and amounts then to $n\left(\frac{1}{n}\right)^2 = \frac{1}{n}$; see Falconer (1989, pp. 388–389).

2.2.3 Two Loci, Each with Two Alleles

In Section 2.2.1 it was shown that a single round of panmictic reproduction produces immediately the Hardy–Weinberg genotypic composition with regard to a single locus. It is immediately attained because the random fusion of pairs of gametes implies random fusion of separate alleles, whose frequencies are constant from one generation to the next. For complex genotypes, *i.e.* genotypes with regard to two or more loci (linked or not), however, the so-called **linkage equilibrium** is only attained after *continued panmixis*. Presence of the Hardy–Weinberg genotypic composition for separate loci does not imply presence of linkage equilibrium! (Example 2.7 illustrates an important exception to this rule.)

In panmictic reproduction the frequencies of complex genotypes follow from the frequencies of the complex haplotypes. Linkage equilibrium is thus attained if the haplotype frequencies are constant from one generation to the next. For this reason 'linkage equilibrium' is also indicated as **gametic phase equilibrium**. In this section it is derived how the haplotypic frequencies approach their equilibrium values in the case of continued panmixis. This implies that the tighter the linkage the more generations are required. However, even for unlinked loci a number of rounds of panmictic reproduction are required to attain linkage equilibrium. The genotypic composition in the equilibrium does not depend at all on the strength of the linkage of the loci involved. The designation 'linkage equilibrium' is thus not very appropriate.

To derive how the haplotype frequencies approach their equilibrium, the notation introduced in Section 2.2.1 must be extended. We consider loci *A*-*a* and *B*-*b*, with frequencies p and q for alleles A and a and frequencies r and s for alleles B and b. The **recombination value** is represented by r_c. This parameter represents the probability that a gamete has a recombinant haplotype (see Section 2.2.4). Independent segregation of the two loci occurs at $r_c = \frac{1}{2}$, absolute linkage at $r_c = 0$. Example 2.6 illustrates the estimation of r_c in the case of a testcross with a line with a homozygous recessive (complex) genotype.

The haplotype frequencies are determined at the meiosis. The haplotypic composition of the gametes produced by generation G_{t-1} is described by

	Haplotype			
	ab	*aB*	*Ab*	*AB*
f	$g_{00,t}$	$g_{01,t}$	$g_{10,t}$	$g_{11,t}$

The last subscript (t) in the symbol for the haplotype frequencies indicates the rank of the generation to be formed in a series of generations generated by panmictic reproduction ($t = 1, 2, \ldots$); see Note 2.3.

Example 2.6 The spinach variety Wintra is susceptible to the fungus *Peronospora spinaciae* race 2 and tolerant to Cucumber virus 1. It was crossed with spinach variety Nores, which is resistant to *P. spinaciae* race 2 but sensitive to Cucumber virus 1. The loci controlling the host-pathogen relations are $A - a$ and $B - b$. The genotype of Wintra is *aaBB* and the genotype of Nores *AAbb*. The offspring, with genotype *AaBb*, were crossed with the spinach variety Eerste Oogst (genotype *aabb*), which is susceptible to *P. spinaciae* race 2 and sensitive to Cucumber virus 1. On the basis of the reaction to both pathogens a genotype was assigned to each of the 499 plants resulting from this testcross (Eenink, 1974):

	Genotype				
	aabb	*aaBb*	*Aabb*	*AaBb*	Total
Frequency					
• Observed	61	190	194	54	499
• Expected	124.75	124.75	124.75	124.75	499

The expected frequencies are calculated on the basis of the null hypothesis stating that the two involved loci are unlinked. The expected $\frac{1}{2}{:}\frac{1}{2}$ segregation ratio was confirmed by a goodness of fit test for each separate locus. The specified null hypothesis is, of course, rejected. The two loci are clearly linked. The value estimated for r_c is

$$\frac{61 + 54}{499} = 0.23$$

Note 2.3 In this book the last subscript in the symbols for the genotype and haplotype frequencies indicate the generation number. If it is t it refers to population G_t, *i.e.* the population obtained by panmictic reproduction of t successive generations.

Population G_1, resulting from panmictic reproduction in a single-cross hybrid, has the same genotypic composition as the F_2 population resulting from selfing plants of the single-cross hybrid. To standardize the numbering of generations of cross-fertilizing crops and those of self-fertilizing crops, the population resulting from the first reproduction by means of selfing might be indicated by S_1 (rather than by the more common indication F_2). To avoid confusion this will only be done when appropriate, *e.g.* in Section 3.2.1.

The last subscript in the symbols for the haplotype frequencies of the gametes giving rise to S_1 are taken to be 1. The same applies to the frequencies of the genotypes in S_1. This system for labelling generations of gametophytes and sporophytes was also adopted by Stam (1977).

Population G_0 is thus some initial population, obtained after a bulk cross or simply by mixing. It produces gametes with the haplotypic composition $(g_{00,1}; g_{01,1}; g_{10,1}; g_{11,1})$.

In the absence of selection, allele frequencies do not change. This implies

$$g_{10,1} + g_{11,1} = g_{10,2} + g_{11,2} = \ldots = p$$

for allele A, and similar equations for the frequencies of alleles a, B and b.

It was already noted that the haplotype frequencies in successive generations will be considered. In the appendix of this section it is shown that the following recurrent relations apply:

$$g_{00,t+1} = g_{00,t} - r_c d_t \tag{2.10a}$$

$$g_{01,t+1} = g_{01,t} + r_c d_t \tag{2.10b}$$

$$g_{10,t+1} = g_{10,t} + r_c d_t \tag{2.10c}$$

$$g_{11,t+1} = g_{11,t} - r_c d_t \tag{2.10d}$$

where the definition of d_t follows from

$$2d_t := f_{11\mathrm{C},t} - f_{11\mathrm{R},t} \tag{2.11}$$

where ':=' means: 'is defined as', and $t = 1, 2, 3, \ldots$

N.B. In Note 3.6 it is shown that Equations (2.10a–d) also apply to self-fertilizing crops. The recurrent equations show that the haplotype frequencies do not change from one generation to the next if $r_c = 0$ or if $d_t = 0$. Such constancy of the haplotypic composition implies constancy of the genotypic

composition. It implies presence of linkage equilibrium. Linkage equilibrium is thus immediately established by a single round of panmictic reproduction for loci with $r_c = 0$. This situation coincides with the case of a single locus with four alleles.

The symbol $f_{11\mathrm{C}}$ indicates the frequency of AB/ab-plants, *i.e.* doubly heterozygous plants in **coupling phase (C-phase)**; the symbol $f_{11\mathrm{R}}$ represents the frequency of Ab/aB-plants, *i.e.* doubly heterozygous plants in **repulsion phase (R-phase)**.

In the case of panmixis the following equations apply:

$$\begin{aligned} f_{11\mathrm{C},t} &= 2(g_{11,t}\, g_{00,t}) \\ f_{11\mathrm{R},t} &= 2(g_{10,t}\, g_{01,t}) \end{aligned}$$

In that case we get

$$d_t = (g_{11,t}\, g_{00,t}) - (g_{10,t}\, g_{01,t}) \tag{2.12}$$

This parameter is called **coefficient of linkage disequilibrium**. It appears in the following derivation:

$$\begin{aligned} g_{11,t} &= g_{11,t}(g_{10,t} + g_{01,t} + g_{11,t} + g_{00,t}) = (g_{10,t}\, g_{01,t} + g_{10,t}\, g_{11,t} \\ &\quad + g_{11,t}\, g_{01,t} + g_{11,t}^2) + (g_{11,t}\, g_{00,t} - g_{10,t}\, g_{01,t}) \\ &= (g_{10,t} + g_{11,t})(g_{01,t} + g_{11,t}) + d_t = pr + d_t \end{aligned}$$

Equation (2.10d) may thus be rewritten as

$$pr + d_{t+1} = (pr + d_t) - r_c d_t$$

which implies not only

$$d_{t+1} = (1 - r_c) d_t$$

but of course also

$$d_t = (1 - r_c)^{t-1} d_1 \tag{2.13}$$

for $t = 2, 3, \ldots$

The derivation above (and similar derivations for the other haplotype frequencies) implies

$$d_t = g_{11,t} - pr = -(g_{10,t} - ps) = -(g_{01,t} - qr) = g_{00,t} - qs$$

Because $\frac{1}{2} \leq (1 - r_c) \leq 1$, continued panmixis implies continued decrease of d_t. The decrease is faster for smaller values of $1 - r_c$, *i.e.* for higher values of r_c. Independent segregation, *i.e.* $r_c = \frac{1}{2}$, yields the fastest reduction, *viz.* halving of d_t by each panmictic reproduction. The value of d_t eventually attained,

i.e. $d_t = 0$, implies that linkage equilibrium is attained, *i.e.* constancy of the haplotype frequencies. The haplotype frequencies have then a special value, *viz.*

$$\begin{aligned} g_{00} &= qs \\ g_{01} &= qr \\ g_{10} &= ps \\ g_{11} &= pr \end{aligned}$$

The equilibrium frequencies of the haplotypes are equal to the products of the frequencies of the alleles involved, and the equilibrium frequencies of the complex genotypes are equal to the products of the Hardy–Weinberg frequencies of the single-locus genotypes for the loci involved. The strength of the linkage between the loci is irrelevant with regard to the genotypic composition in the equilibrium. It only affects the number of generations of panmictic reproduction required to 'attain' the equilibrium.

Table 2.1 presents the equilibrium frequencies of complex genotypes and phenotypes for the simultaneously considered loci *A*-*a* and *B*-*b*.

Table 2.1 Equilibrium frequencies of (a) complex genotypes and (b) phenotypes in the case of complete dominance. The equilibrium is attained after continued panmictic reproduction

(a) Genotypes				
	bb	Bb	BB	
aa	q^2s^2	$2q^2rs$	q^2r^2	q^2
Aa	$2pqs^2$	$4pqrs$	$2pqr^2$	$2pq$
AA	p^2s^2	$2p^2rs$	p^2r^2	p^2
	s^2	$2rs$	r^2	1
(b) Phenotypes				
	bb	$B\cdot$		
aa	q^2s^2	$q^2(1-s^2)$	q^2	
$A\cdot$	$(1-q^2)s^2$	$(1-q^2)(1-s^2)$	$(1-q^2)$	
	s^2	$1-s^2$		

The foregoing is illustrated in Example 2.7, which deals with the production of a single-cross hybrid variety and the population resulting from its offspring as obtained by panmictic reproduction. Example 2.8 illustrates the production of a synthetic variety and a few of its offspring generations as obtained by continued random mating.

Example 2.7 Cross $\frac{AB}{AB} \times \frac{ab}{ab}$ yields a doubly heterozygous genotype in the coupling phase, *i.e.* $\frac{AB}{ab}$, whereas cross $\frac{Ab}{Ab} \times \frac{aB}{aB}$ yields a doubly heterozygous genotype in the repulsion phase, *i.e.* $\frac{Ab}{aB}$. In both cases the single-cross hybrid variety, say population G_0, is heterozygous for the loci *A*-*a* and *B*-*b*. It produces gametes with the following haplotypic composition:

		Haplotype				
		ab	*aB*	*Ab*	*AB*	d_1
f	in general	$g_{00,1}$	$g_{01,1}$	$g_{10,1}$	$g_{11,1}$	
	for G_0 in C-phase:	$\frac{1}{2} - \frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2} - \frac{1}{2}r_c$	$\frac{1}{4}(1-2r_c)$
	for G_0 in R-phase:	$\frac{1}{2}r_c$	$\frac{1}{2} - \frac{1}{2}r_c$	$\frac{1}{2} - \frac{1}{2}r_c$	$\frac{1}{2}r_c$	$-\frac{1}{4}(1-2r_c)$

The quantity d_1 is calculated according to Equation (2.12). This yields for G_0 in C-phase

$$d_1 = \tfrac{1}{4}(1-r_c)^2 - \tfrac{1}{4}r_c^2 = \tfrac{1}{4}(1-2r_c)$$

The value for d_1 is in the interval $(0, \frac{1}{4})$ or in the interval $(-\frac{1}{4}, 0)$. In G_1 the absolute value of d_1 is at a maximum. Continued panmictic reproduction gives, in G_∞, the linkage equilibrium pertaining to $p = q = r = s = \frac{1}{2}$. Table 2.2 presents the genotypic composition of population G_1 resulting from a single panmictic reproduction of either G_0 in C-phase or in R-phase, as well as the genotypic composition of population G_∞ resulting from continued panmixis.

Starting with a single-cross hybrid, the quantity d_1 is equal to zero for loci with $r_c = \frac{1}{2}$. Then a single generation of panmictic reproduction produces a population in linkage equilibrium. This remarkable result applies even in the case of selfing of the hybrid variety. (In Section 2.2.1 it has already been indicated that the result of selfing of F_1 plants coincides with the result of panmixis among F_1 plants). Thus for unlinked loci panmictic reproduction (or selfing) of a single-cross hybrid immediately yields a population in linkage equilibrium. Continued panmictic reproduction does not yield further shifts in haplotype and genotype frequencies. This means that it is useless to apply random mating in the F_2 of a self-fertilizing crop with the goal of increasing the frequency of plants with a recombinant genotype.

On the basis of the frequencies of the phenotypes for two traits (each with two levels of expression) showing qualitative variation, one can easily determine whether or not a certain population is in linkage equilibrium. It is, however, impossible to conclude whether or not the loci involved are linked. Only test crosses between individual plants with the phenotype $A \cdot B \cdot$ and plants with genotype *aabb* will give evidence about this.
N.B. By 'phenotype $A \cdot B \cdot$' is meant the phenotype due to genotype *AABB*, *AaBB*, *AABb* or *AaBb*.

Table 2.2 The genotypic composition of G_1, both for G_0 in coupling phase and in repulsion phase, and of G_∞

	Genotypic composition		
Genotype	G_1 for G_0 in C-phase	G_1 for G_0 in R-phase	G_∞
aabb	$\frac{1}{4}(1-r_c)^2$	$\frac{1}{4}r_c^2$	$\frac{1}{16}$
aaBb	$\frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}r_c(1-r_c)$	$\frac{2}{16}$
aaBB	$\frac{1}{4}r_c^2$	$\frac{1}{4}(1-r_c)^2$	$\frac{1}{16}$
Aabb	$\frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}r_c(1-r_c)$	$\frac{2}{16}$
AB/ab	$\frac{1}{2}(1-r_c)^2$	$\frac{1}{2}r_c^2$	$\frac{2}{16}$
Ab/aB	$\frac{1}{2}r_c^2$	$\frac{1}{2}(1-r_c)^2$	$\frac{2}{16}$
AaBB	$\frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}r_c(1-r_c)$	$\frac{2}{16}$
AAbb	$\frac{1}{4}r_c^2$	$\frac{1}{4}(1-r_c)^2$	$\frac{1}{16}$
AABb	$\frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}r_c(1-r_c)$	$\frac{2}{16}$
AABB	$\frac{1}{4}(1-r_c)^2$	$\frac{1}{4}r_c^2$	$\frac{1}{16}$

Example 2.8 A synthetic variety is planned to be produced by intermating five clones of a self-incompatible grass species. Because crosses within each of the five components are excluded, the synthetic variety is produced by outbreeding. It is, therefore, due to a complex bulk cross. The obtained plant material is designated as Syn_1 (or G_0 in the present context). The five clones have the following genotypes for the two unlinked loci B_1-b_1 and B_2-b_2: clone 1: $b_1b_1b_2b_2$; clones 2 and 3: $B_1B_1b_2b_2$, and clones 4 and 5: $B_1B_1B_2B_2$.
The genotypic composition of Syn_1 can be derived from the following scheme:

♀ \ ♂	$b_1b_1b_2b_2$	$B_1B_1b_2b_2$	$B_1B_1b_2b_2$	$B_1B_1B_2B_2$	$B_1B_1B_2B_2$
$b_1b_1b_2b_2$	-	$B_1b_1b_2b_2$	$B_1b_1b_2b_2$	$B_1b_1B_2b_2$	$B_1b_1B_2b_2$
$B_1B_1b_2b_2$	$B_1b_1b_2b_2$	-	$B_1B_1b_2b_2$	$B_1B_1B_2b_2$	$B_1B_1B_2b_2$
$B_1B_1b_2b_2$	$B_1b_1b_2b_2$	$B_1B_1b_2b_2$	-	$B_1B_1B_2b_2$	$B_1B_1B_2b_2$
$B_1B_1B_2B_2$	$B_1b_1B_2b_2$	$B_1B_1B_2b_2$	$B_1B_1B_2b_2$	-	$B_1B_1B_2B_2$
$B_1B_1B_2B_2$	$B_1b_1B_2b_2$	$B_1B_1B_2b_2$	$B_1B_1B_2b_2$	$B_1B_1B_2B_2$	-

Table 2.3 presents the genotype frequencies in a few relevant generations. When deriving these it was assumed that incompatibility can be neglected when considering continued panmictic reproduction starting in G_0. The portion of homozygous plants in G_0, G_1, G_2 and G_∞ amounts to 0.2; 0.35; 0.3508 and 0.3536, respectively. The excess of heterozygous plants in comparison to the linkage equilibrium amounts therefore to 0.1536; 0.0036 and 0.0028 in G_0, G_1 and G_2, respectively. (This concerns plants which are heterozygous for one or two loci. For each single locus the Hardy–Weinberg genotypic composition occurs in G_1 and all later generations).

Table 2.3 The genotypic composition of plant material obtained when creating and maintaining an imaginary synthetic variety (see Example 2.8). P indicates the parental clones, G_0 indicates population Syn_1, G_1 indicates Syn_2, G_2 indicates Syn_3 and G_∞ indicates Syn_∞

	Frequency				
Genotype	P	G_0	G_1	G_2	G_∞
$b_1b_1b_2b_2$	0.2		0.0225	0.0182	0.0144
$b_1b_1B_2b_2$			0.0150	0.0176	0.0192
$b_1b_1B_2B_2$			0.0025	0.0042	0.0064
$B_1b_1b_2b_2$		0.2	0.1350	0.1256	0.1152
B_1B_2/b_1b_2		0.2	0.1050	0.0904	0.0768
B_1b_2/b_1B_2			0.0450	0.0605	0.0768
$B_1b_1B_2B_2$			0.0350	0.0436	0.0512
$B_1B_1b_2b_2$	0.4	0.1	0.2025	0.2162	0.2304
$B_1B_1B_2b_2$		0.4	0.3150	0.3116	0.3072
$B_1B_1B_2B_2$	0.4	0.1	0.1225	0.1122	0.1024

APPENDIX: The haplotype frequencies in generation t

In this appendix, first is derived an equation relating the frequency of gametes with haplotype ab in generation $t+1$ to its frequency in generation t, i.e. Equation (2.10a). Thereafter an equation describing the haplotype frequencies in generations due to continued panmictic reproduction, starting with a single-cross hybrid, is derived.

The frequency of gametes with haplotype ab

The relevant genotypes, their frequencies (in general, as well as after panmixis) and the haplotypic composition of the gametes they produce are:

	Genotype frequency		Haplotype frequency			
Genotype	in general	after panmixis	*ab*	*aB*	*Ab*	*AB*
aabb	f_{00}	${g_{00}}^2$	1	0	0	0
Aabb	f_{10}	$2g_{00}g_{10}$	$\frac{1}{2}$	0	$\frac{1}{2}$	0
AAbb	f_{20}	${g_{10}}^2$	0	0	1	0
aaBb	f_{01}	$2g_{00}g_{01}$	$\frac{1}{2}$	$\frac{1}{2}$	0	0
$\frac{AB}{ab}$	$f_{11\mathrm{C}}$	$2g_{00}g_{11}$	$\frac{1}{2}-\frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2}-\frac{1}{2}r_c$
$\frac{Ab}{aB}$	$f_{11\mathrm{R}}$	$2g_{10}g_{01}$	$\frac{1}{2}r_c$	$\frac{1}{2}-\frac{1}{2}r_c$	$\frac{1}{2}-\frac{1}{2}r_c$	$\frac{1}{2}r_c$
AABb	f_{21}	$2g_{01}g_{11}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
aaBB	f_{02}	${g_{01}}^2$	0	1	0	0
AaBB	f_{12}	$2g_{01}g_{11}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
AABB	f_{22}	${g_{11}}^2$	0	0	0	1

The frequency of gametes with haplotype ab, produced by generation G_t, are equal to

$$\begin{aligned} g_{00,t+1} &= f_{00,t} + \tfrac{1}{2}f_{10,t} + \tfrac{1}{2}f_{01,t} + \tfrac{1}{2}(1-r_c)f_{11\mathrm{C},t} + \tfrac{1}{2}r_c f_{11\mathrm{R},t} \\ &= f_{00,t} + \tfrac{1}{2}f_{10,t} + \tfrac{1}{2}f_{01,t} + \tfrac{1}{2}f_{11\mathrm{C},t} - r_c d_t \end{aligned}$$

One may derive likewise

$$\begin{aligned} g_{01,t+1} &= f_{02,}t + \tfrac{1}{2}f_{01,}t + \tfrac{1}{2}f_{12,}t + \tfrac{1}{2}f_{11\mathrm{R},}t + r_c dt \\ g_{10,t+1} &= f_{20,}t + \tfrac{1}{2}f_{10,}t + \tfrac{1}{2}f_{21,}t + \tfrac{1}{2}f_{11\mathrm{R},}t + r_c dt \\ g_{11,t+1} &= f_{22,}t + \tfrac{1}{2}f_{21,}t + \tfrac{1}{2}f_{12,}t + \tfrac{1}{2}f_{11\mathrm{C},}t - r_c dt \end{aligned}$$

Panmictic reproduction of generation G_t yields generation G_{t+1}. The genotypic composition of G_{t+1} is described by the frequencies given by the third column of the previous table. Inclusion of these genotype frequencies in the above equation for $g_{00,t+1}$ gives

$$\begin{aligned} g_{00,t+1} &= g_{00,t}^2 + g_{00,t}\, g_{10,t} + g_{00,t}\, g_{01,t} + g_{00,t}\, g_{11,t} - r_c d_t \\ &= g_{00,t}(g_{00,t} + g_{10,t} + g_{01,t} + g_{11,t}) - r_c d_t = g_{00,t} - r_c d_t \end{aligned}$$

where, according to Equation (2.12)

$$d_t = (g_{11,t}\, g_{00,t} - g_{10,t}\, g_{01,t})$$

Similarly one can derive

$$\begin{aligned} g_{01,t+1} &= g_{01,t} + r_c d_t \\ g_{10,t+1} &= g_{10,t} + r_c d_t \\ g_{11,t+1} &= g_{11,t} - r_c d_t \end{aligned}$$

The haplotype frequencies in generations due to continued panmictic reproduction, starting with a single-cross hybrid

In the case of panmictic reproduction starting from a single-cross hybrid there will be a symmetry in the haplotype frequencies such that

$$g_{00,t} = g_{11,t}$$

and

$$g_{01,t} = g_{10,t} = \tfrac{1}{2} - g_{11,t}$$

Derivation of $g_{11,t}$ suffices then to obtain the frequencies of all haplotypes with regard to two segregating loci. An equation presenting $g_{11,t}$ immediately for any value for t will now be derived.

If the genotype of the single-cross hybrid is $\frac{AB}{ab}$, *i.e.* coupling phase, the genotypic composition of the initial population G_0 is simply described by

$f_{11\mathrm{C},0} = 1$, if it is $\frac{Ab}{aB}$ the genotypic composition of G_0 is described by $f_{11\mathrm{R},0} = 1$. Equation (2.11) yields then

$$d_0 = \tfrac{1}{2}$$

in the former case, and

$$d_0 = \tfrac{-1}{2}$$

in the latter case. The frequency of gametes with the AB haplotype among the gametes produced by the single-cross amounts to

$$g_{11,1} = \tfrac{1}{2}(1 - r_c)$$

and

$$g_{11,1} = \tfrac{1}{2} r_c$$

respectively (see Example 2.7). In Example 2.7 it was also derived that

$$d_1 = \tfrac{1}{4}(1 - 2r_c)$$

for G_0 in C-phase and that

$$d_1 = \tfrac{-1}{4}(1 - 2r_c)$$

for G_0 in R-phase.

The frequencies of AB haplotypes in the case of continued panmixis follow from Equation (2.10d) combined with Equation (2.13):

$$\begin{aligned} g_{11,t+2} &= g_{11,t+1} - r_c d_{t+1} = g_{11,t+1} - r_c(1 - r_c)^t d_1 \\ &= g_{11,t} - r_c(1 - r_c)^{t-1} d_1 - r_c(1 - r_c)^t d_1 \\ &= g_{11,1} - r_c d_1[(1 - r_c)^0 + \ldots + (1 - r_c)^t] \end{aligned}$$

The terms within the brackets form a convergent geometric series. The sum of such terms is given by the expression

$$a\frac{1 - q^n}{1 - q}$$

where a is the first term, q is the multiplying factor and n is the number of terms. In the present situation this sum amounts to

$$\frac{1 - (1 - r_c)^{t+1}}{r_c}$$

Thus

$$g_{11,t+2} = g_{11,1} - d_1[1 - (1 - r_c)^{t+1}] \tag{2.14}$$

For $r_c = \frac{1}{2}$ we got $d_1 = 0$. Then

$$g_{11,t+2} = g_{11,1} = \tfrac{1}{4}$$

This implies that linkage equilibrium is present after one generation with panmictic reproduction!

For G_0 in C-phase, Equation (2.14) can be rewritten as

$$g_{11,t+2} = \tfrac{1}{2}(1 - r_c) - \tfrac{1}{4}(1 - 2r_c)[1 - (1 - r_c)^{t+1}] \qquad (2.14C)$$

Thus

$$g_{11,2} = \tfrac{1}{2}(1 - r_c) - \tfrac{1}{4}r_c(1 - 2r_c) = \tfrac{1}{2}r_c^2 - \tfrac{3}{4}r_c + \tfrac{1}{2}$$

For G_0 in R-phase, Equation (2.14) can be transformed into

$$g_{11,t+2} = \tfrac{1}{2}r_c + \tfrac{1}{4}(1 - 2r_c)[1 - (1 - r_c)^{t+1}] \qquad (2.14R)$$

This implies

$$g_{11,2} = \tfrac{1}{2}r_c + \tfrac{1}{4}r_c(1 - 2r_c) = -\tfrac{1}{2}r_c^2 + \tfrac{3}{4}r_c$$

$$g_{11,3} = \tfrac{1}{2}r_c + \tfrac{1}{4}(1 - 2r_c)[1 - (1 - r_c)^2] = \tfrac{1}{2}r_c^3 - 1\tfrac{1}{4}r_c^2 + r_c$$

These equations are of relevance with regard to the question of whether it is advantageous, when it is aimed to promote the frequency of plants with a genotype due to recombination, to apply random mating in an F_2 population of a self-fertilizing crop (see Section 3.2.2).

2.2.4 More than Two Loci, Each with Two or more Alleles

Attention is given to linkage involving three loci. A few aspects which play an important role with regard to linkage maps, for example of molecular markers, are considered along with the frequencies of complex genotypes after continued panmixis.

Linkage involving three loci

Three loci *A*-*a*, *B*-*b* and *C*-*c* are considered. These loci occur in this order along a chromosome. The segments AB, BC and AC are distinguished. Effective recombination of alleles belonging to loci *A*-*a* and *B*-*b* requires that the number of crossover events in segment AB is an odd number. The probability of recombination is called **recombination value**, designated by the symbol r_c, or by the symbol r_{AB} or simply by r (depending on the context).

With an even number of times of crossing-over in segment AB there is no (effective) recombination. The probability of this event is $1 - r_{AB}$.

There is (effective) recombination of alleles belonging to loci *A*-*a* and *C*-*c* if there is either (effective) crossing-over in segment AB, but not in segment BC; or if there is (effective) crossing-over in segment BC, but not in segment AB. If the occurrence of recombination in one chromosome segment has no effect

on the recombination value for the adjacent segment the following relation applies:

$$r_{AC} = r_{AB}(1 - r_{BC}) + r_{BC}(1 - r_{AB}) = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$$

This situation is likely for loci that are not too closely linked. The situation where recombination in one segment depresses the probability of recombination in an adjacent segment is called **chiasma interference**. A more general expression for r_{AC} is thus:

$$r_{AC} = r_{AB} + r_{BC} - 2(1 - \delta)r_{AB}r_{BC},$$

where δ is the interference parameter, ranging from 0 (no interference) through 1 (complete interference). It shows that r_{AC} is higher at higher values for δ.

Recombination values are additive if

$$2(1 - \delta)r_{AB}r_{BC} = 0$$

i.e. if $\delta = 1$ and/or $r_{AB}r_{BC} = 0$. In other cases they are not additive. These conditions imply that recombination values are mostly not additive. They are, consequently, inappropriate to measure distances between loci.

The hypothesis of independence of crossing-over in segments AB and BC, *i.e.* the hypothesis of absence of chiasma interference, can be tested by means of a goodness-of-fit test. Among N plants, the expected number of plants with a genotype which is due to double crossing-over amounts, according to this hypothesis, to $r_{AB}r_{BC}N$. It is compared to the observed number. The ratio

$$\frac{observed\ number}{expected\ number}$$

is called **coefficient of coincidence**. When there is independency it is equal to 1. Its complement, *i.e.*

$$1 - \frac{observed\ number}{expected\ number}$$

estimates δ. Its value is positive if the observed number of plants with the recombinant genotype is smaller than the number expected at independency: the presence of a chiasma in the one segment hinders the formation of a chiasma in the other segment.

The actual distance between loci, say the **map distance** m, measures the total number of cross-over events (both odd and even numbers) between the loci. This distance is an additive measure. It can only approximately be determined from recombination values. Haldane (1919) developed an approximation for the situation in the absence of interference ($\delta = 0$). His **mapping function** is

$$m = -\frac{\ln(1 - 2r_c)}{2},$$

where m represents the expected number of cross-over events in the considered segment (Kearsey and Pooni, 1996; pp. 127–130). As the map distance is mostly expressed in centiMorgans (cM), this function is often written as

$$m = -50 \ln(1 - 2r_c)$$

An approximation which takes interference into account is called Kosambi's mapping function (Kosambi, 1944).

Frequencies of complex genotypes after continued panmixis
It can be shown (Bennett, 1954) that continued panmixis eventually leads to an equilibrium of the frequencies of complex genotypes for three or more loci, each with two or more alleles. The equilibrium is characterized by haplotype frequencies equal to the products of the frequencies of the alleles involved. Linkage equilibrium for one or more pairs of loci does not imply equilibrium of the frequencies of complex genotypes for three or more loci. Equilibrium of the frequencies for complex genotypes implies, however, linkage equilibrium for all pairs of loci.

2.3 Autotetraploid Chromosome Behaviour and Panmixis

The implications of panmixis in an autotetraploid crop will only be considered for a single locus with two alleles. This is to keep the mathematical derivations simple. It will be shown that the equilibrium frequencies of the genotypes are not obtained after a single panmictic reproduction. At equilibrium the frequencies of the genotypes and the haplotypes are equal to the products of the frequencies of the alleles involved.

Among cross-fertilizing autotetraploid crops the more important representatives are alfalfa (*Medicago sativa* L.; $2n = 4x = 32$) and cocksfoot (*Dactylis glomerata* L.; $2n = 4x = 28$). Additionally, highbush blueberry (*Vaccinium corymbosum* L.; $2n = 4x = 48$) might be mentioned. Leek (*Allium porrum* L.; $2n = 4x = 32$) is an autotetraploid crop with a tendency to a diploid behaviour of the chromosomes (Potz, 1987). Among ornamentals several autotetraploid species occur, *e.g. Freesia hybrida, Cyclamen persicum* Mill. ($2n = 4x = 48$) and *Begonia semperflorens*. Also, artificial autotetraploid crops have been made, *e.g.* rye (*Secale cereale* L.; $2n = 4x = 28$) and perennial rye grass (*Lolium perenne* L.; $2n = 4x = 28$). In 1977 about 500,000 ha of autotetraploid rye were grown in the former Soviet Union. Sweet potato, *i.e. Ipomoea batatas* var. *littoralis* ($2n = 4x = 60$) or *I. batatas* var. *batatas* ($2n = 6x = 90$), may be considered as a cross-fertilizing crop (due to self-incompatibility), but it is mainly vegetatively propagated.

Under certain conditions **double reduction** may occur in autotetraploid crops, in which case (parts of) sister chromatids end up in the same gamete. The resulting haplotype is homozygous for the loci involved. The process of

double reduction causes the frequency of homozygous genotypes and haplotypes to be somewhat higher than in absence of double reduction. Blakeslee, Belling and Farnham (1923) discovered the phenomenon in autotetraploid jimson weed (*Datura stramonium* L.; $2n = 4x = 48$): a triplex plant (with genotype $AAAa$) produced some nulliplex offspring after crossing with a nulliplex (genotype $aaaa$). This is only possible if the triplex plant produces aa gametes. The process of double reduction is an interesting phenomenon, but in a quantitative sense it is of no importance. For this reason we assume that double reduction does not occur.

The autotetraploid genotypes to be distinguished for locus A-a are $aaaa$ (**nulliplex**), $Aaaa$ (**simplex**), $AAaa$ (**duplex**), $AAAa$ (**triplex**) and $AAAA$ (**quadruplex**). In each cell these genotypes contain JA alleles and $4 - Ja$ alleles. At meiosis two of these four alleles are sampled to produce a gamete. The haplotypes that can be produced by an autotetraploid plant containing JA alleles can be described by $\underline{j}$, the number of A alleles that they contain, where $j = 0$, 1 or 2. The conditional probability distribution for $\underline{j}$, given that the parental genotype contains JA alleles, is a **hypergeometric probability distribution:**

$$P(\underline{j} = j|J) = \frac{\binom{J}{j}\binom{4-J}{2-j}}{\binom{4}{2}} = \frac{1}{6}\binom{J}{j}\binom{4-J}{2-j}$$

The probability that a triplex plant (*i.e.* $J = 3$) produces a gamete with haplotype Aa (*i.e.* $j = 1$) is therefore

$$P(\underline{j} = 1|J = 3) = \frac{1}{6}\binom{3}{1}\binom{1}{1} = \frac{1}{2}$$

Table 2.4 presents, for each autotetraploid genotype, the haplotypic composition, *i.e.* the probability distribution for the haplotypes produced.

The genotypic composition of a tetraploid population is described like that of a diploid population. Thus in the case of autotetraploid species the row

Table 2.4 The haplotypic composition of the gametes produced by each of the five autotetraploid genotypes that can be distinguished for locus A-a

	Haplotype		
Genotype	aa	Aa	AA
$aaaa$	1	0	0
$Aaaa$	$\frac{1}{2}$	$\frac{1}{2}$	0
$AAaa$	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$
$AAAa$	0	$\frac{1}{2}$	$\frac{1}{2}$
$AAAA$	0	0	1

vector $(f_0, f_1, f_2, f_3, f_4)$ is used. The equilibrium frequencies of the genotypes are attained as soon as the haplotype frequencies are stable. Therefore the haplotypic composition of successive generations with panmictic reproduction will be monitored.

Some initial population G_0 produces gametes with haplotypic composition:

	Haplotype		
	aa	Aa	AA
f	$g_{0,1}$	$g_{1,1}$	$g_{2,1}$

The frequency of a is

$$q = g_{0,1} + \tfrac{1}{2}g_{1,1}$$

and that of A is

$$p = \tfrac{1}{2}g_{1,1} + g_{2,1}$$

Panmictic reproduction of G_0 yields population G_1 with the following genotypic composition:

	Genotype				
	$aaaa$	$Aaaa$	$AAaa$	$AAAa$	$AAAA$
f	${g_{0,1}}^2$	$2g_{0,1}g_{1,1}$	${g_{1,1}}^2 + 2g_{0,1}g_{2,1}$	$2g_{1,1}g_{2,1}$	${g_{2,1}}^2$

The haplotypic composition of the gametes produced by G_1 is:

	Haplotype		
	aa	Aa	AA
f	$g_{0,2}$	$g_{1,2}$	$g_{2,2}$

According to Table 2.4 the following applies:

$$\begin{aligned} g_{1,2} &= \tfrac{1}{2}(2g_{0,1}g_{1,1}) + \tfrac{2}{3}({g_{1,1}}^2 + 2g_{0,1}g_{2,1}) + \tfrac{1}{2}(2g_{1,1}g_{2,1}) \\ &= \tfrac{2}{3}\left(\tfrac{3}{2}g_{0,1}g_{1,1} + \tfrac{3}{2}g_{1,1}g_{2,1} + {g_{1,1}}^2 + 2g_{0,1}g_{2,1}\right) \\ &= \tfrac{2}{3}\left[2(g_{0,1} + \tfrac{1}{2}g_{1,1})(\tfrac{1}{2}g_{1,1} + g_{2,1}) + \tfrac{1}{2}g_{1,1}(g_{0,1} + g_{1,1} + g_{2,1})\right] \\ &= \tfrac{2}{3}(2pq + \tfrac{1}{2}g_{1,1}) \end{aligned}$$

Generally

$$g_{1,t+1} = \tfrac{2}{3}(2pq + \tfrac{1}{2}g_{1,t}) \tag{2.15}$$

The frequencies of the genotypes have attained their equilibrium (e) values as soon as the frequencies of the haplotypes are constant. The latter implies:

$$g_{1,e} = \tfrac{2}{3}(2pq + \tfrac{1}{2}g_{1,e}),$$

i.e.

$$g_{1,e} = 2pq$$

The haplotype frequencies are then

$$\begin{aligned} g_{0,e} &= q - \tfrac{1}{2}g_{1,e} = q - pq = q^2 \\ g_{1,e} &= 2pq \\ g_{2,e} &= p - \tfrac{1}{2}g_{1,e} = p - pq = p^2 \end{aligned}$$

The genotypic composition in equilibrium is consequently

	Genotype				
	aaaa	*Aaaa*	*AAaa*	*AAAa*	*AAAA*
f	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4

This composition is also given by the probability distribution for $\underline{J}$, the number of A alleles in the autotetraploid genotype:

$$P(\underline{J} = J) = \binom{4}{J} p^J q^{4-J}$$

The deviation from the equilibrium is measured by the quantity d_t, which measures the excess or deficit of the frequency of gametes with the *Aa* haplotype with regard to their equilibrium frequency. Thus d_t is defined as follows:

$$d_t := g_{1,t} - g_{1,e} \tag{2.16}$$

The rate of decrease of d_t indicates how fast the equilibrium is approached. Equations (2.16) and (2.15) yield

$$d_{t+1} = g_{1,t+1} - g_{1,e} = \tfrac{2}{3}(2pq + \tfrac{1}{2}g_{1,t}) - 2pq = \tfrac{1}{3}(g_{1,t} - g_{1,e}) = \tfrac{1}{3}d_t$$

One round of panmictic reproduction produces a population in which the deviation amounts only to $\frac{1}{3}$ of the deviation in the preceding population. The equilibrium is approached in an asymptotic way. Example 2.9 gives an illustration.

Example 2.9 The approach of the equilibrium is considered for an initial population G_0 with genotypic composition (0.04; 0; 0.72; 0; 0.24). The haplotype frequencies are:

$$g_{0,1} = 0.04 + 0.12 = 0.16$$
$$g_{1,1} = 0.48$$
$$g_{2,1} = 0.12 + 0.24 = 0.36$$

Thus $q = 0.4$ and $p = 0.6$. This implies that:

$$g_{0,1} = q^2 = g_{0,e}$$
$$g_{1,1} = 2pq = g_{1,e}$$
$$g_{2,1} = p^2 = g_{2,e}$$

Generation G_1 will therefore have the equilibrium composition: (0.0256; 0.1536; 0.3456; 0.3456; 0.1296).

For a more advanced treatment of the population genetic theory of cross-fertilizing crops with an autotetraploid behaviour of the chromosomes the reader is referred to Seyffert (1960). Finally, it is emphasized once again that in this section it was assumed that the population contains only two different alleles for the segregating locus. In fact more alleles may occur in such a way that plants with three or four different alleles per locus are present, *viz.* plants with genotype $A_iA_iA_jA_k$ or $A_iA_jA_kA_l$, respectively. Quiros (1982) reported such genotypes for isozyme loci in alfalfa. Some claims have been made that plants with a heterozygous genotype containing three or four different alleles for the considered locus, are more vigorous than plants with a heterozygous genotype containing one or two alleles (Busbice and Wilsie, 1966).

Chapter 3
Population Genetic Effects of Inbreeding

Because of the agronomic importance of self-fertilizing crops, some population genetic effects of continued selfing will be considered. Also other inbreeding systems, e.g. parent × offspring mating and full sib mating, will get attention. Continued inbreeding yields populations consisting of a mixture of plants with homozygous genotypes. The decrease of the frequency of heterozygous plants is described for both diploid and autotetraploid crops. It is shown that continued inbreeding eventually leads to a genotypic composition which is approximately determined by the initial haplotype frequencies. As perfect selfing is an idealization, also some attention is given to reproduction by means of a mixture of self-fertilization and cross-fertilization.

3.1 Introduction

Inbreeding occurs if mating plants are, on the average, *more* related than random pairs of plants. A more than average relatedness of the mating plants is thus a prerequisite. Relatedness implies, of course, that the plants involved share one or more ancestors. The strength of the inbreeding depends on the degree of relatedness (Note 3.1) of the mating plants. It has already been noted in Section 2.1 that mating of related plants may occur in random mating, but in that case it occurs as a matter of chance.

Note 3.1 Several yardsticks for measuring the degree of relatedness exist, a common one being the probability that an allele of a certain locus in some plant is identical by descent to an arbitrary allele at that same locus in its mate (Falconer and MacKay, 1996, p. 58). In regular systems of inbreeding the degree of relatedness of the mating plants is uniform across all pairs of mating plants. In this book no attention is given to the determination of the degree of relatedness.
Regular systems of inbreeding are far more common in plant breeding than irregular systems. No attention will, therefore, be given to irregular systems of inbreeding.

The counterpart of inbreeding is outbreeding. With **outbreeding** mating plants are on the average *less* related than random pairs of plants. Self-incompatibility is a natural cause for outbreeding as related plants tend to have a similar genotype at the incompatibility locus/loci. After intercrossing,

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 33–58.

such plants will produce no (or few) offspring. Artificial forms of outbreeding are

- Bulk crossing of two unrelated populations (Section 2.2.1)
- Selection of parents to be crossed in such a way that inbreeding is avoided as much as possible

Outbreeding occurs also in the case of immigration.

The population genetic effect of inbreeding is a decrease in the frequency of heterozygous plants. This involves all loci, for all traits. (Random mating, on the other hand, is a mode of reproduction that may occur for certain traits and may simultaneously be absent for other traits). When starting with an F_2 population and considering segregating loci, the frequency of heterozygous plants is the same for all loci. This applies to the successive generations of the superpopulation (see Section 2.1). Each subpopulation consists of few plants: in the case of selfing only a single plant, in the case of full sib mating only pairs of plants. Within these separate subpopulations reproduction is by means of random mating. The random variation of the gene frequencies occurring in small populations (Chapter 7) causes the subpopulations to vary with regard to the frequencies of heterozygous plants: not only for different loci, but also for the same locus. Individual plants of the F_2 (or F_3, etc.) populations vary therefore in the number of heterozygous loci.

In diploid crops procedures for the production of **doubled haploid lines** (DH-lines) allow the production of pure lines from heterozygous parents in a single generation. Doubling of the number of chromosomes of haploid plants, generated by **parthenogenesis** or by **anther culture**, yields immediately complete homozygosity. For dioecious crops as well as for self-fertilizing crops with a long juvenile phase, *e.g. Coffea arabica* L., this approach is an attractive alternative to continued inbreeding.

Tissue culture techniques for the regeneration of plants from anthers or microspores have been developed, for example in wheat, barley, rice and oilseed rape. Also elimination of paternal chromosomes, occurring when making *Hordeum vulgare* L. × *H. bulbosum* L. or *Triticum aestivum* L. × *Zea mays* L. crosses, permits production of DH-lines. (The paternal chromosomes are lost in a few cell divisions of the hybrid zygote/embryo.) Note 3.2 comments further on DH-methods.

Note 3.2 DH-lines are mostly obtained directly from the gametes produced by the F_1-plants. This has a few drawbacks

- Recombination is restricted to the F_1 meiosis
- The proportion of DH-lines that are rejected because of poor performance is high. This is undesirable because of the cost of producing DH-lines.

To avoid these drawbacks one may use gametes from plants obtained by backcrossing the F_1 or one may use F_2- or even F_3-plants. (The latter

allows selection among F_2-plants, followed by selection among F_3-lines in the seedling stage). *In vitro* selection among the haploid embryos appeared to be feasible (Snape, 1997): the size and degree of embryo differentiation predicted which embryos would produce vigorous seedlings. Additionally the growth rate of the embryos was positively correlated with yield performance in the field $r = 0.3$, but this has found little practical application).

Continued self-fertilization is the natural mode of reproduction of self-fertilizing crops. There are many economically important self-fertilizing crops. A number of these are

Barley	*Hordeum vulgare* L.
Oats	*Avena sativa* L.
Wheat	*Triticum aestivum* L.
Rice	*Oryza sativa* L.
Sorghum	*Sorghum bicolor* (L.) Moench.
Finger millet	*Eleusine coracana* (L.) Gaertn.
Pea	*Pisum sativum* L.
Cowpea	*Vigna unguiculata* (L.) Walp.
Dry bean	*Phaseolus vulgaris* L.
Soybean	*Glycine max* (L.) Merr.
Peanut	*Arachis hypogaea* L.
Cotton	*Gossypium* spp.
Arabica coffee	*Coffea arabica* L.
Lettuce	*Lactuca sativa* L.
Tomato	*Lycopersicon esculentum* Mill.
Okra	*Abelmoschus esculentus* (L.) Moench.
Sweet pepper	*Capsicum annuum* L.

Self-fertilization is not always 100% in most of these **autogamous** crops, *e.g.* cotton, okra, sorghum. (The amount of outcrossing in sorghum is about 6%.) Section 3.5 considers the genotypic composition of populations reproducing by a mixture of self-fertilization and cross-fertilization.

Breeders regularly apply inbreeding in cross-fertilizing crops. They may have various reasons for doing this:

- The development of pure lines (mostly by continued selfing) for use as parents in the breeding of hybrid varieties, *e.g.* in maize or cucumber
- To promote the efficiency of elimination of an undesired recessive gene (Section 6.3.2)
- Maintenance of a genic male sterile 'line' (Note 3.3).

Note 3.3 FS-mating occurs also when a maintaining a genic male sterile barley 'line': male sterile plants are harvested after having been pollinated by their male fertile full sibs. (This is also applied in the case of recurrent

selection in self-fertilizing cereals (Koch and Degner, 1977)). Thus the harvesting of a female plant (say genotype mm) implies harvest of seed due to the cross $mm \times Mm$ (where Mm represents the genotype assumed for hermaphroditic plants). The genotypic composition of the obtained FS-family is ($\frac{1}{2}$, $\frac{1}{2}$, 0). Repeated application of this procedure implies repeated FS-mating.

The most powerful form of inbreeding of cross-fertilizing crops, *e.g.* **dioecious** crops, occurs with repeated crossing of the type

(i) full sib × full sib, *i.e.* full sib mating, or
(ii) parent $\overset{\times}{\times}$ offspring.

Full sib mating
The offspring due to a cross of two genotypes constitutes a family. The plants belonging to the family share both their maternal and their paternal parent. With regard to each other these plants are full sibs. Together they form a full sib family **(FS-family)**. Crossing of plants belonging to the same FS-family is called full sib mating **(FS-mating)**.

FS-mating may be used when inbreeding of dioecious crops, such as spinach or asparagus, is the aim. It occurs spontaneously in the case of open pollination within FS-families grown in isolation. This is applied in hermaphroditic, monoecious or dioecious crops in the case of separated FS-family selection (Section 6.3.3). Note 3.3 describes how FS-mating is applied when maintaining a genic male sterile 'line'.

Parent $\overset{\times}{\times}$ offspring mating

In this book the notation A$\overset{\times}{\times}$B indicates the cross A × B and/or the reciprocal cross B × A. Parent $\overset{\times}{\times}$ offspring crosses, *i.e.* so-called **PO-mating**, can only be applied to perennial crops such as oil palm (producing gametes from the age of 4–5 years for many years; see Note 3.4) or asparagus (with a juvenile phase lasting two years). The parent is still alive when its offspring reach the reproductive phase.

Note 3.4 Oil palm (*Elaeis guineensis* Jacq.) is not really a dioecious crop. Each individual palm continuously alternates phases when the palm produces exclusively female inflorescences and then a phase of exclusively male inflorescences. By storing pollen it is possible to apply self-fertilization.

Repeated backcrossing implies continued application of crosses of the type 'recurrent parent $\overset{\times}{\times}$ offspring'. In the absence of selection the genotype of the offspring becomes identical to the genotype of the recurrent parent (if the recurrent parent has a homozygous genotype) or to the genotypic composition of the possible lines obtained by selfing of the recurrent parent (if the recurrent parent is heterozygous, see Section 4.2).

In this chapter only loci segregating for not more than two alleles per locus will be considered. A justification for this was given in Section 2.2.1. For an extensive treatment of the population genetics theory of inbreeding the reader is referred to Allard, Jain and Workman (1968).

3.2 Diploid Chromosome Behaviour and Inbreeding

3.2.1 One locus with two alleles

With continued inbreeding of any (infinitely) large population the genotype frequencies will change from one generation to the other until the frequency of plants with a heterozygous genotype has become zero. Starting from the initial population G_0 with genotypic composition $(f_{0,0}, f_{1,0}, f_{2,0})$, eventually a population with genotypic composition $(q, 0, p)$ will be obtained. Table 3.1 (a)

Table 3.1 The frequency of genotypes aa, Aa and AA in the case of continued selfing

(a) Starting with some arbitrary genotypic composition

	Genotype		
Generation	aa	Aa	AA
S_0	f_0	f_1	f_2
S_1	$f_0 + \frac{1}{4}f_1$	$\frac{1}{2}f_1$	$f_2 + \frac{1}{4}f_1$
S_2	$f_0 + (\frac{1}{4} + \frac{1}{8})f_1$	$\frac{1}{4}f_1$	$f_2 + (\frac{1}{4} + \frac{1}{8})f_1$
S_3	$f_0 + (\frac{1}{4} + \frac{1}{8} + \frac{1}{16})f_1$	$\frac{1}{8}f_1$	$f_2 + (\frac{1}{4} + \frac{1}{8} + \frac{1}{16})f_1$
.			
.			
S_∞	q	0	p

(b) Starting with F_1, *i.e.* a population with genotypic composition (0, 1, 0)

Generation		*Inbreeding*	*Panmictic*	*Genotype*		
(t)	*Population*	*coefficient* ($\mathcal{F}$)	*index* (P)	aa	Aa	AA
0	$S_0(=F_1)$	-1	2	0	1	0
1	$S_1(=F_2)$	0	1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
2	$S_2(=F_3)$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{2}{8}$	$\frac{3}{8}$
3	$S_3(=F_4)$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{7}{16}$	$\frac{2}{16}$	$\frac{7}{16}$
4	$S_4(=F_5)$	$\frac{7}{8}$	$\frac{1}{8}$	$\frac{15}{32}$	$\frac{2}{32}$	$\frac{15}{32}$
5	$S_5(=F_6)$	$\frac{15}{16}$	$\frac{1}{16}$	$\frac{31}{64}$	$\frac{2}{64}$	$\frac{31}{64}$
6	$S_6(=F_7)$	$\frac{31}{32}$	$\frac{1}{32}$	$\frac{63}{128}$	$\frac{2}{128}$	$\frac{63}{128}$
7	$S_7(=F_8)$	$\frac{63}{64}$	$\frac{1}{64}$	$\frac{127}{256}$	$\frac{2}{256}$	$\frac{127}{256}$
∞	$S_\infty(=F_\infty)$	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$

illustrates this for inbreeding by means of continued selfing. It appears that the genotype frequencies approach, in an asymptotic manner, the gene and haplotype frequencies.

Often the frequency of heterozygous plants in generation t, *i.e.* $f_{1,t}$, is written in the form

$$2pq(1-\mathscr{F}_t)$$

(Wright, 1951). In this expression the factor $1-\mathscr{F}_t$ describes the deviation from the Hardy–Weinberg frequency. The factor is called the **panmictic index**, sometimes designated by the symbol P. This implies that $P = 1 - \mathscr{F}_t$. The parameter $\mathscr{F}_t$, say 'script F', is the **inbreeding coefficient** (or **fixation index**) pertaining to generation t.

When starting with an F_1 population, F_2 is the first generation due to self-fertilization. For this reason the F_2 population is chosen to be generation 1. (Its genotypic composition is equal to the genotypic composition of the population obtained by panmictic reproduction of the F_1; Note 2.4.) Successive generations may be indicated by $G_1, G_2, \ldots$, but in the case of continued selfing the designations $S_1, S_2, S_3, \ldots$ are used as well (Table 3.1).

A general description of the genotypic composition of any population (inbred or not) is now given by

	Genotype			(3.1)
	aa	Aa	AA	
f	$q^2 + pq\mathscr{F}_t$	$2pq(1-\mathscr{F}_t)$	$p^2 + pq\mathscr{F}_t$	

In several other books, *e.g.* Falconer and MacKay (1996), the inbreeding coefficient is defined as the probability that the two alleles at any loci of a plant are identical by descent. This would mean that the inbreeding coefficient of an F_2 population obtained from cross $AA \times aa$ is equal to $\frac{1}{2}$, because 50% of the plants contain, for locus A-a, alleles that are identical by descent (this concerns plants with genotype aa or AA). In this book the parameter $\mathscr{F}$ is used to quantify the deviations from the Hardy–Weinberg frequencies. In an F_2 population such deviations are absent and accordingly its inbreeding coefficient is 0. In Note 3.5 it is shown that our definition of the inbreeding coefficient $\mathscr{F}$ can be interpreted as the coefficient of correlation of numerical values, *e.g.* gene-effects, assigned to the haplotypes of the uniting gametes. This is based on the following consideration. With random mating the gene effects of the haplotypes of fusing female and male gametes are independent; in the absence of random mating they are interdependent. With inbreeding they tend to be similar; with outbreeding they tend to be different.

Breeding of self-fertilizing crops starts mostly with crossing of homozygous lines. For all loci for which the parental lines have a different homozygous genotype the genotype of the F_1 is heterozygous. For these loci $p = q = \frac{1}{2}$ and then the expressions in (3.1) simplify to

Note 3.5 When assigning arbitrary numerical values to haplotypes of the gametes one can calculate the coefficient of correlation between the value assigned to the haplotype of an egg and the value assigned to the haplotype of the pollen grain fusing with it. This is elaborated for the multiple allelic locus B_1-B_2-$\cdots$-B_n, with allele frequencies $p_1, p_2, \cdots, p_n$.

The genotypic composition is given in the central part of the following two-way table. The margins of the table present the haplotypic compositions of the gametes, as well as the numerical values $\alpha_1, \cdots, \alpha_n$ assigned to haplotypes $B_1, \cdots, B_n$. (One may, *e.g.*, use the gene effects as defined in Section 8.3.3).

The value of a female gamete is represented by random variable $\underline{x}$, the value of a male gamete by random variable $\underline{y}$.

		Haplotype pollen ($\underline{y}$)			
		$B_1(\alpha_1)$	$B_2(\alpha_2)$	$\cdots$ $B_n(\alpha_n)$	
Haplotype egg ($\underline{x}$)	$B_1(\alpha_1)$	$p_1{}^2 + p_1(1-p_1)\mathscr{F}$	$p_1p_2(1-\mathscr{F})$	$p_1p_n(1-\mathscr{F})$	p_1
	$B_2(\alpha_2)$	$p_1p_2(1-\mathscr{F})$	$p_2{}^2 + p_2(1-p_2)\mathscr{F})$	$p_2p_n(1-\mathscr{F})$	p_2
	.				
	$B_n(\alpha_n)$	$p_np_1(1-\mathscr{F})$	$p_np_2(1-\mathscr{F})$	$p_n{}^2 + p_n(1-p_n)\mathscr{F}$	p_n
		p_1	p_2	p_n	1

The random variables $\underline{x}$ and $\underline{y}$ are isomorous; thus $\mathrm{E}\underline{x} = \mathrm{E}\underline{y}, \mathrm{E}\underline{x}^2 = \mathrm{E}\underline{y}^2$ and $\sigma_x = \sigma_y$. The expression for the coefficient of correlation simplifies therefore as follows:

$$\rho_{x,y} = \frac{cov(\underline{x}, \underline{y})}{\sigma_x \sigma_y} = \frac{E\underline{x}\,\underline{y} - (E\underline{x})^2}{E\underline{x}^2 - (E\underline{x})^2}$$

As $\mathrm{E}\underline{x}\,\underline{y} = \sum_{i=1}^{n} \left[p_i{}^2 + p_i(1-p_i)\mathscr{F}\right] \alpha_i^2 + \sum_{i=1}^{n} \sum_{j=1:j\neq i}^{n} p_i p_j (1-\mathscr{F})\alpha_i\alpha_j, (\mathrm{E}\underline{x})^2 = \left(\sum_{i=1}^{n} p_i\alpha_i\right)^2$, and $\mathrm{E}\underline{x}^2 = \sum_{i=1}^{n} p_i{\alpha_i}^2$ it follows that

$$\mathrm{E}\underline{x}\cdot\underline{y} - (\mathrm{E}\underline{x})^2 = \mathscr{F}\left[\sum_{i=1}^{n} p_i(1-p_i){\alpha_i}^2 - \sum_{i=1}^{n}\sum_{j=1;j\neq i}^{n} p_i p_j \alpha_i\alpha_j\right] = \mathscr{F}(\mathrm{E}\underline{x}^2 - (\mathrm{E}\underline{x})^2).$$

This implies that $\rho = \mathscr{F}$; the coefficient of correlation appears to be equal to the inbreeding coefficient!

	Genotype		
	aa	Aa	AA
f	$\frac{1}{4}(1+\mathscr{F}_t)$	$\frac{1}{2}(1-\mathscr{F}_t)$	$\frac{1}{4}(1+\mathscr{F}_t)$

(3.2)

As $f_{1,0} = \frac{1}{2}(1-\mathscr{F}_0) = 1$, it follows that $\mathscr{F}_0 = -1$, *i.e.* a negative value for the inbreeding coefficient. The panmictic index of the F_1 amounts for heterozygous loci to $P_0 = 2$.

In the remainder of this section the decrease in the frequency of heterozygous plants is considered for the three most important regular inbreeding systems, *viz.* self-fertilization, full sib mating and parent × offspring mating. To measure this decrease the parameter λ is defined:

$$\lambda = \frac{2pq(1-\mathscr{F}_t)}{2pq(1-\mathscr{F}_{t-1})} = \frac{1-\mathscr{F}_t}{1-\mathscr{F}_{t-1}} \tag{3.3}$$

This parameter indicates the frequency of heterozygous plants as a proportion of this frequency in the preceding generation. At a smaller value for λ the decrease of f_1 is stronger. In the case of selfing the values for λ do not depend on t; they are approximately constant when applying full sib mating or parent × offspring. Then $\lambda_1 = \lambda_2 = \cdots = \lambda_t$. This implies

$$f_{1,t} = \lambda f_{1,t-1} = \lambda^2 f_{1,t-2} = \lambda^t f_{1,0}$$

Self-fertilization
In the F_2 generation, the first generation generated by selfing, the genotype frequencies coincide with the Hardy-Weinberg frequencies. Thus $f_{1,1} = 2pq$, implying that $\mathscr{F}_1$, the inbreeding coefficient of F_2, is zero. In population F_∞, approximately obtained after a very large number of generations reproducing by means of selfing, there is complete homozygosity, *i.e.* $f_{1,\infty} = 0$, implying that $\mathscr{F}_\infty$, the inbreeding coefficient of F_∞, is 1.

The decrease of f_1, due to continued selfing, is indicated in Table 3.1(a). The table shows that f_1 is halved by each round of reproduction by means of selfing. Thus

$$1-\mathscr{F}_t = \tfrac{1}{2}(1-\mathscr{F}_{t-1})$$

implying

$$\mathscr{F}_t = \tfrac{1}{2}(1+\mathscr{F}_{t-1}) \tag{3.4}$$

With regard to continued selfing the expression

$$1-\mathscr{F}_t = \tfrac{1}{2}(1-\mathscr{F}_{t-1})$$

or

$$P_t = \tfrac{1}{2}P_{t-1}$$

implies

$$P_t = (\tfrac{1}{2})^t P_0 = (\tfrac{1}{2})^{t-1}$$

i.e.

$$\mathscr{F}_t = 1-(\tfrac{1}{2})^{t-1} \tag{3.5}$$

(see Table 3.1(b)). At all other systems of inbreeding the reduction of f_1 is smaller. The minimum value for λ is thus attained with selfing. It amounts to $\lambda_S = \frac{1}{2}$.

Full sib mating and parent × offspring mating
Li (1976, pp. 312–317) showed that for both full sib mating and parent × offspring mating, the relation

$$f_{1,t+2} = \tfrac{1}{2} f_{1,t+1} + \tfrac{1}{4} f_{1,t} \tag{3.6}$$

applies.

Consider an initial population with genotypic composition (0,1,0), thus $f_{1,0} = 1$. In this population plants are crossed in pairwise combinations. In the next generation the genotypic composition of the population obtained, which consists of full sib families, is expected to be $(\frac{1}{2}, \frac{1}{4}, \frac{1}{2})$, with $f_{1,1} = \frac{1}{2}$. Continued full sib mating, within the continuously generated FS-families, gives, according to Equation (3.6)

$$\begin{aligned} f_{1,2} &= \tfrac{1}{2}(\tfrac{1}{2}) + \tfrac{1}{4}(1) = \tfrac{1}{2},\ i.e.\ \lambda_2 = 1 \\ f_{1,3} &= \tfrac{1}{2}(\tfrac{1}{2}) + \tfrac{1}{4}(\tfrac{1}{2}) = \tfrac{3}{8},\ i.e.\ \lambda_3 = \tfrac{3}{4} = 0.75 \\ f_{1,4} &= \tfrac{1}{2}(\tfrac{3}{8}) + \tfrac{1}{4}(\tfrac{1}{2}) = \tfrac{5}{16},\ i.e.\ \lambda_4 = \tfrac{5}{6} = 0.8333, etc. \end{aligned}$$

The first round of inbreeding (full sib mating or parent × offspring mating) does not give a decrease of the frequency of heterozygous plants ($\lambda_2 = 1$). Indeed, with full sib mating first FS-families have to be generated.

It appears that λ approaches asymptotically the value $\lambda_{FS} = \lambda_{PO} = 0.809$. As $(0.809)^3 = 0.53 \approx \frac{1}{2}$, three generations of reproduction by means of FS-mating or parent × offspring mating give the same reduction in f_1 as a single round of reproduction by selfing.

3.2.2 A pair of linked loci

In Chapter 1 it was shown that linkage may be expected to play a relatively unimportant role in the inheritance of quantitative traits. It was said that, throughout this book, absence of linkage would be assumed. It is, nevertheless, useful to be familiar with some implications of linkage. An important reason for this is the study of the linkage of loci affecting a quantitative trait with molecular markers.

Consider haplotypes *ab*, *aB*, *Ab or AB* for the two loci *A-a* and *B-b* with recombination value r_c. Continued selfing, starting with an F_1 with the heterozygous genotype *AaBb*, yields in the absence of selection 'symmetric' haplotype frequencies:

$$g_{11,t} = g_{00,t}$$

and

$$g_{01,t} = g_{10,t}$$

Because

$$g_{11,t} + g_{10,t} = p_A = \tfrac{1}{2}$$

we get

$$g_{10,t} = \tfrac{1}{2} - g_{11,t}$$

This implies that, when one knows $g_{11,t}$, one also knows $g_{10,t}$, $g_{01,t}$ and $g_{00,t}$. It suffices thus to consider only the frequency of gametes with the AB haplotype. This is particularly of interest when considering F_∞. This population is described by

	Genotype			
	$aabb$	$AAbb$	$aaBB$	$AABB$
f	$f_{00,\infty}$	$f_{20,\infty}$	$f_{02,\infty}$	$f_{22,\infty}$

Only plants with the $AABB$ genotype are capable of producing gametes with the AB haplotype. Thus $g_{11,\infty} = f_{22,\infty}$. The haplotypic composition of the gametes produced by this population is

	Haplotype			
	ab	Ab	aB	AB
g	$g_{00,\infty}(= g_{11,\infty})$	$g_{10,\infty}(= \frac{1}{2} - g_{11,\infty})$	$g_{01,\infty}(= \frac{1}{2} - g_{11,\infty})$	$g_{11,\infty}$

There are thus good reasons to consider the frequency of gametes with the AB haplotype. In Note 3.6 the following relation between the frequencies of AB-haplotypes in two successive generations is derived:

Note 3.6 The frequency of AB haplotypes, *i.e.* g_{11}, is considered for the case of continued autogamous reproduction. (To promote readability the recombination value is – in this section – mostly just indicated by the symbol r). The genotypes capable of producing AB haplotypes, their frequencies in generation t and the haplotypic composition of the gametes they produce are

		Haplotype			
Genotype	f	ab	aB	Ab	AB
$AABB$	$f_{22,t}$	0	0	0	1
$AABb$	$f_{21,t}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$
$AaBB$	$f_{12,t}$	0	$\frac{1}{2}$	0	$\frac{1}{2}$
AB/ab	$f_{11\mathrm{C},t}$	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$	$\frac{1}{2}r$	$\frac{1}{2}(1-r)$
Ab/aB	$f_{11\mathrm{R},t}$	$\frac{1}{2}r$	$\frac{1}{2}(1-r)$	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$

Then

$$\begin{aligned} g_{11,t+1} &= f_{22,t} + \tfrac{1}{2}f_{21,t} + \tfrac{1}{2}f_{12,t} + \tfrac{1}{2}(1-r)f_{11\mathrm{C},t} + \tfrac{1}{2}rf_{11\mathrm{R},t} \\ &= f_{22,t} + \tfrac{1}{2}f_{21,t} + \tfrac{1}{2}f_{12,t} - \tfrac{1}{2}r(f_{11\mathrm{C},t} - f_{11\mathrm{R},t}) \\ &= f_{22,t} + \tfrac{1}{2}f_{21,t} + \tfrac{1}{2}f_{12,t} + \tfrac{1}{2}f_{11\mathrm{C},t} - rd_t \end{aligned} \tag{3.7}$$

where, according to Equation (2.11), d_t is defined as

$$d_t = \tfrac{1}{2}(f_{11\mathrm{C},t} - f_{11\mathrm{R},t})$$

and

$$f_{22,t} = f_{22,t-1} + \tfrac{1}{4}f_{21,t-1} + \tfrac{1}{4}f_{12,t-1} + \tfrac{1}{4}(1-r)^2 f_{11\mathrm{C},t-1} + \tfrac{1}{4}r^2 f_{11\mathrm{R},t-1} \tag{3.8}$$

$$f_{21,t} = \tfrac{1}{2}f_{21,t-1} + \tfrac{1}{2}r(1-r)f_{11\mathrm{C},t-1} + \tfrac{1}{2}r(1-r)f_{11\mathrm{R},t-1} \tag{3.9}$$

$$f_{12,t} = \tfrac{1}{2}f_{12,t-1} + \tfrac{1}{2}r(1-r)f_{11\mathrm{C},t-1} + \tfrac{1}{2}r(1-r)f_{11\mathrm{R},t-1} \tag{3.10}$$

$$f_{11\mathrm{C},t} = \tfrac{1}{2}(1-r)^2 f_{11\mathrm{C},t-1} + \tfrac{1}{2}r^2 f_{11\mathrm{R},t-1} \tag{3.11}$$

$$f_{11\mathrm{R},t} = \tfrac{1}{2}r^2 f_{11\mathrm{C},t-1} + \tfrac{1}{2}(1-r)^2 f_{11\mathrm{R},t-1} \tag{3.12}$$

Thus

$$\begin{aligned} g_{11,t+1} &= f_{22,t-1} + (\tfrac{1}{4} + \tfrac{1}{4})f_{21,t-1} + (\tfrac{1}{4} + \tfrac{1}{4})f_{12,t-1} + [\tfrac{1}{4}(1-r)^2 \\ &\quad + \tfrac{1}{4}r(1-r) + \tfrac{1}{4}r(1-r) + \tfrac{1}{4}(1-r)^2]f_{11\mathrm{C},t-1} \\ &\quad + [\tfrac{1}{4}r^2 + \tfrac{1}{4}r(1-r) + \tfrac{1}{4}r(1-r) + \tfrac{1}{4}r^2]f_{11\mathrm{R},t-1} - rd_t \\ &= f_{22,t-1} + \tfrac{1}{2}f_{21,t-1} + \tfrac{1}{2}f_{12,t-1} \\ &\quad + (\tfrac{1}{2} - r + \tfrac{1}{2}r^2 + \tfrac{1}{2}r - \tfrac{1}{2}r^2)f_{11\mathrm{C},t-1} \\ &\quad + (\tfrac{1}{2}r^2 + \tfrac{1}{2}r - \tfrac{1}{2}r^2)f_{11\mathrm{R},t-1} - rd_t \\ &= f_{22,t-1} + \tfrac{1}{2}f_{21,t-1} + \tfrac{1}{2}f_{12,t-1} + \tfrac{1}{2}(1-r)f_{11\mathrm{C},t-1} \\ &\quad + \tfrac{1}{2}rf_{11\mathrm{R},t-1} - rd_t \\ &= g_{11,t} - rd_t \end{aligned} \tag{3.13}$$

(This equation is identical to Equation (2.10d), derived for the case of continued panmictic reproduction.)

$$g_{11,t+1} = g_{11,t} - r_c d_t \tag{3.13}$$

Equation (3.13) applies at continued self-fertilization. It is identical to Equation (2.10d) applying at continued panmictic reproduction. One should realize, however, that with panmictic reproduction the relation between d_{t+1} and d_t was derived to be

$$d_{t+1} = (1 - r_c)d_t$$

(see Equation (2.13)). For autogamous reproduction, however, the relation between d_t and d_{t-1} can be shown (see Note 3.7) to be

$$d_{t+1} = \left(\frac{1 - 2r_c}{2}\right) d_t \tag{3.14}$$

Note 3.7 In the case of (continued) selfing, plants with a doubly heterozygous genotype, in the coupling phase or in the repulsion phase, can only be produced by doubly heterozygous parents, one can easily derive from Table 2.2 that:

$$f_{11C,t+1} = 2\left(\frac{1-r}{2}\right)^2 f_{11C,t} + 2\left(\frac{r}{2}\right)^2 f_{11R,t} \tag{3.15}$$

$$f_{11R,t+1} = 2\left(\frac{1-r}{2}\right)^2 f_{11R,t} + 2\left(\frac{r}{2}\right)^2 f_{11C,t} \tag{3.16}$$

Thus:

$$\begin{aligned} f_{11,t+1} &= \left[2\left(\frac{1-r}{2}\right)^2 + 2\left(\frac{r}{2}\right)^2\right](f_{11C,t} + f_{11R,t}) \\ &= (r^2 - r + \tfrac{1}{2})f_{11,t} = \left[\left(r - \tfrac{1}{2}\right)^2 + \tfrac{1}{4}\right] f_{11,t} \end{aligned}$$

Equation (2.11), *i.e.*

$$d_{t+1} = \tfrac{1}{2}(f_{11\mathrm{C},t+1} - f_{11\mathrm{R},t+1})$$

yields thus

$$d_{t+1} = \tfrac{1}{4}[(1-r)^2 - r^2](f_{11C,t} - f_{11R,t})$$

This gives Equation (3.14), *viz.*

$$d_{t+1} = \left(\frac{1-2r_c}{2}\right) d_t$$

implying:

$$d_t = \left(\frac{1-2r_c}{2}\right)^{t-1} d_1 \tag{3.17}$$

Equations (3.13) and (3.14) yield for the case of continued selfing:

$$g_{11,t+1} = g_{11,t} - r_c\left(\frac{1-2r_c}{2}\right) d_{t-1} \tag{3.18}$$

The parameter d_t is still, as defined in Equation (2.11), equal to $\frac{1}{2}(f_{11\mathrm{C},t} - f_{11\mathrm{R},t})$. Equation (3.18) shows that, unless $d_t = 0$ or $r_c = \frac{1}{2}$, the haplotype frequencies will change from one generation to the next.

The genotypic composition of F_∞, for F_1 in coupling phase as well as in repulsion phase, depends directly on Equation (3.19), *viz.*

$$g_{11,\infty} = f_{22,\infty} = g_{11,1} - \left(\frac{2r}{1+2r}\right) d_1 \tag{3.19}$$

which is derived in Note 3.8.

Note 3.8 Equation (3.13) combined with Equation (3.17) yields in the case of continued selfing

$$g_{11,t+1} - g_{11,t} = -rd_1 \left(\frac{1-2r}{2}\right)^{t-1}$$

Repeated application of this equation results via

$$\begin{aligned} g_{11,2} - g_{11,1} &= -rd_1 \left(\frac{1-2r}{2}\right)^{0} \\ g_{11,3} - g_{11,2} &= -rd_1 \left(\frac{1-2r}{2}\right)^{1} \\ &\vdots \\ g_{11,t+1} - g_{11,t} &= -rd_1 \left(\frac{1-2r}{2}\right)^{t-1} \end{aligned}$$

in

$$g_{11,t+1} - g_{11,1} = -rd_1 \sum_{j=0}^{t-1} \left(\frac{1-2r}{2}\right)^{j}$$

The sum of the terms of this geometric series is

$$\frac{1 - \left(\frac{1-2r}{2}\right)^{t-1}}{1 - \left(\frac{1-2r}{2}\right)} = \frac{2}{1+2r}\left[1 - \left(\frac{1-2r}{2}\right)^{t-1}\right]$$

Thus

$$g_{11,t+1} = g_{11,1} - r\left(\frac{2}{1+2r}\right) \cdot d_1 \cdot \left[1 - \left(\frac{1-2r}{2}\right)^{t-1}\right]$$

implying

$$g_{11,\infty} = f_{22,\infty} = g_{11,1} - \left(\frac{2r}{1+2r}\right) d_1$$

The quantity to be substituted in Equation (3.19) for d_1 amounts, according to Example 2.7, to $\frac{1}{4}(1-2r)$ for F_1 in the coupling phase and to $\frac{-1}{4}(1-2r)$ for F_1 in the repulsion phase. Equation (3.19) yields thus for F_1 in the coupling phase:

$$g_{11,\infty} = f_{22,\infty} = \left(\frac{1-r}{2}\right) - \left(\frac{2r}{1+2r}\right)\left(\frac{1-2r}{4}\right) = \frac{1}{2(1+2r)} \tag{3.20}$$

For F_1 in the repulsion phase we get

$$g_{11,\infty} = f_{22,\infty} = \left(\frac{r}{2}\right) + \left(\frac{2r}{1+2r}\right)\left(\frac{1-2r}{4}\right) = \frac{2r}{2(1+2r)} \tag{3.21}$$

Table 3.2 The genotypic composition of F_∞ with regard to the complex genotypes for the two linked loci *A-a* and *B-b*

(a) F_1 in coupling phase

	bb	Bb	BB	
aa	$\frac{1}{2(1+2r_c)}$	0	$\frac{2r_c}{2(1+2r_c)}$	$\frac{1}{2}$
Aa	0	0	0	0
AA	$\frac{2r_c}{2(1+2r_c)}$	0	$\frac{1}{2(1+2r_c)}$	$\frac{1}{2}$
	$\frac{1}{2}$	0	$\frac{1}{2}$	1

(b) F_1 in repulsion phase

	bb	Bb	BB	
aa	$\frac{2r_c}{2(1+2r_c)}$	0	$\frac{1}{2(1+2r_c)}$	$\frac{1}{2}$
Aa	0	0	0	0
AA	$\frac{1}{2(1+2r_c)}$	0	$\frac{2r_c}{2(1+2r_c)}$	$\frac{1}{2}$
	$\frac{1}{2}$	0	$\frac{1}{2}$	1

Table 3.2 presents the genotypic composition of F_∞. It may be compared with Table 2.1 presenting the genotypic composition obtained after continued panmixis.

In the case of linkage ($0 < r_c < \frac{1}{2}$) the frequencies of the haplotypes change in the course of the generations. For gametes with the AB haplotype the difference between $g_{11,1}$ and $g_{11,\infty}$ amounts to

$$g_{11,\infty} - g_{11,1} = \left(\frac{2r}{1+2r}\right) d_1$$

This amounts, according to Example 2.7, for F_1 in the coupling phase to

$$\left(\frac{2r}{1+2r}\right)\left(\frac{1-2r}{4}\right) = \frac{r(1-2r)}{2(1+2r)}$$

and for F_1 in the repulsion phase to

$$\left(\frac{2r}{1+2r}\right)\left(\frac{2r-1}{4}\right) = \frac{r(2r-1)}{2(1+2r)}$$

These differences are for $0 < r_c < \frac{1}{2}$ generally quite small. For $r_c = \frac{1}{4}$, for instance, it amounts for F_1 in the repulsion phase to $g_{11,1} - g_{11,\infty} = \frac{1}{8} - \frac{1}{6} = -0.0417$.

We consider now the frequency of plants with a genotype obtained by crossing two parents. It may, for example, be desired to obtain genotype $AABB$ from an initial cross of genotypes $AAbb$ and $aaBB$. The frequency of $AABB$ plants amounts in population F_2 to $f_{22,1} = \frac{1}{4}{r_c}^2$ (Table 2.2). Equation (3.8)

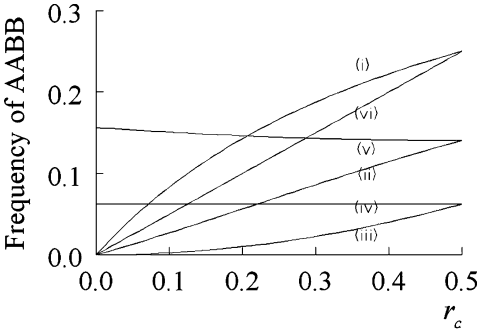


Fig. 3.1 The frequency of plants with genotype $AABB$ as a function of the recombination value r_c. Considered are populations obtained by crossing of genotypes $AAbb$ and $aaBB$ followed by (i) continued self-fertilization until F_∞, (ii) selfing until F_3, (iii) selfing until F_2, (iv) continued panmixis until linkage equilibrium, (v) continued panmixis followed by one round of reproduction by means of selfing, or (vi) doubling of the number of chromosomes in the gametes produced by F_1

yields for $t = 2$ the frequency of plants with genotype $AABB$ in F_3. When substituting the F_2 genotype frequencies presented in Table 2.2 one gets for an F_1 in the repulsion phase:

$$\begin{aligned} f_{22,2} &= \tfrac{1}{4}r^2 + \tfrac{1}{8}r(1-r) + \tfrac{1}{8}r(1-r) + \tfrac{1}{8}r^2(1-r)^2 + \tfrac{1}{8}(1-r)^2r^2 \\ &= \tfrac{1}{4}r + \tfrac{1}{4}r^2 - \tfrac{1}{2}r^3 + \tfrac{1}{4}r^4 \end{aligned} \tag{3.22}$$

This amounts, for unlinked loci, to $f_{22,2} = \frac{9}{64} = \left(\frac{3}{8}\right)^2 = f_{00,2}$. According to Equation (3.21) the frequency of $AABB$ plants in F_∞ is $\frac{2r}{2(1+2r)}$.

Because $\frac{2r}{2(1+2r)} \leq \frac{1}{2(1+2r)}$, plants with one of the parental genotypes will outnumber plants with this recombinant genotype to a greater extent as linkage is stronger, *i.e.* as r_c is smaller. In Figure 3.1 curves (i), (ii) and (iii) show the values for f_{22} in F_∞, F_3 and F_2 as a function of r_c. Recombination of alleles belonging to two different loci can only occur at meiosis of doubly heterozygous genotypes. In populations of cross-fertilizing crops, doubly heterozygous genotypes tend to be permanently present; in populations of self-fertilizing crops they disappear.

One should, however, be careful when speaking about 'the recombining effect of cross-fertilization'. This is illustrated for loci *A-a* and *B-b*.

Continued panmictic reproduction gives eventually, at linkage equilibrium, $f_{22} = p^2r^2$. This amounts for $p = r = \frac{1}{2}$ to $\frac{1}{16}$, whatever the recombination value (Fig. 3.1(iv)). For tightly linked loci, with $r_c < \frac{1}{14}$, genotype $AABB$ will indeed occur with a higher frequency in populations in linkage equilibrium than in populations obtained by continued selfing. For less tightly linked loci, *i.e.* $r_c > \frac{1}{14}$, the frequency of $AABB$ will, however, be higher in F_∞. Thus one should not decide rashly to increase the frequency of plants with a recombinant genotype by the application of random mating in $F_2, F_3, \ldots$ populations of a self-fertilizing crop (Bos, 1977). With regard to unlinked loci continued

random mating will only result in the genotypic composition of F_2, because for unlinked loci the F_2 population obtained by selfing will have the linkage equilibrium composition (see Example 2.7).

Selection in a cross-fertilizing crop is more efficient when increasing the frequency of homozygous recombinant genotypes by selfing. According to Note 3.9 a single round of reproduction by means of self-fertilization in a population in linkage equilibrium gives

$$f_{22} = \frac{5 - 2r + 2r^2}{32}$$

(Fig. 3.1(v))

Note 3.9 Consider a population in linkage equilibrium. It is obtained by panmictic reproduction starting with a single-cross hybrid variety. With regard to loci *A*-*a* and *B*-*B* a single round of reproduction by means of selfing results, according to Equation (3.8), in the following frequency of plants with genotype *AABB*:

$$f_{22} = \tfrac{1}{16} + \tfrac{1}{4} \cdot \tfrac{1}{8} + \tfrac{1}{4} \cdot \tfrac{1}{8} + \tfrac{1}{4} r^2 \cdot \tfrac{1}{8} + \tfrac{1}{4}(1-r)^2 \cdot \tfrac{1}{8} = \tfrac{5-2r+2r^2}{32}$$

For $r = \frac{1}{2}$ this amounts to $\frac{9}{64}$, *i.e.* $\left(\frac{3}{8}\right)^2$. It is the same value as obtained, from Equation (3.22), for an F_3. The single reproduction by means of selfing gives thus the genotypic composition of an F_3. This illustrates that the genotypic composition of the population in linkage equilibrium is equal to the genotypic composition for pairs of unlinked loci in an F_2.

In a diploid crop, doubling the number of chromosomes of haploid plants is the fastest way to attain complete homozygosity. The frequency of plants with the desired recombinant genotype then amounts to $\frac{1}{2} r_c$, *i.e.* $\frac{2}{r_c}$ times as high as in F_2 (Fig. 3.1(vi)).

The frequency of doubly heterozygous plants is greatly reduced with reproduction by means of selfing. Depending on the recombination value, a single round of selfing reduces this frequency to only $\frac{1}{4}$ to $\frac{1}{2}$ of the frequency of plants with the *AaBb* genotype in the preceding generation. Note 3.8 shows that the remaining portion of doubly heterozygous plants amounts to $\frac{f_{11,t+1}}{f_{11,t}} = (r - \frac{1}{2})^2 + \frac{1}{4}$, which amounts to $\frac{1}{4}$ for $r_c = \frac{1}{2}$ and to $\frac{1}{2}$ for $r_c = 0$. This reduction of the frequency of heterozygous plants is even stronger for more complex genotypes: a single round of selfing reduces the frequency of the complex genotype consisting of a heterozygous single-locus genotype for each of k unlinked loci to the portion $(\frac{1}{2})^k$ of its preceding value.

3.2.3 Two or more unlinked loci, each with two alleles

Independent segregation occurs when the recombination value is equal to $\frac{1}{2}$. Some population genetical implications of continued selfing with regard to unlinked loci are thus easily obtained from results derived in Section 3.2.2.

Two unlinked loci
Consider the haplotypes ab, aB, Ab or AB for the two unlinked loci A-a and B-b. Equation (3.18) shows that absence of linkage implies constancy of the haplotype frequencies:

$$\begin{aligned} g_{00,t+1} &= g_{00,t} \\ g_{01,t+1} &= g_{01,t} \\ g_{10,t+1} &= g_{10,t} \\ g_{11,t+1} &= g_{11,t} \end{aligned}$$

This applies for any genotypic composition of the initial population. An application is described in Note 3.10. The haplotypic composition of the gametes produced by populations $S_0, S_1, \ldots, S_\infty$ remains thus constant across the generations. This implies that the genotypic composition of S_∞ immediately follows from the haplotypic composition of the gametes produced by S_0:

Note 3.10 When breeding a *non-perennial* cross-fertilizing crop, selection among plants on the basis of a progeny test (see Section 6.3.6) is impossible because the candidate plants cannot be maintained. In such cases these plants are selfed: their S_1-lines produce gametes with the same haplotypic composition as they do themselves. Indeed: haplotypic compositions can be maintained by means of selfing. This is applied in *recurrent selection for general combining ability* as well as in *reciprocal recurrent selection* (see Section 11.3.2).

	Genotype			
	$aabb$	$aaBB$	$AAbb$	$AABB$
f	g_{00}	g_{01}	g_{10}	g_{11}

The constancy of the haplotypic composition in the case of continued selfing is in striking contrast to the continuous change, until linkage equilibrium is attained, of the haplotypic composition in the case of continued panmixis. Notwithstanding the stability of the haplotype frequencies the genotype frequencies change drastically: the frequencies of heterozygous plants decrease and those of homozygous plants increase. The frequencies of the complex genotypes only become stable if heterozygous plants no longer occur.

When starting with an F_1 the frequencies of the complex genotypes follow directly from the frequencies of the single-locus genotypes given by Equation (3.2). (It should be realized that in cross-fertilizing crops this rule applies only

Table 3.3 The frequencies of complex and single-locus genotypes for the unlinked loci *A-a* and *B-b* in generation $t(= 1, 2, 3, \ldots, \infty)$ produced by selfing for t generations since the F_1 population

		Genotype for locus *B-b*			
		bb	*Bb*	*BB*	
Genotype for locus *A-a*:	*aa*	$\frac{1}{16}(1+\mathscr{F}_t)^2$	$\frac{1}{8}(1-\mathscr{F}_t{}^2)$	$\frac{1}{16}(1+\mathscr{F}_t)^2$	$\frac{1}{4}(1+\mathscr{F}_t)$
	Aa	$\frac{1}{8}(1-\mathscr{F}_t{}^2)$	$1/4(1-\mathscr{F}_t)^2$	$\frac{1}{8}(1-\mathscr{F}_t{}^2)$	$\frac{1}{2}(1-\mathscr{F}_t)$
	AA	$\frac{1}{16}(1+\mathscr{F}_t)^2$	$\frac{1}{8}(1-\mathscr{F}_t{}^2)$	$\frac{1}{16}(1+\mathscr{F}_t)^2$	$\frac{1}{4}(1+\mathscr{F}_t)$
		$\frac{1}{4}(1+\mathscr{F}_t)$	$\frac{1}{2}(1-\mathscr{F}_t)$	$\frac{1}{4}(1+\mathscr{F}_t)$	1

in linkage equilibrium). Thus Table 3.3 presents the genotypic composition with regard to the complex genotypes for two unlinked loci of any generation obtained by (continued) selfing starting with an F_1.

K unlinked loci

It is, in general, impossible to determine how many loci control the phenotypic expression of a certain trait, *e.g.* culm length in wheat. The reason for this is that the contribution due to non-segregating loci cannot be assessed: if one crosses some line P_1 with genotype *AabbccDD* with regard to the trait under consideration with line P_2 with genotype *aabbCCdd* then the contribution due to locus *B-b* cannot be assessed. Thus it might appear that three instead of four loci are responsible for the genetic control of the trait. In fact only the number of segregating loci, *i.e.* the number of loci for which the two homozygous parents have a different genotype with regard to the trait under consideration, can be studied. This number is an interesting quantity, upon which the size of an F_2 generation (or a later generation) may be based. It is speculated that the analysis of (quantitative trait) loci based on molecular markers is going to substitute biometrical methods for estimating the number of segregating loci. When generating a large number of molecular markers one can localize (and count) polygenes with relatively large phenotypic effects on the studied trait.

We consider, for the case of K unlinked loci, the probability that a plant contains for $\underline{k}$ of these loci a heterozygous single-locus genotype and for the remaining $K - \underline{k}$ loci a homozygous genotype. This probability is given by the binomial probability distribution function:

$$P(\underline{k} = k) = \binom{K}{k} \cdot \left(\frac{1-\mathscr{F}_t}{2}\right)^k \left(\frac{1+\mathscr{F}_t}{2}\right)^{K-k}$$

The probability of a completely homozygous plant is

$$P(\underline{k} = 0) = \left(\frac{1+\mathscr{F}_t}{2}\right)^K$$

Table 3.4 The probability of a completely homozygous plant in generation $G_t (t = 1, \ldots, 7)$, obtained after t successive generations with reproduction by means of selfing, when considering $K = 1, \ldots, 14$ unlinked loci. G_t corresponds to generation F_{t+1}

	t						
K	1	2	3	4	5	6	7
1	0.500	0.750	0.875	0.938	0.969	0.984	0.992
2	0.250	0.563	0.766	0.879	0.938	0.969	0.984
3	0.125	0.422	0.670	0.824	0.909	0.954	0.977
4	0.063	0.316	0.586	0.772	0.881	0.939	0.969
5	0.031	0.237	0.513	0.724	0.853	0.924	0.962
6	0.016	0.178	0.449	0.679	0.827	0.910	0.954
7	0.008	0.133	0.393	0.637	0.801	0.896	0.947
8	0.004	0.100	0.344	0.597	0.776	0.882	0.939
9	0.002	0.075	0.301	0.559	0.751	0.868	0.932
10	0.001	0.056	0.263	0.524	0.728	0.854	0.925
11	0.000	0.042	0.230	0.492	0.705	0.841	0.917
12	0.000	0.032	0.201	0.461	0.683	0.828	0.910
13	0.000	0.024	0.176	0.432	0.662	0.815	0.903
14	0.000	0.018	0.154	0.405	0.641	0.802	0.896

or, when applying Equation (3.5)

$$\left[\frac{1+1-(\frac{1}{2})^{t-1}}{2}\right]^K = \left[1-(\tfrac{1}{2})^t\right]^K = \left(\frac{2^t-1}{2^t}\right)^K \tag{3.23}$$

Table 3.4 presents this probability for $K = 1, \ldots, 14$ and $t = 1, \ldots, 7$. Allard (1960, Fig. 6.1) gives a graphical presentation of these probabilities.

The expected value of $\underline{k}$, the number of loci with a heterozygous single-locus genotype in a random plant, is

$$\mathrm{E}\underline{k} = K \cdot \tfrac{1}{2}(1 - \mathscr{F}_t) = \tfrac{1}{2}K(\tfrac{1}{2})^{t-1} = (\tfrac{1}{2})^t K$$

It is $\frac{1}{2}K$ in an F_2 plant, $\frac{1}{4}K$ in an F_3 plant, *etc.*

The variance of $\underline{k}$ is

$$\begin{aligned}\mathrm{var}(\underline{k}) &= K \cdot \tfrac{1}{2}(1-\mathscr{F}_t) \cdot \tfrac{1}{2}(1+\mathscr{F}_t) \\ &= \tfrac{1}{4}K(1-\mathscr{F}_t^2) = \tfrac{1}{4}K[1-\{1-(\tfrac{1}{2})^{t-1}\}^2] \\ &= \tfrac{1}{4}K[1-\{1-(\tfrac{1}{2})^t+(\tfrac{1}{4})^{t-1}\}] = [(\tfrac{1}{2})^{t-2}-(\tfrac{1}{4})^t]K\end{aligned}$$

Example 3.1 illustrates an application to an F_5 population.

Example 3.1 The probability distribution for $\underline{k}$, the number of loci with a heterozygous single-locus genotype, among $K = 3$ loci is derived for plants belonging to an F_5 population. The relevant inbreeding coefficient is then $\mathscr{F}_4 = 1 - (\frac{1}{2})^3 = \frac{7}{8}$. The probability distribution is then

$$P(\underline{k} = k) = \binom{3}{k} \cdot \left(\frac{1}{16}\right)^k \left(\frac{15}{16}\right)^{K-k}$$

This gives:

$$\begin{aligned} \mathrm{P}(\underline{k} = 0) &= 0.8240 \\ \mathrm{P}(\underline{k} = 1) &= 0.1648 \\ \mathrm{P}(\underline{k} = 2) &= 0.0110 \\ \mathrm{P}(\underline{k} = 3) &= 0.0002 \end{aligned}$$

The expected value of $\underline{k}$, $\mathrm{E}\underline{k}$, is $(\frac{1}{2})^4 \cdot 3 = 0.1875$ and the variance of $\underline{k}$ across the F_5-plants amounts to $\mathrm{var}(\underline{k}) = [(\frac{1}{2})^4 - (\frac{1}{4})^3] \cdot 3 = 0.176$. (Otherwise: $\mathrm{var}(\underline{k}) = \mathrm{E}\underline{k}^2 - (\mathrm{E}\underline{k})^2 = [0.1648 + 0.0110 \times 2^2 + 0.0002 \times 3^2] - (0.1875)^2 = 0.176$).

3.3 Autotetraploid Chromosome Behaviour and Self-Fertilization

Spontaneous self-fertilization as the natural mode of reproduction occurs rather rarely among crops with an autotetraploid chromosome behaviour. The somatic chromosome number of quinoa (*Chenopodium quinoa*) is $2n = 36$. The basic chromosome number for the genus *Chenopodium* is $x = 9$. This suggests that quinoa is a tetraploid. Ward (2000) found for the same locus both diploid and tetraploid behaviour. Simmonds (1976) reported that selfing predominates, without evident inbreeding depression.

Quite a few autotetraploid crops, *e.g.* durum wheat (*Triticum durum*; $2n = 4x = 28$) or coffee (*Coffea arabica*; $2n = 4x = 44$), have a diploid chromosome behaviour. For other crops, *e.g.* European potato (*Solanum tuberosum*; $2n = 4x = 48$) or wild barley (*Hordeum bulbosum*; $2n = 4x = 28$), there may be a more or less perfect autotetraploid chromosome behaviour, implying that exclusively quadrivalents are being formed at meiosis. Artificial self-fertilization may be applied in a man-made autotetraploid crop such as rye (*Secale cereale*; $2n = 4x = 28$), which is self-incompatible in its natural diploid condition.

In this section attention is only given to the simple situation of a single segregating locus with two alleles. It is assumed that double reduction does not occur.

The genotypic composition of some initial generation, say S_0, is

	Genotype				
	aaaa	*Aaaa*	*AAaa*	*AAAa*	*AAAA*
	nulliplex	simplex	duplex	triplex	quadruplex
f	f_0	f_1	f_2	f_3	f_4

Its gene frequencies are

$$p = \tfrac{1}{4}f_1 + \tfrac{1}{2}f_2 + \tfrac{3}{4}f_3 + f_4 \tag{3.24}$$

and

$$q = 1 - p$$

It is first verified that the gene frequencies remain constant from one generation to the next (such constancy is to be expected in the absence of selection). In order to do this, Table 3.5 is used. This table presents, for each possible autotetraploid genotype, and according to the haplotype frequencies presented in Table 2.4, the genotypic composition of the line obtained by selfing.

The allele frequencies in the parental population follow from Equation (3.24). Across the total of the lines obtained from this parental population the frequency of allele A is

$$\tfrac{1}{4}\left(\tfrac{1}{2}f_1 + \tfrac{2}{9}f_2\right) + \tfrac{1}{2}\left(\tfrac{1}{4}f_1 + \tfrac{1}{2}f_2 + \tfrac{1}{4}f_3\right) + \tfrac{3}{4}\left(\tfrac{2}{9}f_2 + \tfrac{1}{2}f_3\right)$$
$$+ \left(\tfrac{1}{36}f_2 + \tfrac{1}{4}f_3 + f_4\right) = \tfrac{1}{4}f_1 + \tfrac{1}{2}f_2 + \tfrac{3}{4}f_3 + f_4$$

This is equal to the frequency in the parental population. The genotypic composition of S_∞ will thus be:

	Genotype				
	aaaa	*Aaaa*	*AAaa*	*AAAa*	*AAAA*
f	q	0	0	0	p

How fast do the frequencies of plants with a heterozygous genotype and of gametes with a heterozygous haplotype decrease with (continued) selfing?

Table 3.5 The genotypic composition of the line obtained by selfing an autotetraploid genotype

Parent		Genotypic composition of line				
genotype	f	*aaaa*	*Aaaa*	*AAaa*	*AAAa*	*AAAA*
aaaa	f_0	1	0	0	0	0
Aaaa	f_1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	0
AAaa	f_2	$\frac{1}{36}$	$\frac{2}{9}$	$\frac{1}{2}$	$\frac{2}{9}$	$\frac{1}{36}$
AAAa	f_3	0	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
AAAA	f_4	0	0	0	0	1

In order to answer this question, first the decrease of g_1, *i.e.* the frequency of gametes with haplotype Aa is considered and thereafter the decrease of f_h. *i.e.* the frequency of heterozygous plants. From Table 2.4 it can be derived that

$$g_{1,t+1} = \tfrac{1}{2} f_{1,t} + \tfrac{4}{6} f_{2,t} + \tfrac{1}{2} f_{3,t} \tag{3.25}$$

Thus, similarly

$$\begin{aligned} g_{1,t+2} &= \tfrac{1}{2} f_{1,t+1} + \tfrac{4}{6} f_{2,t+1} + \tfrac{1}{2} f_{3,t+1} = \tfrac{1}{2} \left(\tfrac{1}{2} f_{1,t} + \tfrac{2}{9} f_{2,t} \right) \\ &\quad + \tfrac{4}{6} \left(\tfrac{1}{4} f_{1,t} + \tfrac{1}{2} f_{2,t} + \tfrac{1}{4} f_{3,t} \right) + \tfrac{1}{2} \left(\tfrac{2}{9} f_{2,t} + \tfrac{1}{2} f_{3,t} \right) \\ &= \tfrac{5}{12} f_{1,t} + \tfrac{5}{9} f_{2,t} + \tfrac{5}{12} f_{3,t} = \tfrac{5}{6} g_{1,t+1} \end{aligned} \tag{3.26}$$

This implies that each population obtained by selfing still produces $\frac{5}{6}$ of the proportion of gametes with the Aa haplotype which was produced by the previous generation.

Now the frequency of plants with a heterozygous genotype is considered. This frequency is designated by f_h. Thus

$$f_{h,t} := f_{1,t} + f_{2,t} + f_{3,t}$$

As

$$\begin{aligned} f_{1,t+2} &= \tfrac{1}{2} f_{1,t+1} + \tfrac{2}{9} f_{2,t+1} \\ f_{2,t+2} &= \tfrac{1}{4} f_{1,t+1} + \tfrac{1}{2} f_{2,t+1} + \tfrac{1}{4} f_{3,t+1} \\ f_{3,t+2} &= \tfrac{2}{9} f_{2,t+1} + \tfrac{1}{2} f_{3,t+1} \end{aligned}$$

the decrease of f_h at (continued) selfing is described by:

$$\begin{aligned} f_{h,t+2} &= \tfrac{3}{4} f_{1,t+1} + \tfrac{17}{18} f_{2,t+1} + \tfrac{3}{4} f_{3,t+1} \\ &= f_{h,t+1} - \left(\tfrac{1}{4} f_{1,t+1} + \tfrac{1}{18} f_{2,t+2} + \tfrac{1}{4} f_{3,t+1} \right) \\ &= f_{h,t+1} - \Bigg[\tfrac{1}{4} \left(\tfrac{1}{2} f_{1,t} + \tfrac{2}{9} f_{2,t} \right) + \tfrac{1}{18} \left(\tfrac{1}{4} f_{1,t} + \tfrac{1}{2} f_{2,t} + \tfrac{1}{4} f_{3,t} \right) \\ &\quad + \tfrac{1}{4} \left(\tfrac{2}{9} f_{2,t} + \tfrac{1}{2} f_{3,t} \right) \Bigg] = f_{h,t+1} - \tfrac{5}{36} \left(f_{1,t} + f_{2,t} + f_{3,t} \right) \\ &= f_{h,t+1} - \tfrac{5}{36} f_{h,t} \end{aligned} \tag{3.27}$$

We consider the decrease of the frequency of heterozygous plants for an initial population consisting exclusively of duplex plants. The genotypic composition of S_0 is then (0, 0, 1, 0, 0), with $f_{h,0} = 1$. According to Table 3.5, $f_{h,1}$ amounts

Table 3.6 The frequency in generation t of plants with a heterozygous genotype, *viz.* $f_{h,t}$, in the case of continued self-fertilization in an autotetraploid population, starting with a population exclusively consisting of duplex plants. The parameter λ_s indicates the portion of heterozygous plants which remained

Generation	t	$f_{h,t}$	$\lambda_S = \frac{f_{h,t}}{f_{h,t-1}}$
S_0	0	1	
S_1	1	$\frac{17}{18} = 0.9444$	0.9444
S_2	2	$\frac{29}{36} = 0.8056$	0.8529
S_3	3	$\frac{437}{648} = 0.6744$	0.8372
S_4	4	$\frac{729}{1296} = 0.5625$	0.8341

then to $\frac{2}{9} + \frac{1}{2} + \frac{2}{9} = \frac{17}{18}$. Table 3.6 presents the frequency of plants with a heterozygous genotype in successive generations, as calculated from Equation (3.27).

The frequency of heterozygous plants as a proportion of the frequency in the preceding generation, *i.e.*

$$\lambda_S = \frac{f_{h,t}}{f_{h,t-1}}$$

is also presented in Table 3.6. It appears that λ_S converges to a constant value, *viz.* to $\frac{5}{6} = 0.8333$. This implies, per round of reproduction by selfing, the same constant (relative) decrease in the frequency of heterozygous plants as derived from the frequency of heterozygous gametes; see Equation (3.26).

In this phase, reproduction by means of self-fertilization for n successive generations reduces $f_{h,t}$ to

$$f_{h,t+n} = \left(\frac{5}{6}\right)^n f_{h,t}$$

The frequency of heterozygous plants is halved if $\left(\frac{5}{6}\right)^n = 0.5$, *i.e.* if

$$n = \frac{\ln(0.5)}{\ln(0.8333)} = 3.8$$

Starting with an initial population with genotypic composition (0, 0, 1, 0, 0) the decrease of the frequency of heterozygous plants is even less: in S_4, $f_{h,4}$ is still larger than $\frac{1}{2}$ (Table 3.6).

When comparing the decrease in the frequency of plants with a heterozygous genotype occurring at selfing of a diploid crop and such decrease at selfing of a diploid crop and such decrease at selfing of a tetraploid crop it is clear that the decrease is quite slow in the case of tetraploidy. Continued FS-mating in a diploid crop gives a somewhat faster decrease in the frequency of heterozygous plants than continued selfing of a tetraploid crop.

A more comprehensive treatment of population genetical effects of selfing in an autotetraploid population is given by Seyffert (1959).

3.4 Self-Fertilization and Cross-Fertilization

There are many crops which are neither completely autogamous nor allogamous:

Broad bean	*Vicia faba* L.
Oil-seed rape	*Brassica napus* L.
Lupin	*Lupinus luteus* L.
Sorghum	*Sorghum bicolor* (L.) Moench.
Cotton	*Gossypium hirsutum* L.
Safflower	*Carthamus tinctorius* L.

The genotypic composition resulting from this mixture of modes of reproduction is considered. The portion of the eggs which develops into a zygote after selfing is represented by s and the portion which develops into a zygote after cross-fertilization by $k = 1 - s$.

A general description of the genotypic composition of the plants of generation t is

	Genotype		
	aa	Aa	AA
f	$q^2 + pq\mathscr{F}_t$	$2pq(1 - \mathscr{F}_t)$	$p^2 + pq\mathscr{F}_t$

The portion $s = 1 - k$ of the plants in generation $t + 1$ originates then from selfing. Its genotypic composition is

	Genotype		
	aa	Aa	AA
f	$q^2 + pq\mathscr{F}_t + \frac{1}{2}pq(1 - \mathscr{F}_t)$	$pq(1 - \mathscr{F}_t)$	$p^2 + pq\mathscr{F}_t + \frac{1}{2}pq(1 - \mathscr{F}_t)$

The portion k of the plants in generation $t+1$ originates from random mating. Its genotypic composition is

	Genotype		
	aa	Aa	AA
f	q^2	$2pq$	p^2

Among all offspring the frequency of plants with a heterozygous genotype is then

$$f_{1,t+1} = 2pq(1 - \mathscr{F}_{t+1}) = (1 - k) \cdot pq(1 - \mathscr{F}_t) + k \cdot 2pq$$

implying

$$\begin{aligned} 1 - \mathscr{F}_{t+1} &= \tfrac{1}{2}(1-k)(1-\mathscr{F}_t) + k \\ 2 - 2\mathscr{F}_{t+1} &= 1 - k - \mathscr{F}_t + k\mathscr{F}_t + 2k \\ 2\mathscr{F}_{t+1} &= 1 - k + \mathscr{F}_t - k\mathscr{F}_t = (1-k)(1+\mathscr{F}_t) \\ \mathscr{F}_{t+1} &= \tfrac{1}{2}s(1+\mathscr{F}_t) \end{aligned} \tag{3.28}$$

As required, this expression coincides at $s = 1$ with Equation (3.4).

We now consider the situation that s is constant from one generation to the next. In the case of equilibrium, successive generations have identical genotypic compositions. Then $\mathscr{F}_t = \mathscr{F}_{t+1} = \mathscr{F}_{t+2} = \ldots = \mathscr{F}_e$. Equation (3.28) implies then

$$2\mathscr{F}_e = s(1+\mathscr{F}_e) = s + s\mathscr{F}_e$$

i.e.

$$\mathscr{F}_e(2-s) = s$$

Thus

$$\mathscr{F}_e = \frac{s}{2-s} \tag{3.29}$$

In the equilibrium (e) the genotypic composition is

	Genotype		
	aa	Aa	AA
f	$q^2 + pq\mathscr{F}_e$	$2pq(1-\mathscr{F}_e)$	$p^2 + pq\mathscr{F}_e$

The relation between $\mathscr{F}_e$ and s, *i.e.* Equation (3.29), is almost linear in the range of possible values for s (Fig. 3.2): $\mathscr{F}_e$ roughly equals s.

We now consider, for the case of $p = q = \frac{1}{2}$, the effect on the genotypic composition of a continued change in the mode of reproduction. First the

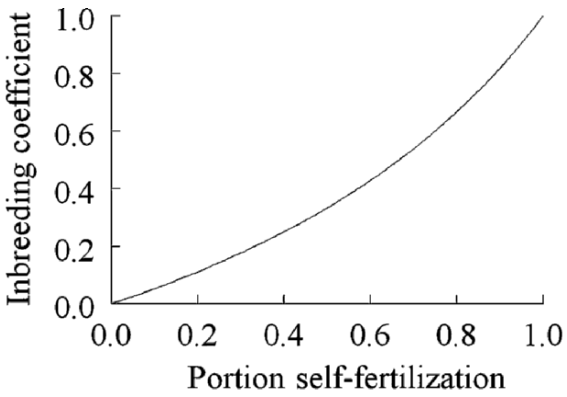


Fig. 3.2 The equilibrium value of the inbreeding coefficient as a function of the portion of reproduction by means of self-fertilization

population genetical effect of some cross-fertilization, *i.e.* $k > 0$, in an – until then - exclusively self-fertilizing crop (*e.g.* wheat) is considered; thereafter we consider the population genetical effect of some selfing, *i.e.* $s > 0$, in an – until then – exclusively cross-fertilizing crop.

Some cross-fertilization in a self-fertilizing crop
Assume that in an F_∞-population, with genotypic composition $(\frac{1}{2},\ 0,\ \frac{1}{2})$, from some generation onward always 10% of the offspring result from cross-fertilization (*i.e.* $k = 0.1$), *e.g.* because the population is maintained in a different environment. In this case the frequency of heterozygous plants increases from $f_1 = 0$ to $f_{1,e} = 0.09$. Some cross-fertilization in a self-fertilizing crop gives thus a non-negligible increase in the frequency of heterozygous plants. According to Equation (3.28) the successive generations will have the following coefficients of inbreeding:

$$\begin{aligned} \mathscr{F}_1 &= 0.900 \\ \mathscr{F}_2 &= 0.855 \\ \mathscr{F}_3 &= 0.835 \\ \mathscr{F}_4 &= 0.826 \\ &\cdot \\ \mathscr{F}_\mathrm{e} &= 0.818 \end{aligned}$$

It is concluded that equilibrium is approached slowly.

Some self-fertilization in a cross-fertilizing crop
We consider a panmictic population with genotypic composition $(\frac{1}{4},\ \frac{1}{2},\ \frac{1}{4})$. From some generation onward always 10% of the offspring is due to selfing (*i.e.* $s = 0.1$). This results in a reduction of the frequency of heterozygous plants: at $s = 0.1$ it reduces from $f_1 = 0.50$ to $f_{1,e} = 0.47$. It can be derived that

$$\begin{aligned} \mathscr{F}_1 &= 0.050 \\ \mathscr{F}_2 &= 0.053 \\ &\cdot \\ \mathscr{F}_\mathrm{e} &= 0.053 \end{aligned}$$

In this situation the equilibrium is attained almost immediately.

Workman and Allard (1962) studied the equilibrium with regard to two segregating loci, attained in the case of simultaneous occurrence of selfing and cross-fertilization, for unlinked loci. Weir and Cockerham (1973) did so for linked loci.

Chapter 4
Assortative Mating and Disassortative Mating

It is reasonable to assume that if two intermating plants resemble each other more, with regard to some trait, than two random plants, then their genotypes for the involved loci will tend to be similar. The population genetic effect of such assortative mating is a decrease of the frequency of plants with a heterozygous genotype. With disassortative mating the intermating plants will tend to resemble each other less than two random plants. The population genetic effect of repeated backcrossing is also considered in this chapter as repeated backcrossing may be considered as a particular application of disassortative mating.

4.1 Introduction

Assortative mating occurs if intermating plants tend to resemble each other more, with regard to some trait, than two random plants. It implies a positive correlation between the mating plants of their phenotypic values for the trait involved. The genotypes of these plants for the loci controlling the expression for the trait will therefore tend, in general, to be similar. With **disassortative mating**, the mating plants will have a negative correlation of their phenotypic values for the considered trait: the mating plants tend to resemble each other less than random plants.

It is obvious that the trait involved in the resemblance should be expressed before pollen distribution. Thus assortative and disassortative mating are only conceivable for traits such as colour of hypocotyls (*e.g.* in radish, *Raphanus sativus* var. *radicula* L.), flower colour (*e.g.* in Brussels sprouts, *Brassica oleracea* L. var. *gemmifera* DC., Example 4.1), anther colour (*e.g.* in maize, *Zea mays* L.), number of tillers (*e.g.* in rye, *Secale cereale* L.), date of flowering (Example 4.2).

Example 4.1 When producing hybrid seed of Brussels sprouts, by making use of sporophytic self-incompatibility, rows of plants representing inbred line A, with genotype S_aS_a, are intermixed with rows of plants representing inbred line B, with genotype S_bS_b. The pure lines involved may differ with regard to shape and size of the ultraviolet-coloured honey guide (which is invisible for the human eye). However, bees, responsible for the pollination, observe such differences. They tend to visit either flowers of the S_aS_a pure line or flowers of the S_bS_b pure line. Thus the bees apply assortative mating, which is counter-productive when the aim is to produce hybrid seed.

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 59–67.

Example 4.2 Assortative mating occurs in cross-fertilizing crops, *e.g.* perennial ryegrass (*Lolium perenne* L.), spontaneously with regard to date of flowering. This phenomenon has attracted a lot of attention in ecological population genetics. The rare, very early flowering plants on the one hand, and the rare, very late flowering plants, on the other hand, are then at a disadvantage. In the case of self-incompatibility, these plants will have a reduced seed-set, due to the scarcity of nearby cross-compatible plants. Such selection against both extreme phenotypes is called **stabilizing selection**.

Plants may produce flowers over an extended period of time. This applies especially to wild plant species, but also to certain cultivated grass species or rye, certainly when grown at a low plant density. The crossing between flowers, or inflorescences, flowering at the same time does then, due to the overlap of flowering periods of different plants, imply rather imperfect assortative mating.

Some authors, *e.g.* Allard (1960, p. 203) and Strickberger (1976, p. 789), have used the term 'phenotypic assortative mating' when considering the present form of assortative mating. They used the term 'genotypic assortative mating' where this book deals with inbreeding. It is questionable whether it is useful to distinguish between two forms of assortative mating: phenotypic resemblance implies at least some genotypic resemblance, especially in the case of qualitative variation. Li (1976) used the terms 'positive' and 'negative assortative mating' instead of assortative and disassortative mating.

The population genetic effect of assortative mating with regard to some trait is a decreased frequency of plants with a heterozygous genotype for the loci affecting the trait, as well as their linked neighbours. Experience shows that for loci controlling traits that have no relationship with fitness (Section 6.1), a decreased frequency of plants with a heterozygous genotype is not associated with inbreeding depression. Inbreeding gives for all loci a decrease in the frequency of plants with a heterozygous genotype and so affects fitness traits and so may result in inbreeding depression. Assortative mating, however, exclusively decreases heterozygosity for loci controlling the expression for the trait involved in the resemblance.

Selection efficiency is promoted by an increased frequency of homozygous genotypes (Section 6.3.2). Assortative mating may thus be a useful tool: in the case of self-incompatibility or dioecy a breeder could apply assortative mating to increase the frequency of homozygous plants, *e.g.* with respect to the locus controlling the colour of the hypocotyl of radish.

With qualitative variation the small number of different phenotypes can easily be distinguished. Thus for the colour of the hypocotyl of radish one may distinguish white and red. The plants can be classified according to the expression for the considered trait. The phenotypes of the plants belonging to the same class are equivalent. Then, with assortative mating, the coefficient of correlation of the phenotypic values of the mating plants will approach the

value 1. The rate of decrease of the frequency of plants heterozygous for the loci involved will then be similar to this rate in the case of self-fertilization.

With quantitative variation the level of expression may behave as a continuous, random variable. This applies to traits such as single plant yield, plant height, or (to a lesser degree) date of flowering or number of tillers. Plants grouped into the same class of phenotypic values have roughly the same phenotype. In this case the coefficient of correlation of the phenotypic values of the mating plants will tend to be less than 1.

It should be clear that the rate of decrease of heterozygosity due to assortative mating strongly depends on the nature of the variation: qualitative or quantitative.

Qualitative variation

In the case of qualitative variation the relation between genotype and phenotype is more direct than in the case of quantitative variation: the classification of plants according to their phenotype tends to reflect the underlying genotypes. The population genetic effect of assortative mating resembles then the population genetic effect of selfing and the frequency of heterozygous plants decreases rather fast.

Quantitative variation

With quantitative variation the relation between genotype and phenotype is disturbed by variation in the quality of the growing conditions: in that situation it is impossible to classify plants on the basis of their phenotype in such a way that all plants in some class have the same genotype, or to distinguish genotypes in such a way that all plants with a specified genotype belong to the same class of phenotypes. In addition, the same phenotype can be produced by a wide range of different genotypes and thus, from both causes, it implies only a loose relationship between phenotype and genotype, which rules out attainment of complete homozygosity by means of continued assortative mating.

For both categories of variation the relation between genotype and phenotype is additionally disturbed by dominance, because different genotypes may then give rise to the same phenotype.

Disassortative mating implies crossing of plants belonging to different phenotypic classes; especially the two extreme classes. It may result in plant material with phenotypes mainly distributed around the mid-parent value.

Maintenance of small populations, *e.g.* accessions in a gene bank, requires care to prevent inconspicuous change of the genotypic composition, due to random variation of the allele frequencies (Chapter 7). Disassortative mating of early flowering plants with late flowering plants may be applied to maintain the typical average flowering time of some accession. In natural populations

plants with extreme phenotypes, *e.g.* very early flowering plants and very late flowering plants, may have a reduced fitness (Example 4.2).

Mating of plants with a different sex may be considered as disassortative mating. In this book some population genetic theory dealing with sex-expression is developed in Chapter 5.

Some authors classify the phenomenon of **incompatibility** among disassortative mating (Karlin, 1968; Crow and Kimura, 1970, p. 166) Two forms of incompatibility may be distinguished: **homomorphic** and **heteromorphic**. In contrast to heteromorphic incompatibility, homomorphic incompatibility is not associated with anatomical differences. In cabbages homomorphic incompatibility is used to produce hybrid varieties (Example 4.1). Heteromorphic incompatibility may occur as **heterostyly**, *e.g.* in primrose (*Primula* sp.). This provision indeed leads to disassortative mating with regard to flower structure (Note 4.1).

Note 4.1 In primrose and buckwheat (*Fagopyrum esculentum* Moench.) heterostyly occurs: there are short-styled plants ('thrum') and long-styled plants ('pin'). Darwin noted that *Primula* spp. plants are pollinated by bees or moths possessing a long proboscis. If an insect collects nectar from a plants producing the *thrum* type of flowers it will pick up pollen around the base of its proboscis. Upon further feeding this pollen may be deposited on the long stigma of plants producing the *pin* type of flowers. If so, the insect may pick up pollen near the tip of its proboscis. This might later be deposited on the short stigma of *thrum* flowers of other plants.

The heterostyly is in fact associated with sporophytic self-incompatibility. Primrose and buckwheat are thus both obligatory allogamous crops.

Often two populations that compensate each other with regard to the expression for one or more traits are crossed. The aim of this initial cross is to introduce from one parent the gene(s) inducing a desired expression for some trait into the other parent, which is an otherwise acceptable genotype (or population). The initial cross is followed by a programme of repeated backcrossing, in which plants with the improved expression are, generation after generation, selected to be crossed with the parent to be improved. Because of the disassortative mating involved in this procedure, repeated backcrossing is treated in this chapter (Section 4.2). In fact disassortative mating is a mode of reproduction that may occur within some populations. Repeated backcrossing could therefore also have been considered in Section 2.2.1, where bulk crossing was introduced.

In some crops **sexual dimorphism** (Chapter 5) occurs. It is possible that each plant can be classified as either a female or as male plant (this situation is called dioecy); or one may distinguish female plants and hermaphroditic plants, which may be monoecious or not.

4.2 Repeated Backcrossing

A breeder may wish to improve an otherwise acceptable genotype by the incorporation of a specific major gene. For example

- It may be desired to improve the resistance of a rice variety or a lettuce variety against a new race of some disease.
- When breeding a hybrid variety it might be useful to develop a **male sterile pure line** which is genotypically identical to the pure line used as the paternal parent of the hybrid, except for its **idiotype** at the locus and cytoplasm controlling pollen development. Then one should transform the male fertile pure line into a male sterile line. This is done by pollination of a male sterile line by the paternal pure line parent. The obtained progeny is repeatedly, *i.e.* generation after generation, backcrossed with the male fertile pure line. (The latter line is called: **maintainer line**. It is, of course, maintained by continued selfing. In Note 3.3 a somewhat different procedure for maintaining a male sterile line was mentioned, *viz.* full sib mating followed by harvesting of the male sterile plants. This procedure is applied with recurrent selection in self-fertilizing crops).

The genotype to be improved is called (for reasons that will become clear hereafter): **recurrent parent**. It may be a pure line (possibly a variety of a self-fertilizing crop or a pure line used in the production of a hybrid variety of a cross-fertilizing crop) or a clone. The allele determining the desired trait is designated by R. It belongs to locus R-r and is to be incorporated into the recurrent parent. The latter is therefore crossed with a **donor** line containing the desired allele, but otherwise resembling the recurrent parent as much as possible. For all loci for which the recurrent parent and the donor line have a different genotype (save locus R-r), one wants to retain the genotype of the recurrent parent. These loci may or may not be linked with locus R-r.

With the introduction of the desired allele R, alleles belonging to other loci – which are possibly linked to locus R-r – are introduced as well. This phenomenon is called **linkage drag**. Many of these unintentionally introduced alleles will be undesirable. Often the breeder is not even aware of the introduction of such undesirable alleles, *e.g.* alleles belonging to loci controlling bitterness of the seeds).

Repeated backcrossing of the material under development with the recurrent parent, is applied in order to replace the dragged alleles step by step with the alleles of the recurrent parent. In this way a so-called **near isogenic line** is developed.

The rate of the replacement is considered for the simple situation of dominance of the desired allele, to be introduced from the donor, over the recurrent parent allele that is to be replaced. Each of all the other loci, for which a possibly unfavourable allele was introduced, is represented by locus B-β. The actual (and favoured) genotype of the variety is represented by BB; the

genotype of the donor by $\beta\beta$. For the time being it is assumed that selection is only applied with regard to the trait controlled by locus R-r. Then it does not matter which allele of locus B-β is dominant, or whether the locus controls a trait that is expressed before or after pollen distribution. The recombination value for loci R-r and B-β is r_c. Its value depends on the specific locus which is represented by B-β. For most loci r_c will amount to $\frac{1}{2}$. The slower the (rate of) replacement of allele β by allele B, the higher the number of backcross generations required to restore genotype BB for all loci represented by B-β.

Allele R is introduced by crossing the recurrent parent (say $\mathbf{P}_1$, with genotype rB/rB) with a donor (say P_2, with genotype $R\beta/R\beta$). The obtained F_1 has genotype $rB/R\beta$. The haplotypic composition of the gametes produced by F_1 is

	Haplotype			
	rB	$r\beta$	RB	$R\beta$
f	$\frac{1}{2}(1-r_c)$	$\frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2}(1-r_c)$

The first backcross, $P_1 \times F_1$, results in a population (usually designated as BC_1) with genotypic composition:

	Genotype			
	rB/rB	$r\beta/rB$	RB/rB	$R\beta/rB$
f	$\frac{1}{2}(1-r_c)$	$\frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2}(1-r_c)$

Elimination of plants with genotype rr transforms population BC_1 into population BC_1'. The genotypic composition of BC_1' and the haplotypic composition of the gametes produced by each genotype in BC_1' are

Genotypic composition of BC_1		Haplotypic composition of the gametes produced by each genotype			
genotype	f	rB	$r\beta$	RB	$R\beta$
$RB/r\beta$	r_c	$\frac{1}{2}$	0	$\frac{1}{2}$	0
$R\beta/rB$	$1-r_c$	$\frac{1}{2}(1-r_c)$	$\frac{1}{2}r_c$	$\frac{1}{2}r_c$	$\frac{1}{2}(1-r_c)$

The haplotypic composition of the gametes produced by BC_1' as a whole is

	Haplotype			
	rB	$r\beta$	RB	$R\beta$
f	$\frac{1}{2}r_c + \frac{1}{2}(1-r_c)^2$	$\frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}r_c + \frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}(1-r_c)^2$

The second backcross, *i.e.* $P_1 \times BC_1$, yields population BC_2 with genotypic composition:

	Genotype			
	rB/rB	$r\beta/rB$	RB/rB	$R\beta/rB$
f	$\frac{1}{2}r_c + \frac{1}{2}(1-r_c)^2$	$\frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}r_c + \frac{1}{2}r_c(1-r_c)$	$\frac{1}{2}(1-r_c)^2$

Because all BC_1'-plants have genotype Rr, half of the BC_2-plants will have genotype rr. Elimination of the latter plants yields population BC_2' with genotypic composition:

	Genotype	
	RB/rB	$R\beta/rB$
f	$1-(1-r_c)^2$	$(1-r_c)^2$

Likewise, population BC_t' contains genotype $R\beta/rB$ with frequency $(1-r_c)^t$. The frequency of plants with genotype $R\beta/rB$ in population $BC_t{}'$ is thus $(1-r_c)^t$. For $r_c = \frac{1}{2}$ this amounts to $(\frac{1}{2})^t$. The frequency of genotype RB/rB amounts then to $1-(\frac{1}{2})^t$. The probability that a line, obtained by selfing in population $BC_t{}'$ a random plant, might segregate for locus $B-\beta$ is $(1-r_c)^t$.

We consider now the K unlinked loci $B_1-\beta_1, B_2-\beta_2, \ldots, B_K-\beta_K$. Locus R-r is not linked with any of these. Then in population BC_t' the frequency of plants with the desired complex genotype will amount to

$$\left[1-\left(\tfrac{1}{2}\right)^t\right]^K = \left[\frac{2^t-1}{2^t}\right]^K \tag{4.1}$$

This expression is equal to Expression (3.23), tabulated in Table 3.4 for $K = 1, \ldots, 14$ and $t = 1, \ldots, 7$. When considering $K = 7$ loci Table 3.4 shows that in population $BC_5{}'$ the frequency of plants with the complex genotype $RrB_1B_1B_2B_2 \ldots B_7B_7$ amounts to 0.801. In population $BC_6{}'$ it is already 0.896. When considering $K = 14$ loci the frequency of plants with genotype $RrB_1B_1 \ldots B_{14}B_{14}$ amounts to 0.641 in population $BC_5{}'$ and to 0.802 in population $BC_6{}'$.

The frequency of plants with a complex genotype deviating for one or more of the loci B_1-$\beta_1, \ldots, B_K$-β_K from the genotype of the recurrent parent will amount to:

$$1-\left[\frac{2^t-1}{2^t}\right]^K$$

This equation gives the probability that a line, obtained by selfing a random plant taken from population $BC_t{}'$, might segregate for one or more of the K loci. Such segregation will also appear from a difference, for at least one trait, between plants of the line and the recurrent parent.

It may be concluded that, even for unlinked loci, five generations of backcrossing yield an insufficient reduction in the frequency of plants containing

at one or more loci an undesired allele. One or more additional backcross generations already implies a considerable reduction, especially for 'large' values for K. One should, of course, minimize K. This can be done by using as the donor a genotype that resembles the recurrent parent as closely as possible.

An additional criterion for choosing a donor, follows from the dominance relationships among the alleles at the B-β loci. With regard to loci for which the recurrent parent allele B is not dominant over the donor allele β, one might distinguish, among the plants with genotype Rr, plants with genotype $RrBB$ and plants with genotype $RrB\beta$. Selection of plants with genotype $RrBB$ implies then elimination of allele β. Selection, particularly marker-assisted selection (Section 12.3.2), among the plants with genotype Rr, of plants with the genotype of the recurrent parent (BB) reduces consequently the number of backcross generations required to attain the desired frequency of plants with genotype $RrBB$. Markers strongly linked to locus B-β and/or locus R-r are particularly useful. Among donor lines which differ from the recurrent parent with regard to their genotype for K loci, one should choose the donor containing a dominant allele at the highest number of these loci. Different donor lines can, in this respect, be compared by considering the similarity of the F_1 and the donor: the greater the similarity, the larger the number of dominant donor alleles.

Until now the recurrent parent was assumed to have a homozygous genotype. When dealing with vegetatively propagated crops (such as apple, rhubarb, shallots, asparagus) the recurrent parent may be heterozygous for some locus B-b-β. The cross between the recurrent parent (with genotype Bb) and a donor (with genotype $\beta\beta$) yields an F_1 with the following genotypic composition

	Genotype	
	$B\beta$	$b\beta$
f	$\frac{1}{2}$	$\frac{1}{2}$

The frequencies of genotypes and alleles in $BC_1{}'$, $BC_2{}'$ and $BC_3{}'$ then amount to:

		Genotype					Allele		
		bb	Bb	BB	$b\beta$	$B\beta$	b	B	β
f	in $BC_1{}'$:	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}$
	in $BC_2{}'$:	$\frac{3}{16}$	$\frac{3}{8}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{7}{16}$	$\frac{7}{16}$	$\frac{1}{8}$
	in $BC_3{}'$:	$\frac{7}{32}$	$\frac{7}{16}$	$\frac{7}{32}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{15}{32}$	$\frac{15}{32}$	$\frac{1}{16}$

It will be clear that repeated backcrossing to a heterozygous recurrent parent is expected to result in a $BC_\infty{}'$ population with genotypic composition;

	Genotype		
	bb	Bb	BB
f	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

with regard to locus B-b-β. $BC_\infty{}'$ is thus not identical to the recurrent parent, but to its S_1 lines. The same applies to the two loci B_1-b_1-β_1 and B_2-b_2-β_2, which may be linked or not, if the genotype of the recurrent parent is $B_1b_1B_2b_2$.

Bos (1980) considered backcrossing in autotetraploid crops. In population $BC_t{}'$ the frequency of plants containing the unintentionally introduced allele β was derived to be $(\frac{1}{2})^{t-1}$ if loci R-r and B-β are unlinked. Thus, compared with diploid crops, one additional backcross generation is required in order to obtain the same degree of replacement of β by B.

Chapter 5
Population Genetic Effect of Selection with regard to Sex Expression

Breeders may consider the use of male sterility when developing hybrid varieties or when making complex bulk crosses. The frequency of male sterile plants is then an interesting topic, especially when the involved crop is grown because of seed yield. Male sterile plants may have a reduced seed-set and consequently a reduced fitness as compared to male fertile plants. Selection with regard to sex expression is therefore an issue of practical relevance.

5.1 Introduction

The types of sex expression distinguished for our purposes are

- Hermaphroditism, in contrast to
- Sex differentiation (sexual dimorphism)

Hermaphroditism is the most common form of sex expression among plant species. It means that the reproductive organs of both sexes are present in the same flower, *i.e.* a **bisexual** flower (this situation is indicated by the symbol ☿), or in different flowers occurring on the same plant. In the latter case a flower contains either male or female organs; this situation is called **monoecy**, indicated by the symbol ♂♀. Monoecy occurs in crops such as

Maize	*Zea mays* L.
Castorbean	*Ricinus communis* L.
Cucumber	*Cucumis sativus* L.
Plane trees	*Platanus occidentalis* L.
Alder	*Alnus glutinosa* Gartn.
Hazelnut	*Corylus avellana* L.

The types of sex differentiation to be distinguished are

- Dioecy
- Gynodioecy

Dioecy means that plants either exclusively produce female flowers (these are female plants, indicated by ♀), or exclusively male flowers (these are the male plants, indicated by ♂).

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 69–76.

Well-known dioecious crops are

Spinach	*Spinacia oleracea* L.
Asparagus	*Asparagus officinalis* L.
Hemp	*Cannabis sativa* L.
Hops	*Humulus lupulus* L.
Poplar	*Populus nigra* L.
Date	*Phoenix dactylifera* L.
Kiwi	*Actinidia deliciosa* (A. Chev.) [C.F. Liang & A.R. Ferguson]
Papaya	*Carica papaya* L.

Gynodioecy means that female plants as well as hermaphroditic plants occur. Thus a gynodioecious maize population consists of male sterile plants, *i.e.* female plants, as well as 'normal' plants. This situation is considered in Section 5.2.

It has been demonstrated that sex expression, both in plants and animals, is due to rather diverse mechanisms, ranging from a more or less clear-cut XY-XX-mechanism to sex expression determined by environmental conditions (Example 5.1).

Example 5.1 In cucumber four types of sex expression may occur: monoecy, gynoecy, and andromonoecy (plants have male and hermaphroditic flowers) and hermaphroditism. Modern cucumber cultivars produce exclusively female flowers: their fruits develop parthenocarpic. The sex expression is affected by treatment with gibberellic acid or silvernitrate. These substances promote the development of male flowers. This allows the selfing required for maintenance of pure lines used in hybrid varieties.

The population genetic effect of selection with regard to sex expression is thus necessarily derived on the basis of simplifying assumptions about the genetic control of sex expression. In this chapter implications of specific assumptions about the genetic control of dioecy or gynodioecy are elaborated.

Assumed genetic control of dioecy

A 'homozygous' genotype is assumed to give rise to a female plant, *viz.* XX in the case of sex chromosomes or mm in the case of a locus M-m controlling sex expression. A 'heterozygous' genotype (XY or Mm) is assumed to give rise to a male plant:

	Genotype	
	mm (or: XX)	Mm (or: XY)
sex	♀	♂
f	$\frac{1}{2}$	$\frac{1}{2}$

The genotypic composition $(\frac{1}{2}, \frac{1}{2}, 0)$ results from the harvesting of female plants which have been pollinated by male plants. This genotypic composition will apply whatever the initial frequencies of male and female plants.

Assumed genetic control of gynodioecy

Gynodioecy occurs in the situation of **cytoplasmic male sterility** or in the situation of **genic male sterility**. The **idiotypic** basis for cytoplasmic male sterility is assumed to be

	Idiotype		
	(S)rr	(·)Rr	(·)RR
sex	♀	☿ or ♂	☿ or ♂

The symbol (S) designates presence of male-sterility-inducing cytoplasm, the symbol (·) presence of any cytoplasm. The latter symbol represents thus both (S) and (N), *i.e.* the presence of normal cytoplasm. Locus *R*-*r* is the male fertility restoring locus.

The genetic basis for genic male sterility is assumed to be

	Genotype		
	mm	Mm	MM
sex	♀	☿ or ♂	☿ or ♂

In the case of gynodioecy there is selection against the male-sterility-inducing allele (this is allele m; or – in the presence of (S) cytoplasm – allele r). Male sterile plants are unable to transmit this allele to the next generation via pollen. The decrease in the frequency of male sterile plants is considered in Section 5.2.

5.2 The Frequency of Male Sterile Plants

Allogamous crops

In cross-fertilizing crops male sterile plants may have a normal (complete) seed set. The selection against the male-sterility-inducing allele, say m, is then due to the incapability of plants with genotype mm to transmit allele m via the pollen to the next generation. Only plants with genotype MM or Mm produce pollen. Eggs are produced by all plants, whatever the genotype. The frequency of male sterile plants in this situation is considered in Section 5.2.1.

Elimination of male sterility may be a breeding objective because of a low seed-set on the male sterile. Male sterile plants, which may be conspicuous because of their low seed-set, are then not harvested. This implies that plants with genotype mm not only fail to produce pollen, but – effectively – then

also fail to produce eggs. Only male fertile plants are harvested. In successive generations the genotypic composition with regard to locus M-m coincides then with the genotypic composition with regard to locus A-a in the case of continued mass selection, before pollen distribution, against plants with genotype aa. The decrease in the frequency of gene m proceeds, therefore, as in Example 6.11.

Autogamous crops

Incomplete seed-set is certainly to be expected for male sterile plants belonging to a self-fertilizing crop. In Section 5.2.2 attention is given to natural selection against male sterility in an autogamous crop.

In the case of recurrent selection in a self-fertilizing crop (Note 3.3), only male sterile plants are harvested. This guarantees that the harvested seeds resulted from intercrossing. Then, effectively, plants with genotype MM or Mm produce the pollen and plants with genotype mm the eggs. This situation coincides effectively with dioecy. It leads immediately to the equilibrium frequencies $(\frac{1}{2}, \frac{1}{2}, 0)$, whatever the seed-set of male sterile plants may be.

5.2.1 Complete seed-set of the male sterile plants

The situation of complete seed-set of male sterile plants of a cross-fertilizing crop resembles the case of mass selection, after pollen distribution, against plants with genotype aa: such plants are not harvested and, consequently, do not transmit allele a via eggs; pollen, however, is produced by all plants, whatever the genotype. In successive generations the genotypic composition with regard to locus M-m is, consequently, equal to the genotypic composition with regard to locus A-a in the case of mass selection, after pollen distribution, against plants with genotype aa. This is illustrated in Example 6.12.

Consider now a gynodioecious population of a cross-fertilizing crop, *e.g.* maize: female plants have idiotype (S)rr and hermaphroditic plants idiotype (N)rr. The relative frequencies of female plants and hermaphroditic plants will then not change if these two categories of plants have equal seed-set. The problem described in Note 5.1 pertains to this situation.

Note 5.1 In a gynodioecious population of a cross-fertilizing crop the female plants are assumed to have idiotype (S)rr and the hermaphroditic plants idiotype (N)rr. Derive, for this situation, how the idiotypic composition with regard to some locus A-a is expected to develop if the initial frequencies of (N)aa and (S)AA are both $\frac{1}{2}$.

5.2.2 *Incomplete seed-set of the male sterile plants*

In the case of cytoplasmic male sterility in a self-fertilizing crop the incomplete seed-set of the male sterile plants, due to insufficient pollination, implies reduction of the frequency of plants with the (S) cytoplasm. With **cleistogamy**, *i.e.* the flowers remain closed at pollination time, there is no seed-set at all. Plants with the (S) cytoplasm do then not produce any offspring. The (S) cytoplasm will then not be transmitted to the next generation. It is immediately lost.

In the remainder of this section attention is given to genic male sterility in a self-fertilizing crop. It is assumed that all seeds produced by hermaphroditic plants, *i.e.* by plants with genotype Mm or MM, are due to self-fertilization. For these plants the value for k, *i.e.* the portion of the eggs that develop into a zygote after cross-fertilization (Section 3.5) is zero. The seeds produced by male sterile plants, *i.e.* plants with genotype mm, are due to cross-fertilization. It is rather common that male sterile plants produce flowers that are more widely opened than flowers produced by male fertile plants, but nevertheless they tend to produce less seeds than male fertile plants. The relative seed-set or – in more general population genetic terms – the relative fitness of plants with genotype mm is represented by the factor w_0. (The relative fitness is also designated by $1 - s_0$, or briefly by $1 - s$, where s represents the so-called selection coefficient for plants with genotype mm; see also Section 6.1.) Example 5.2 gives an example.

Example 5.2 Even for a crop like spring barley, k appears to be positive. Jain and Allard (1960) observed $k = 0.02$ for hermaphroditic barley plants. The seed-set of male sterile barley plants is rather variable. For the conditions in Davis, California, Jain and Suneson (1964) reported a maximum seed-set of 0.40; *i.e.* $s \geq .6$. For Wageningen, The Netherlands, Baltjes (1975) reported a maximum seed-set of 0.20; *i.e.* $s \geq 0.8$.

Different parental genotypes produce different numbers of offspring. The effective (relative) frequencies (f_e) of parental genotypes are calculated from their actual frequencies in the following way:

	Genotype		
	mm	Mm	MM
f	$f_{0,t}$	$f_{1,t}$	$f_{2,t}$
w	$1 - s$	1	1
f_e	$\frac{(1-s)f_{0,t}}{1-sf_{0,t}}$	$\frac{f_{1,t}}{1-sf_{0,t}}$	$\frac{f_{2,t}}{1-sf_{0,t}}$

Plants with genotype Mm or MM are assumed to produce offspring by spontaneous self-fertilization:

- The genotypic composition of the offspring of plants with genotype Mm is $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$.
- The genotypic composition of the offspring of plants with genotype MM is $(0, 0, 1)$.

Plants with genotype mm produce offspring by cross-fertilization. The haplotypic composition of the pollen produced by generation t is

	Haplotype	
	m	M
f	$g_{0,t+1}$	$g_{1,t+1}$

where

$$g_{0,t+1} = \frac{\frac{1}{2}f_{1,t}}{1 - f_{0,t}} \text{ and } g_{1,t+1} = \frac{\frac{1}{2}f_{1,t} + f_{2,t}}{1 - f_{0,t}}$$

The genotypic composition of the offspring of plants with genotype mm is $(g_{0,t+1}, g_{1,t+1}, 0)$. Altogether the genotypic composition of generation $t+1$, in terms of the genotype frequencies in generation t is

	Genotype		
	mm	Mm	MM
f	$\frac{\frac{\frac{1}{2}f_{1,t}(1-s)f_{0,t}}{1-f_{0,t}} + \frac{1}{4}f_{1,t}}{1-sf_{0,t}}$	$\frac{\frac{(\frac{1}{2}f_{1,t}+f_{2,t})(1-s)f_{0,t}}{1-f_{0,t}} + \frac{1}{2}f_{1,t}}{1-sf_{0,t}}$	$\frac{\frac{1}{2}f_{1,t}+f_{2,t}}{1-sf_{0,t}}$

(5.1)

The frequency of plants with genotype Mm decreases due to self-fertilization but, on the other hand, it increases due to cross-fertilization of plants with genotype mm. The frequency of plants with genotype MM can only increase. The eventual genotypic composition is thus (0, 0, 1). This limit is approached more quickly when the seed-set of plants with genotype mm is lower, *i.e.* s is larger. Example 5.3 illustrates the reduction of f_0 for a few values for s.

Example 5.3 Table 5.1 presents f_0, *i.e.* the frequency plants with genotype mm. It does so for several values of s and for successive generations, starting with an initial population with the genotypic composition of an F_2, *i.e.* $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. The column headed by '$s = 0$' represents complete seed-set of male sterile plants. The column headed by '$s = 1$', representing complete sterility, illustrates how f_0 is reduced by mass selection in a self-fertilizing crop against plants with genotype mm. The column headed 'Observed frequency' presents actual data obtained from barley, Composite Cross XXI (Example 5.4). The frequencies presented in this column and in the column headed '$s = 0.8$' are depicted in Fig. 5.1. It appears that f_0 decreased in later generations less than calculated for $s = 0.8$: from population F_8 onward the actual values for f_0 were somewhat higher than the calculated values. Some tentative explanations for this are given at the end of the present section.

Suneson (1956) advocated the so-called *evolutionary plant breeding method*. It is based on the thought that natural selection in a genetically heterogeneous population favours, for certain traits, the same phenotypes as preferred by the breeder. The improvement of the population will be slow, but in the long run sufficient for obtaining attractive plant material. Example 5.4 provides some results.

Table 5.1 The (expected) frequency of male sterile plants (with genotype mm) in successive generations. The genotypic composition of the initial population is $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. The relative fitness of the male sterile plants is $1-s$. The column headed by 'Observed frequency' presents actual data obtained from barley (Baltjes, 1975)

	Frequency of male sterile plants expected for				Observed
Population	$s = 0$	$s = 0.6$	$s = 0.8$	$s = 1$	frequency
F_2	0.250	0.250	0.250	0.250	
F_3	0.208	0.186	0.177	0.167	
F_4	0.159	0.124	0.122	0.100	0.060
F_5	0.125	0.082	0.069	0.056	
F_6	0.098	0.054	0.042	0.029	0.037
F_7	0.078	0.035	0.025	0.015	0.023
F_8	0.062	0.023	0.015	0.008	0.020
F_9		0.016	0.009		0.010
F_{10}		0.010	0.005		0.013
F_{11}		0.003			
F_{12}		0.002			0.010
F_{13}		0.001			0.006

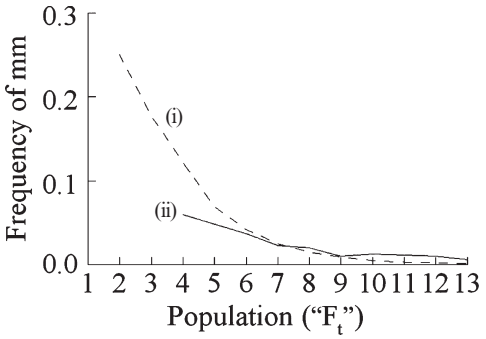


Fig. 5.1 The frequency of male sterile plants, with genotype mm, in successive generations. The genotypic composition of the original population was $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. (i) Data calculated for a relative fitness of the male sterile plants equal to $1 - s = 0.2$, and (ii) observed data in barley (Baltjes, 1975)

Example 5.4 To test the '*evolutionary plant breeding method*' hypothesis, Suneson developed broad base populations by open pollination of male sterile lines. He developed Composite Cross XXI by growing 6200 spring barley varieties next to male sterile barley plants. The seed harvested from the male sterile plants was used as the source population. This population was grown for many years/generations. Baltjes (1975) studied, within the same growing season, many generations, as derived in Wageningen, The Netherlands. A significant improvement in resistance to powdery mildew appeared. As for yield, however, no clear effect was observed: relative to the check variety Zephyr, the F_4 population yielded 75.7% and the F_{13} population 83.7%.

Baltjes (1975) observed that f_0 decreased in later generations less than calculated for $s = 0.8$: from F_8 onward the actual frequency of plants with genotype mm was somewhat higher than the calculated frequency. Two tentative explanations are presented:

1. The relative fitness of male sterile plants may increase in the course of the generations. Thus seed-set improves. This could be due to more intense pollination because of the increase in the frequency of male fertile plants. Indeed, Jain and Suneson (1964) reported a seed-set of 40% in generation F_{18} and a seed-set of 60% in generation F_{21}. They, therefore, assumed a higher relative fitness of male sterile plants at a lower frequency of such plants: $1 - s$ was taken to be $0.6 - f_0$.
2. Male sterile plants (genotype mm) produce offspring heterozygous for many loci. Due to this highly heterozygous background-genotype these offspring (genotype mm or Mm), may tend to be more vigorous than the more homozygous plants (genotype mm, Mm or MM) obtained after selfing. Constancy of q, the frequency of gene m, may occur if its potential decrease, because of reduced fertility of mm plants, is offset by its potential increase, due to greater vitality of mm plants belonging to the heterozygous offspring of plants with genotype mm (Jain and Suneson, 1964).

Chapter 6
Selection with Regard to a Trait with Qualitative Variation

Plant breeding aims at the genetic improvement of plant material. Thus among candidates for selection (clones, (pure) lines, hybrids, families or individual plants) those resembling most closely the ideal of the breeder are selected. The genetic improvement due to selection often deviates from the ultimate goal. One of the causes is that natural selection interferes with the artificial selection. Thus the phenotype(s) favoured by the breeder (under artificial selection) may differ from the phenotype(s) best prepared for 'the struggle for life' (under natural selection). Another cause for a disappointing result from artificial selection is the fact that the phenotype of a candidate is a poor indicator of the quality of its genotype. The phenotype may give a misleading impression of the genotype because of dominance, of epistasis or because of the growing conditions.

This chapter considers impacts of artificial selection on the genotypic composition with regard to traits with qualitative variation. Some attention is given to effects of natural selection. Selection with regard to traits with quantitative variation is considered in later chapters.

6.1 Introduction

The genotypic composition of a population may change from one generation to the next because of

- The mode of reproduction
  This cause for a change in the genotypic composition was considered in Chapters 2, 3 and 4. The change is not associated with changes of the allele frequencies.
- Selection
  This cause was briefly considered in the previous chapter. It will be thoroughly further elaborated in the present chapter, as well as in later chapters. The change is associated with changes of the allele frequencies.
- Random variation of allele frequencies
  This cause is due to a small population size. It is elaborated in Chapter 7.

In Chapter 1 it was indicated that all traits can show qualitative variation as well as quantitative variation. Nevertheless, the effect of selection will be considered separately for these two types of variation. Thus in the present chapter impacts of selection on the genotypic composition for traits exhibiting exclusively qualitative variation are considered.

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 77–106.

In practice, selection often aims at improvement of traits with quantitative variation. Then one may apply within lines or families, that are acceptable for the considered trait, additional single-plant selection for that trait (this is called: **combined selection**, see Section 14.3.1). Alternatively, one may select with regard to an additional trait among the acceptable lines or families (this is called: **simultaneous selection**, see Section 12.1). The efficiency of selection for traits with quantitative variation is often (very) low. For such selection special procedures may be considered which are dealt with separately, especially from Chapter 12 onward.

In Chapters 2 and 3 the development, in the course of the generations, of the genotypic composition of a population was derived on the basis of the implicit assumption that different genotypes possess the same vitality and the same fertility. In the present chapter this assumption is dropped: genotypes are assumed to differ with regard to their vitality and/or fertility. This is done with the intention of allowing models more accurately describe the development of the genotypic composition. A drawback is that such models will apply in a narrower range of situations, as different selection strategies, *i.e.* different patterns of genetic variation in vitality and fertility, require different models.

Selection occurs if genotypes of the zygotes differ with regard to **fitness**, *i.e.* the expected number of (viable) seeds to be produced in the adult plant stage of these genotypes. The expected number of seeds is, of course, the product of the probability that a zygote with the considered genotype develops into an adult, reproducing a plant and the average number of seeds produced by such a plant. The probability that a zygote with a certain genotype survives until the adult plant stage is the so-called **vitality** (v) component of the fitness (W) of this genotype. It depends on the success of germination, the competitive ability as a seedling, the growth rate, *etc.* The average number of seeds produced by an adult plant with the considered genotype is the so-called **fertility** (ϕ) component of the fitness of this genotype. This number depends on the number of ovules, the number of pollen grains, the efficiency of fertilization, *etc.* Variation among genotypes with regard to fitness implies selection.

To derive the impact of selection on the genotypic composition we consider the fitnesses (W) of the genotypes for some locus A-a. This locus may, for example, control the taste of fruits or seeds (sweet or bitter). The fitnesses of these genotypes are considered for the situation where genotypes aa, Aa and AA have the same background genotypes, which do not interact differentially with the genotypes for locus A-a. As in Section 2.2.1 the suffix j of the fitness parameter W_j indicates the number of A alleles in the involved genotype. Example 6.1-a shows how differences between genotypes with regard to vitality and fertility affect the genotypic composition.

The fitnesses of genotypes aa and AA are often related to the fitness of genotype Aa. This yields **relative fitness**, say w_j, where $w_1 = 1$. Instead of w_j one may write $1 - s_j$, where s_j is the so-called **selection coefficient**.

Example 6.1-a An imaginary example of natural selection with regard to a trait with qualitative variation is elaborated for the F_2 and F_3 generations of a self-fertilizing species. The initial cross involved genotypes aa and AA. All plants of population F_1 have genotype Aa and have, therefore, the same fitness. The vitalities of zygotes with genotype aa, Aa and AA are assumed to be $\frac{1}{2}$, 1 and $\frac{1}{2}$, respectively. The fertilities of adult plants with these genotypes are arbitrarily assumed to be 32, 48 and 24, respectively. The fitnesses of the three genotypes are thus 16, 48 and 12. The genotypic compositions, expressed in absolute numbers of plants (#), in successive phases are

	Genotype		
	aa	Aa	AA
F_1: # zygotes	–	1	–
# reproducing plants	–	1	–
# seeds per plant	–	48	–
F_2: # zygotes	12	24	12
# reproducing plants	6	24	6
# seeds per plant	32	48	24
F_3: # zygotes	$6\times 32+\frac{1}{4}(24\times 48)$ $= 480$	$\frac{1}{2}(24\times 48)$ $= 576$	$6\times 24+\frac{1}{4}(24\times 48)$ $= 432$
f: zygotes	0.3226	0.3871	0.2903

The zygotic frequency of allele A in F_2 is 0.5. In F_3 it is $\frac{1}{2}(0.3871)+0.2903 = 0.4839$. The frequency of allele A is thus a little bit reduced due to natural selection: genotype AA has a smaller fitness than genotype aa.

In the absence of selection the genotypic composition of F_3 would have been (0.375, 0.250, 0.375). Due to the high fitness of plants with genotype Aa, the reduction of the frequency of plants with genotype Aa due to selfing is considerably diminished.

With regard to the fitness-affecting locus A-a the considered population in its initial state, prior to the selection, is described by

	Genotype		
	aa	Aa	AA
f	f_0	f_1	f_2
W	W_0	W_1	W_2
w	$w_0 = \frac{W_0}{W_1} = 1 - s_0$	1	$w_2 = \frac{W_2}{W_0} = 1 - s_2$

Example 6.1-b gives a numerical illustration.

Example 6.1-b The 12 F_2 zygotes with genotype aa, see Example 6.1-a, contributed eventually $6 \times 32 = 192$ seeds to the F_3. The expected number of seeds eventually to be produced by a zygote with genotype aa is thus 16. Equally, the fitness of a zygote with genotype Aa amounts to $\frac{24\times 48}{24} = 48$; of a zygote with genotype AA it is $\frac{6\times 24}{12} = 12$. The relative fitnesses of zygotes with genotype aa, Aa or AA are $\frac{1}{3}$, 1 and $\frac{1}{4}$, respectively, implying that $s_0 = \frac{2}{3}$ and $s_2 = \frac{3}{4}$.

The expected relative fitness of a zygote can easily be derived from the above scheme:

$$\mathrm{E}\underline{w} = f_0 w_0 + f_1 w_1 + f_2 w_2 \tag{6.1}$$

For a specific zygote, the product of its zygotic frequency and its fitness measures the **effective genotype frequency**, f_e. To induce the sum of these effective genotype frequencies to be equal to 1, one should calculate $f_{e,j}$ as:

$$f_{e,j} = \frac{w_j f_j}{\mathrm{E}\underline{w}} \tag{6.2}$$

Example 6.1 is expressed in absolute numbers of plants. Example 6.2 presents the same data in terms of (relative) effective genotype frequencies.

Example 6.2 The expected relative fitness of an F_2-zygote is $\mathrm{E}\underline{w} = \frac{1}{3} \times \frac{1}{4} + 1 \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{4} = 0.6458$. It is used to calculate, according to Equation (6.2), the effective genotype frequencies in F_2. The zygotic genotype frequencies in F_3 are derived from the effective genotype frequencies in F_2 as for normal self-fertilization. This proceeds as follows

			Genotype		
			aa	*Aa*	*AA*
		w	$\frac{1}{3}$	1	$\frac{1}{4}$
F_2:	zygotes:	f	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
		f_e	0.1290	0.7742	0.0968
F_3:	zygotes:	f	0.3226	0.3871	0.2903

The resulting figures are equal to those derived in Example 6.1-a on the basis of absolute numbers of plants.

In the case of artificial selection certain genotypes do not produce offspring at all, whereas other genotypes produce the 'normal' number of offspring. Such selection is said to be **complete**. With natural selection certain genotypes produce systematically more offspring than others. Such selection is said to be **incomplete** (Example 6.3).

Example 6.3 Locus *A*-*a* controls the taste of fruits. Plants with genotype *aa* produce sweet fruits, whereas plants with genotype *Aa* or *AA* produce bitter fruits. The relative fitnesses (w) of the genotypes, in the case of natural selection as well as in the case of artificial selection, could consequently be

		Genotype		
		aa	*Aa*	*AA*
w:	With natural selection:	$\frac{1}{2}$	1	1
	With artificial selection:	1	0	0

In self-fertilizing crops the number of offspring of a plant can be determined unambiguously. For cross-fertilizing crops, however, it is virtually impossible

to control and/or to count the number of offspring of a plant via its pollen. It is much easier to determine the number of offspring of a plant via its eggs. Therefore in the following, attention is primarily given to the number of offspring of a plant via its eggs. The term **complete selection**, as mentioned above, applies to this situation. Thus the expected number of seeds produced by a genotype, *i.e.* offspring via the female gametes, is taken to be decisive for the fitness of the genotype.

For traits with quantitative variation the actual selection will generally fail to be complete. Thus when it is aimed to select plants with genotype Aa or AA, due to the growing conditions, several (or many) of the selected plants will have genotype aa. For traits with qualitative variation, however, the ideal of complete selection may be closely approached (Example 6.4).

Example 6.4 In order to select plants with a genotype yielding resistance to some disease one may inoculate seedlings representing a segregating population with the pathogen. The susceptible plants (possibly with genotype rr) are eliminated and the resistant plants (possibly with genotype Rr or RR) survive.

A somewhat hidden form of natural selection concerns selection among haplotypes (in the gametophytic phase). An extreme form of such selection is gametophytic self-incompatibility. In this case the fitness to be associated with some haplotype, specified by its S-allele, depends on the frequency of the considered allele. (This is an example of frequency-dependent fitness selection, see Section 6.2.) Another example of gametophytic selection is **certation**, *i.e.* different haplotypes have different pollen tube growth rates (Example 6.5).

Example 6.5 For maize plants with genotype $Rf_1rf_1Rf_2rf_2$ it has been observed that pollen grains containing two male-fertility-restoring alleles in their haplotype, *i.e.* pollen grains with haplotype Rf_1Rf_2, were more likely to fertilize an egg than pollen grains containing only one male-fertility-restoring allele (with haplotype rf_1Rf_2) (Josephson, 1962).

Apart from incompatibility systems, gametophytic selection is a rare phenomenon. This is no surprise because such selection eliminates alleles, endowing the pollen with a low vitality. Thus in this book it is assumed that gametophytic selection does not produce disturbing effects and hence will be ignored.

Selection implies that different genotypes differ (systematically) in fitness. Indeed, Lerner (1958, p. 5) spoke about 'non-random differential reproduction of genotypes'. It results in a change in allele frequencies. Selection within a single pure line or within a single clone is useless as a breeding procedure, because it will not yield a change in allele frequencies. For sanitary reasons such selection may, however, be very useful: elimination of virus-infected plants

from a seed potato field contributes greatly to the performance of the crop grown from the seed potatoes.

The goal of artificial selection, *i.e.* the production of a cultivar better adapted to demands of growers or consumers, has seldom coincided with the goal of natural selection, *i.e.* improvement of fitness (Example 6.6).

Example 6.6 In the breeding of lettuce or cabbage, artificial selection aims at a well-developed head, whereas natural selection may aim at an undisturbed development of the inflorescence. Similarly, artificial selection favours short culms in wheat or rice, whereas natural selection may favour long culms endowing a high competitive ability. Seed shattering is advantageous under natural conditions, but in a cultivar it is an undesired trait. The goals of artificial selection and natural selection may coincide for other traits, such as winter hardiness of cereals or mildew resistance in barley.

Especially when applying the **bulk breeding method** in self-fertilizing crops, natural selection may be a 'nuisance' to the breeder. In the bulk breeding procedure the phase of inbreeding (about five generations of selfing) precedes the phase of selection. During the inbreeding phase artificial selection is not applied, but natural selection may eliminate attractive genotypes. Effects of natural selection may be minimized during this phase, for example by applying a wide interplant distance and/or harvesting the same number of inflorescences, fruits or seeds from each of a large number of plants. In the selection phase artificial selection is expected to be relatively efficient, because the genotypes of the offspring obtained from the selected plants are identical to the (homozygous) genotypes of the selected plants. (For this reason selection in the case of **identical reproduction**, see Section 8.1, is relatively efficient).

The single goal of the inbreeding phase is indeed development of homozygous plant material, because such material allows selection among plants with identical reproduction. It is attractive to shorten the duration of the inbreeding phase. This is possible by application of the so-called **single seed descent** (SSD-) method, proposed by Goulden (1939), and especially by means of doubling the number of chromosomes of haploid plants (DH-method, see Section 3.1).

The SSD-method was not applied until about 1970. To avoid selection, from each plant (in F_2 and later generations) only a single seed is used to grow the next generation. Since the plants are not required to produce more than just a single seed they may be grown in a regime allowing a fast succession of the generations. Thus in spring cereals three or four generations may be grown in one year. Natural selection will not occur in as far as it is due to differences in fertility.

The SSD- and the DH-methods have the following advantages over the conventional way of attaining complete homozygosity:

- The development of homozygous plant material requires less time and space
- The methods avoid, when developing pure lines, unintentional selection of (possibly vigorous) heterozygous plants as parents for the next generation (such a selection would delay the progress of the inbreeding process; see Example 6.1-a).

Example 6.7 shows that differences between SSD and DH lines cannot easily be explained.

Example 6.7 Caligari, Powell and Jinks (1987) compared for each of five spring barley crosses 20 pure lines, obtained from the DH-method, with 40 pure lines obtained from the SSD-method. The means of the DH-lines and the SSD-lines were different for a number of characters. Differential (natural) selection during the production of the two types of lines was shown to be less likely as a cause. It was concluded that linked, epistatic loci controlling these traits were the main cause for these differences. Apparently (natural) selection was avoided by the application of the SSD-method.

The former conclusion may be questioned as linkage does only give rise to small differences between the genotypic compositions of the DH-lines and the SSD-lines. (This follows from the comparison of $g_{11,1}$ and $g_{11,\infty}$; see Section 3.2.2.)

The conclusions drawn when comparing results of application of the SSD-method with results of application of conventional breeding procedures appear to be divergent: in some cases the SSD-method was superior (see Example 6.8), in other cases the two approaches were equivalent or the SSD-method was inferior.

Example 6.8 Van Oeveren (1993; p. 91) compared

(i) 'Early selection, with early generation cross selection'; and
(ii) Bulk breeding 'where selection is postponed to a more homozygous generation' (obtained by application of the SSD-method).

In procedure (i) the choice of the crosses ('cross selection') was based on F_3-derived estimates of both the cross mean and the between line variance (Section 11.2.3). It was followed by line selection. This study led to the conclusion (p. 97; *loc. cit.*) that 'early cross selection is not an efficient way of breeding. $\cdots$ the main source of error is the difference in growing conditions between the F_3-selection environment and the predicted F_∞-environment'.

With procedure (ii) effects of intergenotypic competition were largely avoided because the differences in growing conditions between the selection environment and the commercial production environment were relatively small. Van Oeveren (1993; p. 97) concluded: 'The procedure of single seed descent can produce superior inbred lines in a more consistent, cheaper and faster way'.

6.2 The Maintenance of Genetic Variation

In applied plant breeding there is continuous interest in the introduction of new genetic variation. Sources for extending the genetic variation with regard to some crop species are natural populations of the same species or of related species. (Genetic transformation is a rather recently developed way for extending the genetic variation to be exploited for crop improvement.) Often such natural populations appear to accommodate a wealth of genetic diversity. Genetic variation may also be maintained in breeding populations of cultivated crops. This is remarkable, because natural (and/or artificial) selection occurs generation after generation and one might speculate that this implies a continuous reduction of genetic variation. In the absence of human intervention genetic variation is/was, however, often maintained, notwithstanding the continuous selection. With regard to cultivated crops one might even state that plant breeding has stimulated the development and maintenance of a wide genetic diversity. It seems that human interference promotes an increase of the genetic diversity in the involved crop. (In contrast to this, wild plant and animal species suffer from genetic erosion because of annihilation of ecological niches due to human activities. In recent times many species have become completely extinct.)

Ecological population genetics studies the mechanisms responsible for the maintenance genetic diversity. In this section four mechanisms (tentatively) explaining this seemingly paradoxical situation are elaborated, namely

1. overdominance,
2. frequency-dependent fitness,
3. recurrent mutations and
4. immigration of pollen or plants.

Overdominance

Crumpacker (1967) and Allard, Jain and Workman (1968) have presented, for cross-fertilizing and self-fertilizing crops respectively, examples of overdominance with regard to traits controlled by a single locus. Reduced probability of recombination alongside a certain chromosome segment gives rise to a gene cluster. If the loci belonging to the cluster control the same trait, an oligogenic basis for overdominance is present. (In humans such a gene cluster has been shown to control the immune system). These few examples do not represent the common situation.

A more realistic concept is **pseudo-overdominance**, due to alleles linked in repulsion phase. An example is a chromosome segment behaving as a single allele (because recombination within the segment hardly ever occurs). Crossing of two homozygous genotypes, differing for such segment, yields an offspring heterozygous for this segment which, consequently, may exceed both homozygous parents; see Example 9.10.

In 1917 Jones had already stated that hybrid vigour could be due to the assembling of favourable alleles from each of both parents in one genotype. Linkage of such favourable alleles to unfavourable alleles hampers fixation of the superior heterozygous F_1-genotype into an equivalent homozygous genotype. However, it does not exclude such fixation. Results of experiments using electrophoresis substantiate the concept of pseudo-overdominance.

Notwithstanding the previous remarks, many population genetical models, aimed at explaining genetic polymorphisms, have been developed on the basis of a single locus. Population genetic theory (Li, 1976, p. 419) shows that for loci with overdominance, *i.e.* $s_0 > 0$ and $s_2 > 0$, a **stable equilibrium** of the genotypic composition may occur, notwithstanding the selection. Thus a **genetic polymorphism** is maintained, and – in contrast to what was said at the beginning of this chapter – the genotypic composition may be stable, notwithstanding selection. The equilibrium allele frequencies can be derived to be

$$q_e = \frac{s_2}{s_0 + s_2} \text{ and } p_e = \frac{s_0}{s_0 + s_2} \tag{6.3}$$

thus $0 < p_e < 1$ (see, however, Note 6.1).

Note 6.1 One may criticize the derivation undertlying Equation (6.3) on two grounds:

1) It is based on the assumption that the preceding generation had the Hardy–Weinberg genotypic composition. This composition applies in the case of mass selection occurring before pollen distribution. Selection with regard to vitality is thus, implicitly, assumed not to occur.
2) Overdominance with regard to a single locus is a rare event.

Frequency-dependent fitness

The concept of frequency-dependent fitness is based on the fascinating observation that it is, under constant ecological conditions, both rare for plants (or animals) with a certain genotype to be completely extinct as well as rare that the frequency of plants with the considered genotype grows unrestricted. Apparently, there are mechanisms regulating the number of individuals with a certain genotype in such a way, that the number increases if it is low and that it decreases if it is high (see Example 6.9).

Example 6.9 Two examples of frequency-dependent fitness are mentioned here:

1. The seed-set of male sterile barley plants (with genotype mm) may depend on the frequency of such plants. Section 5.2.2 refers to the relation $w_0 = 0.6 - f_0$.

2. In the case of self-incompatibility, a low frequency of a genotype for the incompatibility locus/loci tends to be associated with a higher fitness of the genotype than the fitness of a genotype with a higher frequency.

A tentative explanation for genotypes to have a frequency-dependent fitness is as follows. Plants with the same genotype tend to have similar demands, at the same time. These demands are specific for the genotype. Among the plants with a certain genotype, more plants will survive the 'struggle' for the same, restrictedly available resources, as the genotype's frequency is lower. Plants with a genotype with a relatively low frequency may thus tend to have a relatively high fitness. This phenomenon might apply to genotypes adapted to rare environmental conditions. Such genotypes are favoured by selection. Mather (1973) called such selection **disruptive selection**. It may lead to distinct types or it may be balanced by **stabilizing selection**, for example by the genotype adapted to rare environmental conditions becoming increasingly common.

Recurrent mutations

Mutations are, in fact, the ultimate source of all genetic diversity. However, their frequencies are generally very low (see Note 6.2). Thus in the equilibrium between the production of a new allele and its elimination, if it does not give rise to a better adapted phenotype, the new allele will have a (very) low frequency. It is concluded that recurrent mutations should not be considered as a quantitatively important factor for maintenance of genetic diversity.

Note 6.2 The frequency of the occurrence of a mutation is very low. Furthermore, one should realize that a mutant allele is not transmitted to the next generation when the mutation occurs outside the chain of cells connecting two generations, the so-called **germ-line**. Such mutations have no population genetical implications. This concerns mutations in cells of roots, stems, leaves, style, stigma, seed coat, connectivum, *etc.*

Immigration of pollen or plants

The effect of immigration of pollen or plants on the genotypic composition of the considered population depends on

- the difference in the allele frequencies of 'donor' and 'recipient' and
- the extent of the immigration

Both factors may play a role in legislation concerning mutual isolation distances required at the multiplication of seed of varieties of cross-fertilizing crops.

It is emphasized here that **introgression** means the incorporation by crossing and repeated backcrossing of alleles originating from a *different* species. This may occur spontaneously or as a breeding activity.

Alleles may immigrate into a population in different ways:

(i) Flow of pollen, transported by wind or by insects
(ii) Mixing, intended or not, of seed lots representing different varieties

Flow of pollen

We define q as the frequency of allele a in the recipient, q_m as the frequency of a among the immigrating pollen, and m as the proportion of immigrating pollen among the effective male gametes. The frequency, q', of the effective pollen grains with haplotype a is

$$q' = (1 - m)q + mq_m$$

The case of immigrating pollen situation can be considered as a form of bulk crossing (Section 2.2.1). According to Equation (2.2) the frequency of a in the 'hybrid' population will be

$$q_1 = \tfrac{1}{2}(q + q') = \tfrac{1}{2}[q + (1 - m)q + mq_m] = q + \tfrac{1}{2}m(q_m - q)$$

Thus

$$\Delta q = q_1 - q = \tfrac{1}{2}m(q_m - q)$$

This expression contains both factors mentioned before. For $q_m = q$ or for $m = 0$ the allele frequency will not change. For $m > 0$ the expression yields of course $\Delta q > 0$ if $q_m > q$ and $\Delta q < 0$ if $q_m < q$.

If immigration occurs generation after generation, selection aiming at the elimination of allele a will never succeed. Then, notwithstanding selection, a genetic polymorphism is maintained.

Mixing of seed

This case is considered as immigration of sporophytes. For a diploid crop one can then derive:

$$\Delta q = m(q_m - q)$$

In certain situations immigration of sporophytes is applied intentionally, *e.g.* as a remedy against genetic erosion in populations of a small size.

6.3 Artificial Selection

6.3.1 Introduction

When applying selection in a self-fertilizing crop it is irrelevant whether the trait is expressed before or after pollen distribution: the plants selected are

simultaneously selected both as female and as male plants. For annual cross-fertilizing crops, however, the time of expression of the trait of interest, *i.e.* before or after pollen distribution, and consequently the time of the selection, has important impact on the efficiency of the selection. If the trait is expressed after pollen distribution, there is no selection with regard to the plants as male parents. All plants contribute pollen from which the next generation is generated. The selection implies selection among plants as female parents. Only the selected plants contribute eggs from which the next generation is generated. Example 6.10 mentions for each of a few cross-fertilizing crops a trait that is expressed either before or after pollen distribution.

Example 6.10 Traits of cross-fertilizing crops expressed before pollen distribution are

- The colour of the midrib of leaves of maize plants: brown-midrib plants have a lower lignin content than green-midrib plants and are more easily digested as silage maize (Barrière and Argillier, 1993)
- The coleoptile colour of seedlings of rye
- The reaction of spinach plants to inoculation with *Perenospora spinaciae*

Traits of these crops expressed after pollen distribution are

- The colour of the cob of the ears of maize plants
- The colour of the kernels produced by rye plants
- The shape of the seeds produced by spinach plants (they can be smooth or prickly)

If the genetic control of the trait of interest is characterized by **incomplete dominance** the genotype of each plant (be it aa, Aa or AA) can be derived from its phenotype. A population exclusively consisting of plants with the desired genotype can then, under certain conditions, easily be obtained. These conditions concern the mode of reproduction of the crop and/or the time of the expression of the trait. Such easy and successful selection is possible:

- If the crop is a self-fertilizing species
- If the crop is a cross-fertilizing species, and if the trait is expressed before pollen distribution
- If the crop is a cross-fertilizing species, if the trait is expressed after pollen distribution and if the species permits selfing to be carried out successfully. (If the latter is impossible, *e.g.* due to dioecy or self-incompatibility, one could cross random plants in pairwise combinations. Later, after expression of the trait, one may harvest the seeds due to crosses where both plants involved appear to have the desired genotype.)

Because the case of incomplete dominance will not impose problems, in the present chapter attention is only given to procedures for selection with regard to a trait with qualitative variation, controlled by a single locus

accommodating an allele with **complete dominance**. The desired expression for the considered trait may be due to

(i) Genotype aa
In this case allele A is to be eliminated from the population
(ii) Genotypes Aa and AA
In this case allele a is to be eliminated from the population.

Initially, it will be assumed that the candidates (lines, families or populations) consist of an infinitely large number of plants. In practice, however, the candidates will consist of a limited number of plants. Thus the minimal acceptable number of plants per candidate will also be considered.

Selection for genotype aa

If the trait is expressed before pollen distribution, mass selection before pollen distribution suffices to eliminate the undesired allele A at once. If the trait is expressed after pollen distribution selfing of a large number of plants is most appropriate. As soon as the trait is expressed, one may harvest the plants that appear to have genotype aa. If selfing is impossible, one can cross random plants pairwise. After expression of the trait one may harvest the seed due to crosses where both involved plants appear to have genotype aa.

To reduce the probability of a non-negligible shift in the frequencies of alleles at loci not affecting the selected trait, a high number of plants with genotype aa should be retained.

Selection for genotype AA

If the desired trait expression is due to genotype AA or Aa, selection is required to eliminate the recessive allele a, which may hide in heterozygous genotypes. Sections 6.3.2 to 6.3.6 are dedicated to this task. In these sections procedures are elaborated for different situations, *i.e.* whether

- Self-fertilization is possible or not
- The trait is expressed before or after pollen distribution

Line selection (Section 6.3.2) is the most efficient selection method if self-fertilization is possible. It allows for complete elimination of allele a within a short period of time. If self-fertilization is impossible, a less efficient selection method should be used. Ranked according to decreasing efficiency (in a genetical sense) attention will be given to

- Full sib family selection (Section 6.3.3)
- Half sib family selection (Section 6.3.4)
- Mass selection (Section 6.3.5)

A somewhat different approach is genotype assessment on the basis of a progeny test (Section 6.3.6): selection among the candidate plants only takes place after having determined their genotype from their offspring.

The general features of **line selection** are the following:

1. In as far as they are cultivated, the lines are evaluated as a whole. Lines containing plants with genotype *aa* are eliminated.
2. Within retained lines, single-plant selection is either applied (**combined selection**) or omitted.
3. The next generation is grown in separate plots tracing back to:
   - seed produced by separate plants selected in retained lines (this procedure is called **pedigree selection**) or
   - seed produced by separate accepted lines.

The general features of **family selection** are

1. In as far as they are cultivated, the families are evaluated as a whole. Families containing plants with genotype *aa* are eliminated.
2. Within retained families, single-plant selection is either applied or omitted (the latter situation is elaborated in Sections 6.3.3 and 6.3.4).
3. The next generation is grown on separate plots tracing back to:
   - seed produced by separate plants belonging to the evaluated (and retained) families,
   - seed produced by the evaluated (and retained) families or
   - seed produced by sibs of the evaluated (and retained) families (**sib selection**; see Note 6.3)

Note 6.3 Reasons to apply sib selection are

1. The evaluation is destructive or requires a cultivation procedure deviating from the one preferred for seed production, *e.g.* radish.
2. At the evaluation, possibly at several locations, interfamily pollination may occur spontaneously. It is, of course, preferable to prevent pollination of retained families by eliminated families. This is applied in the remnant seed procedure (Section 6.3.4), as well as at modified ear-to-row selection (Section 14.3.1).

In Section 3.1, the terms full sib family (FS-family) and full sib mating (FS-mating) were defined. In the case of self-incompatibility, the pairwise crossing, required to produce an FS-family, occurs spontaneously by growing together, but isolated from other plant material, two cross-compatible, synchronously flowering genotypes. In grass breeding this is applied by growing pairs of clones in isolation. Each FS-family constitutes a subpopulation in the sense of Section 2.1. Thus FS-mating occurs if, within each of a number of FS-families, either plants are crossed in pairs or if open pollination occurs. FS-family selection is applied predominantly in crops such as sugar beet (*Beta vulgaris* L.), grasses and oil palm.

Open pollination yields, after separate harvesting of the involved plants, half sib families. These HS-families consist of plants that are each other's

half sibs because they descend from the same maternal parent, but possibly from different paternal parents. (In animal breeding it is common that the individuals belonging to the same HS-family descend from the same father. The situation of a common father is, of course, also possible in plant breeding.) HS-family selection is commonly applied in crops like rye, maize or grasses.

The general features of **mass selection** are

1. Individual plants are rejected or selected on the basis of their phenotype. (For traits with quantitative variation each plant's phenotype might be evaluated on the basis of a comparison with the phenotypes of other, unrelated plants.)
2. The offspring of all selected plants are grown in bulk.

To describe the effect of selection, the meaning of the notation introduced in Note 2.4 is somewhat modified. *The last subscript in a symbol representing a haplotype or a genotype frequency still refers to the rank of the generation to be generated, but in Section 6.3 this rank indicates the number of preceding generations exposed to selection. The symbol designating a population as retained after selection, differs from the symbol designating the original population (before the selection), by addition of a prime.*

6.3.2 Line selection

The trait is expressed before pollen distribution

In the source population, say G_0, plants with the acceptable phenotype, due to genotype Aa or AA, are selfed. These plants are separately harvested. The line selection starts thus with mass selection. The offspring are grown and evaluated ear-to-row, *i.e.* as separate lines. Segregating lines in this generation, *i.e.* in population G_1, descend from parents with genotype Aa. These lines are eliminated before pollen release. The retained subset of lines constitutes population G_1'. It does not anymore contain allele a.

This efficient selection procedure can be applied to self-fertilizing crops as well as to cross-fertilizing crops. In strictly self-fertilizing crops, it does not even matter whether the trait under selection is expressed before or after pollen distribution. In cross-fertilizing crops the non-segregating lines may interpollinate to cancel the decrease of the frequency of heterozygous plants due to the selfing. This eliminates possible inbreeding effects with regard to quantitative traits.

The trait is expressed after pollen distribution

It was stated above that in strictly self-fertilizing crops the time of the expression of the trait under selection, *i.e.* before or after pollen release, does not matter. The present paragraph concerns, therefore, cross-fertilizing crops.

The procedure starts with the selfing of many plants of population G_0. After expression of the trait of interest, one can distinguish plants with genotype AA or Aa from plants with genotype aa. Elimination of plants with genotype aa yields population G_0'. The line selection starts thus with mass selection.

The further pathway of the procedure depends on whether a 'small' or a 'large' number of seeds are obtained after selfing of a retained plant. Note 6.4 considers the question 'What is a large number of seeds?'

Note 6.4 The number of plants evaluated per line, say N, is often small; possibly simply due to the fact that the enforced selfings yield small numbers of seeds. Hopefully it is large enough for the probability of absence of plants with genotype aa, in a line obtained from an Aa plant, to be small. The value for N, such that this probability is not more than 0.01, is interesting. Say, $\underline{k}$ = the number of plants with genotype aa among the N plants in a line. The probability of absence of plants with genotype aa, in a line obtained from an Aa plant, is:

$$P(\underline{k} = 0|\text{parental genotype } Aa) = \left(\frac{3}{4}\right)^N$$

For $N > 16$, this probability is less than 0.01.

- A small number of seeds are available per line
  Population G_1 consists of ear-to-row grown, mutually isolated lines. Open pollination occurs spontaneously within each line. After expression of the trait of interest, one can distinguish segregating lines, descended from plants with genotype Aa, from non-segregating lines, descended from plants with genotype AA. The set of non-segregating lines constitute population G_1'. Allele a is absent in this population.
  Population G_1' is harvested in bulk. The seeds constitute population G_2. Spontaneous open pollination in G_2 eliminates the deficit of heterozygous plants, which is due to the selfing and/or within-line open pollination.
- A large number of seeds are available per line
  If the selfing of the plants yields large numbers of seeds, the remnant seed procedure can be applied. Per line a part of the seed representing the line is grown and evaluated ear-to-row. Open pollination among the lines constituting population G_1 may occur. After expression of the trait of interest, one can identify the non-segregating lines. (These constitute population G_1'). Allele a is absent in G_1'. Remnant seed representing the lines constituting population G_1' is bulked. Spontaneous open pollination among the plants constituting the bulk removes the deficit of heterozygous plants which is due to the selfing.

In both the above procedures allele a is absent already in population G_1'. However, the second approach avoids the laborious mutual isolation of the lines required for the first approach.

A trait of an autotetraploid crop expressed after pollen distribution

In generation G_0 many plants are selfed. After expression of the trait of interest, but before harvest time, plants with genotype *aaaa* are discarded. Population G_1 consists thus of lines originating from plants with genotype *Aaaa*, *AAaa*, *AAAa* or *AAAA*. (Table 3.5 presents for each parental genotype the genotypic composition of the line). The lines constituting generation G_1 are grown in mutual isolation. Lines obtained from a parental plant with genotype *Aaaa* or *AAaa* will segregate (see, however, Note 6.5).

Note 6.5 In population G_1 the number of plants per line, say N, should of course be large enough to ensure that the probability of absence of nulliplex plants in lines obtained from *Aaaa* or *AAaa* plants is small.

Say, $\underline{k}$ = the number of nulliplex plants among the N plants in the line. Then:

$$P(\underline{k} = 0|\text{parental genotype } Aaaa) = \left(\frac{3}{4}\right)^N$$

$$P(\underline{k} = 0|\text{parental genotype } AAaa) = \left(\frac{35}{36}\right)^N$$

These probabilities are less than 0.01 for $N > 16$, and $N > 163$, respectively. The number of plants per line should thus amount at least to 163 to identify (and consequently eliminate) lines descending from *Aaaa* or *AAaa*.

Population G_1' consists of the subset of lines obtained from plants with genotype *AAAa* or *AAAA*. Random mating occurs within each line belonging to G_1'. The haplotypic composition of the gametes produced by a line obtained from a *AAAa* plant can be derived to be

	Haplotype		
	aa	*Aa*	*AA*
f	$\frac{1}{24}$	$\frac{10}{24}$	$\frac{13}{24}$

The genotypic composition of the progeny of this line is

	Genotype				
	aaaa	*Aaaa*	*AAaa*	*AAAa*	*AAAA*
f	$\frac{1}{576}$	$\frac{20}{576}$	$\frac{126}{576}$	$\frac{260}{576}$	$\frac{169}{576}$

This implies that the probability that not a single *aaaa* plant occurs in the progeny is high if the progeny size is (rather) small. One may accept that risk

and bulk the progenies from lines descending from $AAAa$ with the progenies from lines descending from $AAAA$. (Complete elimination of allele a may be pursued by genotype assessment, see Note 6.6.)

Note 6.6 Lines descending from $AAAa$ can be distinguished from lines descending from $AAAA$, by separate pollination of $aaaa$ plants with pollen collected from each line.

The genotypic composition of families obtained from $AAAa$ is

	Genotype				
	$aaaa$	$Aaaa$	$AAaa$	$AAAa$	$AAAA$
f	$\frac{1}{24}$	$\frac{10}{24}$	$\frac{13}{24}$	0	0

Families consisting of at least 109 plants are then required to ensure that

$$P(\underline{k} = 0|\text{line from } AAAa)$$

is less than 0.01.

6.3.3 *Full sib family selection*

FS-family selection is a very efficient procedure. It deserves application whenever the efforts required to produce the families are not unsurmountable. The crossing should thus not be too laborious. In crops where a successful pollination yields only one seed one might consider the application of half sib family selection to half sib families obtained by open pollination, but one should realize that this cheap alternative is rather inefficient (see Section 6.3.4). In self-incompatible crops yielding only one seed after a successful pollination (like in grasses or rye) the production of large numbers of seed per cross does not require large efforts if one bags together one or more inflorescences of the two plants to be crossed.

The trait is expressed before pollen distribution

The genotypic composition of the original population G_0 is $(f_{0,0}, f_{1,0}, f_{2,0})$. Plants with genotype aa will not be involved in a pairwise cross. This implies that mass selection, transforming G_0 into $G_0{}'$, with genotypic composition $(0, f_{1,0}{}', f_{2,0}{}')$, is applied prior to the pairwise crossing generating the FS-families.

With regard to pairwise crosses between plants with genotype Aa or AA one can distinguish three types of crosses. Table 6.1 presents for each type of cross its frequency and the genotypic composition of the obtained FS-family.

Table 6.1 Pairwise crosses between plants with genotype Aa or AA: the types of crosses, their frequencies and the genotypic composition of the obtained FS-families

		Genotype			
Type of cross	Frequency	aa	Aa	AA	Segregation visible
1. $Aa \times Aa$	${f_1}'^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	yes
2. $AA \overset{\times}{\times} Aa$	$2f_1'f_2'$	0	$\frac{1}{2}$	$\frac{1}{2}$	no
3. $AA \times AA$	$f_2'^2$	0	0	1	no

FS-families of type 1 will segregate before pollen distribution with a probability of at least 0.99 if they consist of at least 16 plants. Elimination of such families transforms population G_1 into population ${G_1}'$. The families constituting ${G_1}'$ are grown in mutual isolation. (The reason for this is explained in Note 6.7). Population G_2 consists then of family-derived bulks. In contrast to bulks tracing back to a cross of type 3, bulks tracing back to a type 2 cross may contain aa plants. For this reason, the bulks are separately grown and evaluated. The genotypic composition of a bulk descending from a type 2 FS-family is $\left(\frac{1}{16}, \frac{6}{16}, \frac{9}{16}\right)$. If such bulks consist of at least 72 plants, they will segregate before pollen distribution with a probability of at least 0.99 (Why?). Elimination of these bulks before pollen distribution transforms population G_2 into population ${G_2}'$, consisting of bulks descending from type 3 FS-families.

This procedure leads to absence of allele a in generation ${G_2}'$. (With line selection, Section 6.3.2, this goal is already attained in population ${G_1}'$.) The slight inbreeding in generation ${G_1}'$ is undone by random mating (across bulks) in population ${G_2}'$. FS-family selection involving a single generation with FS-mating is thus an attractive selection procedure for obligatory cross-fertilizing crops.

Note 6.7 Mutual isolation of the FS-families is applied because type 2 families contain the a allele to be eliminated. Such families should not pollinate type 3 families.

Isolation enforces random mating within each of the families constituting ${G_1}'$, *i.e.* FS-mating at the level of the superpopulation. It may be replaced by a number of pairwise crosses within each acceptable family. The seeds resulting from these crosses are bulked per family. For the rest the procedure proceeds as described in this section.

The effect of avoiding FS-mating, by not applying in population ${G_1}'$ mutual isolation of the non-segregating families of type 2 and 3, is now considered. The genotypic compositions of populations G_1 and ${G_1}'$ are $(f_{0,1}, f_{1,1}, f_{2,1})$

and $(0, {f_{1,1}}', {f_{2,1}}')$, respectively, where

$${f_{1,1}}' = \frac{\frac{1}{2}\left(2{f_{1,0}}'{f_{2,0}}'\right)}{1 - {f_{1,0}}'^2} = \frac{{f_{1,0}}'}{1 + {f_{1,0}}'}$$

because ${f_{2,0}}' = 1 - {f_{1,0}}'$ and, consequently, ${f_{2,1}}' = \frac{1}{1+{f_{1,0}}'}$.

The haplotypic composition of the gametes produced by population ${G_1}'$ is $(g_{0,2}, g_{1,2})$, where

$$g_{0,2} = {q_1}' = \frac{1}{2}{f_{1,1}}' = \frac{\frac{1}{2}{f_{1,0}}'}{1 + {f_{1,0}}'} = \frac{{q_0}'}{1 + 2{q_0}'}$$

This implies

$${q_t}' = \frac{{q_{t-1}}'}{1 + 2{q_{t-1}}'} = \frac{\frac{{q_{t-2}}'}{1+2{q_{t-2}}'}}{1 + 2\left(\frac{{q_{t-2}}'}{1+2{q_{t-2}}'}\right)} = \frac{{q_{t-2}}'}{1 + 4{q_{t-2}}'}$$

thus

$${q_t}' = \frac{{q_0}'}{1 + 2t{q_0}'} \tag{6.4}$$

Effectively the absence of mutual isolation implies pairwise crossing of plants, belonging to non-segregating families, with genotype Aa or AA. It is an ineffective procedure: complete elimination of allele a is only asymptotically attained! Application of this procedure in practical breeding, *e.g.* in sugar beet breeding aiming at quantitative traits like sugar content and root weight, is in fact inefficient.

We consider now h, *i.e.* the number of generations with FS-family selection with regard to a trait expressed before pollen distribution required to half ${q_0}'$, the initial frequency of allele a, when avoiding FS-mating. The above equation implies

$${q_h}' = \frac{{q_0}'}{1 + 2h{q_0}'} = \frac{{q_0}'}{2}$$

Thus

$$1 + 2h{q_0}' = 2$$

if

$$h = \frac{1}{2{q_0}'} \tag{6.5}$$

To reduce the probability of random fixation (see Chapter 7), the number of non-segregating bulks should amount to at least 25.

The trait is expressed after pollen distribution

A large number of plants belonging to population G_0 is used for making pairwise crosses. After expression of the trait, crosses involving one or two plants with genotype *aa* are eliminated. The plants involved in the other crosses are retained as population G_0'. In this way only the three types of FS-families distinguished in Table 6.1 occur in population G_1. Because these types differ with regard to the frequency of allele *a*, the families constituting G_1 are grown in mutual isolation to enforce FS-mating. (Note 6.7 indicates that the mutual isolation of the families may be replaced by controlled pairwise crossing within each FS-family).

FS-families of type 1 will segregate after pollen distribution. These families are eliminated. The retained families constitute generation G_1'. They are separately harvested as family-derived bulks. In generation G_2 these bulks are grown in mutual isolation. Bulks descending from a type 2 cross will segregate after pollen distribution. These bulks are to be eliminated. The other bulks, constituting generation G_2', do not contain allele *a*. The seeds produced by these bulks can be pooled. This selection procedure leads to absence of allele *a* in population G_2'. (With line selection, Section 6.3.2, this goal is already attained in population G_1'.)

Open pollination in generation G_3 will eliminate the homozygosity due to the inbreeding enforced by the mutual isolation of the FS-families and the bulks.

The mutual isolation of the family-derived bulks constituting population G_2 may be omitted if each family-derived bulk is represented by a large amount of seed. A part of this seed (at least 72 seeds per bulk) is used to identify in generation G_2 bulks not containing allele *a*. After expression of the trait, mixing of remnant seed representing non-segregating bulks yields generation G_2', in which allele *a* is absent.

In the present as well as in the previous section a few efficient selection procedures were described in just a few words. One should realize, however, that their execution can be quite laborious. Three aspects are briefly considered:

(i) Mutual isolation implies a lot of additional work.
It is interesting to compare procedures employing mutual isolation of the FS-families (and implying enforced FS-mating) with procedures avoiding such isolation. In Note 6.7 the comparison was elaborated for traits expressed before pollen distribution. We now consider FS-family selection with regard to a trait expressed after pollen distribution in the absence of mutual isolation of the families.

In each generation pairwise crosses are made at random, within as well as between FS-families. After expression of the trait only crosses involving plants belonging to non-segregating families are retained. Thus, effectively only plants with genotype *Aa* or *AA* belonging to families of

type 2 or 3 are crossed. This coincides with the ineffective procedure described in Note 6.7.

(ii) To reduce the probability of random fixation with regard to loci not involved in the genetic control of the considered trait, one should start in generation G_0 with making a lot of selfings (when applying line selection) or a lot of crosses (when applying FS-family selection).

(iii) To identify – with some minimum probability – potentially segregating lines, families or family-derived bulks, the number of plants representing such entries should not be to small. Above it was said that family-derived bulks should consist of at least 72 plants. For oil palm this requires, at a commercial plant density, about 5,000 m^2 per entry!

6.3.4 *Half sib family selection*

The trait is expressed before pollen distribution

As with FS-family selection with regard to a trait expressed before pollen distribution, the genotypic composition of the initial population G_0, *i.e.* $(f_{0,0}, f_{1,0}, f_{2,0})$, is first transformed by mass selection into that of $G_0{}'$, *i.e.* $(0, f_{1,0}{}', f_{2,0}{}')$. Open pollination among the plants constituting $G_0{}'$ yields two types of HS-families at harvest. Table 6.2 gives their genotypic compositions.

These families are grown and evaluated ear-to-row. Elimination, before pollen distribution, of segregating HS-families, *i.e.* type 1 families, transforms population G_1 into $G_1{}'$. The genotypic composition of $G_1{}'$ is $(0, f_{1,1}{}', f_{2,1}{}')$ with

$$q_1{}' = \tfrac{1}{2} f_{1,1}{}' = \tfrac{1}{2} q_0{}'$$

A single generation with HS-family selection leads thus to halving of the frequency of allele a. This implies for continued HS-family selection:

$$q_t{}' = (\tfrac{1}{2})^t q_0{}' \tag{6.6}$$

Complete elimination of allele a is only asymptotically attained. The effort required for a progressively smaller decrease of the frequency of allele a

Table 6.2 Open pollination among plant with genotype Aa or AA: the maternal genotypes, their frequencies and the genotypic composition of the obtained HS-families

Maternal genotype	Frequency	Genotypic composition of the obtained HS-family			Segregation visible
		aa	Aa	AA	
1. Aa	$f_{1,0}{}'$	$\frac{1}{2} q_0{}'$	$\frac{1}{2}$	$\frac{1}{2} p_0{}'$	yes
2. AA	$f_{2,0}{}'$	0	$q_0{}'$	$p_0{}'$	no

becomes progressively greater, see Note 6.8. This approach (and the procedure described hereafter) is very inefficient when the aim is to eliminate completely a recessive allele.

Note 6.8 In population G_{t+1} the genotypic composition of a type 1 HS-family is $(\frac{1}{2}q_t', \frac{1}{2}, \frac{1}{2}p_t')$. The probability that a type 1 HS-family consisting of N plants does not segregate is $(1 - \frac{1}{2}q_t')^N$. Identification of a type 1 HS-family with a probability of at least 0.01 requires that the family size is at least $\frac{\log(0.01)}{\log(1-\frac{1}{2}q_t')}$. The smaller q_t' the higher the required number of plants per HS-family. For $q_t' = 0.05$ it should be 182 plants, and for $q_t' = 0.01$ it should be as many as 919 plants.

Identification of potentially segregating HS-families requires thus ever increasing family sizes!

The trait is expressed after pollen distribution

If the trait is expressed after pollen distribution one should prevent inter-pollination between type 1 and type 2 HS-families (Table 6.2). This may be done by:

1. mutual isolation of the HS-families or
2. application of the remnant seed procedure.

Mutual isolation of the HS-families

Mutual isolation of the HS-families constituting population G_1 imposes HS-mating within each family. After expression of the trait, type 1 families and type 2 families can be distinguished. Elimination of type 1 families transforms population G_1 into G_1'. Plants in G_1' are separately harvested and their seed is grown ear-to-row in generation G_2. Mutual isolation induces again HS-mating. Effectively only type 2 families, harvested from in type 2 families from plants with genotype AA, are retained. Type 1 families are eliminated.

The initial population G_0 is transformed by mass selection into G_0' with genotypic composition $(0, f_{1,0}', f_{2,0}')$. HS-family selection after expression of the trait transforms population G_1 into G_1' with genotypic composition $(0, f_{1,1}', f_{2,1}')$, with

$$q_1' = \tfrac{1}{2} f_{1,1}' = \tfrac{1}{2} q_0$$

Within the type 2 families of population G_1', the frequency of pollen with haplotype a is q_1'. This implies that the frequency of Aa plants in the type 2 families constituting population G_2 is q_1'. Thus

$$q_2' = \tfrac{1}{2} q_1'$$

Except after the HS-family selection in population G_1, this procedure implies

$$q_{t+1}' = \tfrac{1}{2} q_t'$$

The reduction of the frequency of allele a is thus 50% per generation when applying the present procedure for HS-family selection with regard to a trait expressed after pollen distribution. The efforts required for such progressively smaller reductions become progressively larger. The reduction requires continued HS-mating. The eventual goal, *i.e.* complete elimination of allele a is only asymptotically attained. It is concluded that this procedure is not to be recommended.

Application of the remnant seed procedure

Application of the **remnant seed procedure** is quite common for traits expressed after pollen distribution. With this procedure each HS-family is sown at two dates in such a way that the first sown part of each family can be evaluated before the later sown part distributes pollen. On the basis of observations concerning the first sown set of families, one eliminates, before pollen distribution, all type 1 families from the later sown set. For annual crops the sowing of the two sets of families may occur in two successive years. The progress is then rather slow. A faster procedure is cultivation of the first and the second set in such a way that an additional growing season is not required. This may imply use of a greenhouse or cultivation in the other hemisphere.

The reduction of the frequency of allele a is the same as the reduction at selection with regard to a trait expressed before pollen distribution. The frequency of allele a thus obeys Equation (6.6). However, the procedure requires more effort than selection with regard to a trait expressed before pollen distribution, and it tends to last longer.

In comparison to mutual isolation of the HS-families, the remnant seed procedure has the advantage of avoiding continued HS-mating as well as the efforts required for mutual isolation. Note 6.9 concerns some historical facts as well as some concluding remarks concerning HS-family selection.

Note 6.9 The terms 'ear-to-row selection' (Allard, 1960, p. 189) and 'modified ear-to-row selection' (Lonnquist, 1964) only imply separate cultivation of progenies. Because mutual isolation is not necessarily required these terms are meaningless in the context of breeding procedures. Poehlman and Sleper (2006) used the term 'ear-to-row breeding' for a procedure (in fact for the so-called Ohio-method for ear-to-row breeding), that we refer to as remnant seed procedure. This procedure is originally due to the German breeder Roemer. With the so-called Illinois-method of ear-to-row breeding the best plants are selected from the best families (in this book this is called: combined selection). One should, consequently, be careful with using the term 'ear-to-row selection'. The separate sowing of lines or families may, however, efficiently be called 'ear-to-row planting'.

None of the HS-family selection procedures leads to complete elimination of allele a within a few generations. The frequency of a approaches the value 0

asymptotically. Certainly application of line selection or FS-family selection in stead of HS-family selection is to be advised.

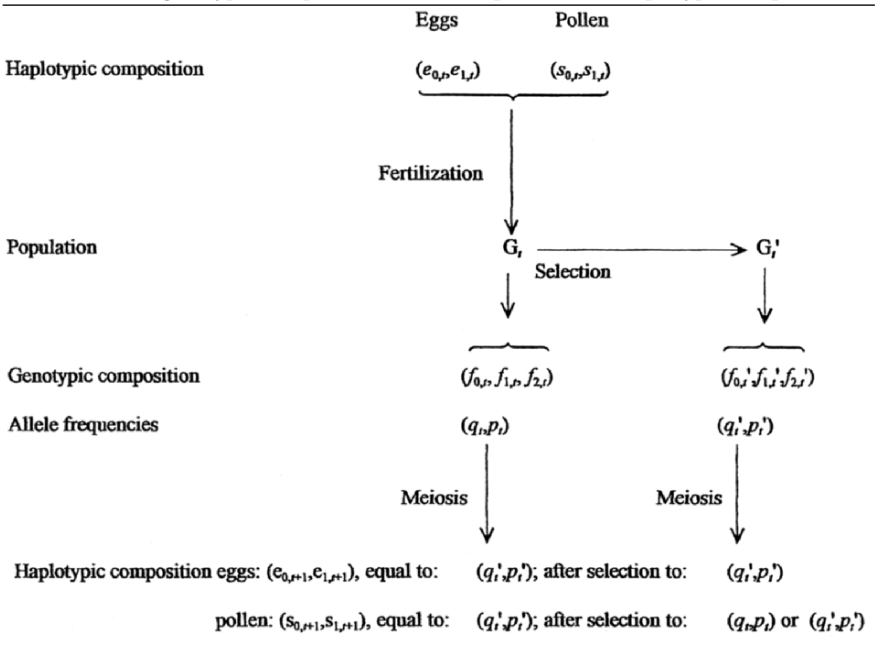
Again (like at the end of Section 6.3.3) attention is drawn to the probability of fixation: to keep this probability small the number of type 2 HS-families should never be less than 25.

6.3.5 *Mass selection*

In the case of mass selection, open pollination occurs. The haplotype frequencies among the female gametes may then deviate from the haplotype frequencies among the male gametes. Thus parameters are introduced to designate female and male haplotype frequencies. Table 6.3 describes the process of selection in terms of these parameters.

For the eggs giving rise to population G_{t+1}, the frequencies of haplotypes a and A are represented by $e_{0,t+1}$ and $e_{1,t+1}$, respectively. They are equal to the allele frequencies in population G_t', the part of parental population G_t surviving the mass selection. For the pollen giving rise to population G_{t+1},

Table 6.3 The process of mass selection and the notation used to indicate generations and to describe genotypic compositions, allele frequencies and haplotypic compositions



the frequencies of haplotypes a and A are represented by $s_{0,t+1}$ and $s_{1,t+1}$, respectively. They adopt the following values:

- In the case of selection with regard to a trait expressed before pollen distribution they are equal to the allele frequencies in generation $G_t{}'$.
- In the case of selection with regard to a trait expressed after pollen distribution they are equal to the allele frequencies in generation G_t, the original parental population.

The trait is expressed before pollen distribution

The initial population G_0, with genotypic composition $({q_0}^2, 2p_0q_0, {p_0}^2)$ is transformed before pollen distribution into population $G_0{}'$, with genotypic composition $(0, {f_{1,0}}', {f_{2,0}}')$ and allele frequencies:

$$q_0{}' = \tfrac{1}{2}{f_{1,0}}' = \frac{p_0q_0}{1-{q_0}^2} = \frac{q_0}{1+q_0}$$

and

$$p_0{}' = 1 - q_0{}' = \frac{1}{1+q_0}$$

The haplotypic composition of the gametes produced by $G_0{}'$ is $(g_{0,1}, g_{1,1})$, where $g_{0,1} = q_0{}'$ and $g_{1,1} = p_0{}'$. Thus q_1, the frequency of allele a in population G_1, is equal to $q_0{}'$, or

$$q_1 = \frac{q_0}{1+q_0}$$

Likewise one can derive

$$q_2 = \frac{q_1}{1+q_1} = \frac{\frac{q_0}{1+q_0}}{1+\frac{q_0}{1+q_0}} = \frac{q_0}{1+2q_0}$$

For G_t this means

$$q_t = \frac{q_0}{1+tq_0} \tag{6.7}$$

This equation resembles Equation (6.4), derived for continued FS-family selection with regard to a trait expressed before pollen distribution at avoidance of FS-mating.

As in Note 6.7, the number of generations required to half the initial frequency of allele a is considered. Equation (6.7) implies

$$q_h = \frac{q_0}{1+hq_0} = \tfrac{1}{2}q_0$$

This applies if

$$h = \frac{1}{q_0} \tag{6.8}$$

When $q_0 \approx 1$ the frequency of allele a is approximately halved when applying mass selection for a single generation, but if $q_0 \approx 0$ mass selection should be

applied for numerous generations for that (which then implies a very small actual reduction of q). It is noteworthy that the present value for h is twice that derived for FS-family selection in absence of FS-mating (Equation (6.5)).

The reduction of the frequency of allele a due to elimination, before pollen distribution, of plants with genotype aa is illustrated in Example 6.11.

Example 6.11 A trait expressed before pollen distribution and controlled by locus A-a is considered. Plants with genotype aa are eliminated prior to pollen distribution. The frequency of allele a in populations G_1, G_2, G_3 and G_4 is calculated by means of Equation (6.7) for each of three values of q in the initial population (see also Example 6.12). This yields

	q		
G_0	0.80	0.50	0.20
G_1	0.44	0.33	0.17
G_2	0.31	0.25	0.14
G_3	0.24	0.20	0.13
G_4	0.19	0.17	0.11

It appears that the reduction of the frequency of allele a is greater as q is higher. For $q_0 = 0.2$, four generations with mass selection do not yet suffice to halve the initial allele frequency.

The lessening in the reduction of the frequency of a is caused by the fact that relatively more and more a alleles remain hidden in heterozygous genotypes. The total frequency of a alleles is $q^2 + pq$. An ever increasing portion, *i.e.*

$$\frac{pq}{q^2 + pq} = p$$

occurs in heterozygous plants, which are not eliminated.

Complete elimination of allele a is achieved asymptotically. Mass selection is only efficient in improving a population as long as the population contains plants with the undesired phenotype in a high frequency.

The trait is expressed after pollen distribution

Population G_t, with genotypic composition $(f_{0,t}, f_{1,t}, f_{2,t})$, is transformed by selection into $G_t{}'$, with genotypic composition $(0, f_{1,t}{}', f_{2,t}{}')$. According to Table 6.3, the haplotypic composition of the effective pollen produced by G_t, *i.e.* $(s_{0,t+1}, s_{1,t+1})$, is equal to (q_t, p_t). The effective eggs are produced by $G_t{}'$. Their haplotypic composition, *i.e.* $(e_{0,t+1}, e_{1,t+1})$, is equal to $(q_t{}', p_t{}')$, where $q_t{}' = \frac{1}{2} f_{1,t}{}'$. The genotypic composition of G_{t+1} is $(q_1 q_t{}', q_t p_t{}' + q_t{}' p_t, p_t p_t{}')$.

Example 6.12 A trait expressed after pollen distribution and controlled by locus A-a is considered. Plants with genotype aa are eliminated after pollen distribution. The frequency of gene a in populations G_1, G_2, G_3 and G_4 is calculated for each of three values of q in the original population. This yields

	q		
G_0	0.80	0.50	0.20
G_1	0.62	0.42	0.18
G_2	0.52	0.36	0.17
G_3	0.43	0.31	0.16
G_4	0.37	0.28	0.15

According to Equation (2.2), derived for the population resulting from a bulk cross, the frequency in G_{t+1} of allele a is $q_{t+1} = \frac{1}{2}(q_t + q_t')$.

A simple formula to express q_t in terms of t and q_0 does not exist. Calculations corresponding to the selection process should thus be carried out repeatedly in order to derive q_t. Results of such calculations are given by Example 6.12.

Comparison of Examples 6.11 and 6.12 shows that, for the same value for q_0, the reduction of the frequency of the undesired allele a, $\Delta q = q_0 - q_1$, is twice as large as at mass selection after pollen distribution. For example the reduction from 0.50 to 0.33 for mass selection before pollen distribution is twice as large as that from 0.50 to 0.42 for mass selection after pollen distribution.

Generally, it may be stated that mass selection with regard to a trait expressed after pollen distribution should only be applied as long as the frequency of a is larger than $\frac{1}{2}$. For smaller values of q its reduction due to selection is too small to be of practical significance. (By the way the reduction of the frequency of allele **m**, which conditions in homozygous state male sterility, see Section 5.2.1, proceeds like the reduction of allele a under the conditions considered here.)

6.3.6 Progeny testing

With the remnant seed procedure, the genetic quality of a (parental) plant is derived from the performance of its progeny. When dealing with an annual plant species, the parent plants do not exist any more at the time when the performance of their offspring is known. The selection, on the basis of

the observed performances, is then necessarily among sibs of the evaluated progenies. With recurrent selection procedures the selection programme is continued on the basis of S1-lines representing the parent plants producing well-performing families. (A justification for this was given in Section 3.2.3, see Note 3.10.)

When, however, vegetative maintenance of the parent plants is possible, the parents might still be available after the evaluation of their progeny. In this situation it does not matter whether the trait is expressed before or after pollen distribution. The selection among the (parental) candidate plants is based on the performances of their offspring.

For many crops, vegetative maintenance after the first reproductive phase is possible. It occurs spontaneously with **perennial crops**, but it may also be imposed by applying some intervention, *e.g.* **tissue culture**. In the case of vegetative maintenance one may decide, on the basis of the performance of their offspring, which parental plants deserve to be selected. The selection is based on a **progeny test**. In animal breeding this is a frequently applied procedure. Among crops the procedure may be applied to herbaceous species (such as grasses, potato (*Solanum tuberosum* L.), asparagus), but especially to woody species, such as coconut (*Cocos nucifera* L.), oil palm (*Elaeis guineensis* Jacq.), or Robusta coffee (*Coffea canephora* Pierre ex Froener).

The offspring to be evaluated can be of different types, *viz.*

- S_1-lines
- FS-families obtained from pairwise crosses, *e.g.* in the case of a **diallel set of crosses** or when **test-crossing** candidate plants with a homozygous recessive genotype
- HS-families obtained after open pollination, possibly as part of a **polycross**

To reduce the probability of random fixation the number of progenies should be high enough to retain for continued breeding work at least about 25 parental genotypes.

S_1-*lines*

Progeny testing involving S_1-lines is a very effective procedure. It allows for easy and complete elimination of allele a, because it allows for discrimination between parental plants with genotype AA and parental plants with genotype Aa.

FS-families

FS-families are obtained by pairwise crosses between parental plants with genotype Aa or AA. On the basis of the progenies one can distinguish parental plants with genotype AA from parental plants with genotype Aa (see Example 6.13).

Example 6.13 FS-families resulting from a diallel set of crosses, excluding selfings and reciprocal crosses, may segregate (s) or may not segregate (ns) with regard to their genotype for locus *A*-*a*.

Consider the FS-families from such set of crosses involving parental plants $P_1, \ldots, P_5$, all with phenotype $A\cdot$,

♂ \ ♀	P_2	P_3	P_4	P_5
P_1	ns	ns	ns	ns
P_2		s	ns	s
P_3			ns	s
P_4				ns

If both parents are heterozygous, the involved FS-family will segregate. Thus parents P_2, P_3 and P_5 must have genotype Aa. These parents should be eliminated. Further breeding work is done with the remaining parents. (If none of the FS-families segregates, no more than one of the parents will have genotype Aa.)

Test-crossing of each of N parental plants with a plant with the recessive genotype aa is a simpler procedure for identifying parents with genotype AA among parents with phenotype $A\cdot$. Instead of $\frac{1}{2}N(N-1)$ FS-families obtained with a diallel set of crosses, only N FS-families have to be produced and evaluated. Furthermore the family size required for identification of potentially segregating families is only 7 (instead of 16).

HS-families

In the case of a **polycross**, a HS-family is harvested for each participating parental genotype, represented either by a single plant or by a clone. On the basis of an evaluation of the HS-families one can distinguish parents with genotype AA from parents with genotype Aa. Allele a can be completely eliminated by a single generation with application of progeny testing. In the case of a dioecious crop both female and male genotypes/clones should function as a polygamic parent. (Why?)

In fact polycrosses or diallel crosses are predominantly applied to determine general and specific combining ability with regard to quantitative variation. They are applied when the aim is to develop a synthetic variety or a hybrid variety. Test-crossing is mainly applied in linkage studies. Thus the procedures described in this section are hardly used in practice when the aim is to eliminate allele a. Progeny testing is, however, an important procedure for improving traits with quantitative variation, *e.g.* in oil palm.

Chapter 7
Random Variation of Allele Frequencies

A small population size is due to a small number of effective fusions between a female and a male gamete. In this case the population is based on a small sample of male and female gametes. The sampling process implies that the allele frequencies behave as random variables. The probability that the frequency of a certain allele becomes either zero or one, this is called fixation, is larger as the population size is smaller. Due to the process of sampling of a small number of gametes, the genetic diversity becomes inevitably smaller in course of the generations. The probability of gene fixation will be shown to depend on the population size and on the mode of reproduction.

7.1 Introduction

In the preceding chapters it was mostly (implicitly) assumed that the considered population consisted of infinitely large numbers of plants. In this chapter, population genetic effects of a restricted number of plants, which constitute a genetically heterogeneous population, are considered. At a small population size the allele frequencies for loci controlling traits not under selection pressure behave as random variables. This applies to all loci in the case of lines or families maintained, at a breeding institution or in a gene bank, in the absence of selection. It also applies to loci controlling traits which are not under selection pressure, and which are not linked to other loci controlling traits under selection pressure.

Random variation of the allele frequencies implies variation in the genotypic composition from one generation to the next. The smaller the population size, the higher the probability of a certain difference between the actual allele and/or genotype frequencies and their values expected when assuming that the population size is infinite (see Example 7.1 and 7.2).

In the course of the generations, the probability that the frequency of some allele of some locus assumes either the value 0 or 1, say: the probability of gene **fixation**, increases steadily. Such fixation implies loss of genetic variation. This may be conspicuous with regard to a trait with qualitative variation (*e.g.* the colour of cabbage heads), or inconspicuous with regard to a trait with quantitative variation (*e.g.* protein content of the achenes of sunflower).

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 107–117.

Example 7.1 An F_2-population consists of N plants, $\underline{n}$ of which have a homozygous genotype (aa or AA). The random variable $\underline{n}$ has a binomial probability distribution with parameters p, equal to $\frac{1}{2}$, and N. In shorthand

$$\underline{n} \simeq \underline{b}(\tfrac{1}{2}, N)$$

The expected value of $\underline{n}$ is

$$\mathrm{E}\underline{n} = \tfrac{1}{2}N$$

The probability that $\underline{n}$ deviates more than 10% from its expected value amounts to

$$P\left(\frac{|\underline{n} - \frac{1}{2}N|}{N} > 0.1\right) = 2P(\underline{n} - \tfrac{1}{2}N > 0.1N) = 2P(\underline{n} > 0.6N)$$

For $N = 10$ this amounts to 0.344 (Pearson and Hartley, 1970, Table 37). For large values of N the probability distribution for $\underline{n}$ can satisfactorily be approximated by

$$E\underline{n} + \sqrt{\tfrac{1}{2} \times \tfrac{1}{2} \times N}\underline{\chi} = \tfrac{1}{2}N + \tfrac{1}{2}\sqrt{N}\underline{\chi}$$

where $\underline{\chi}$ represents the standard normal distribution $N(0, 1)$. This implies that, for $N = 100$, the above probability can be approximated by

$$2P(\underline{n} > 60) \approx 2P(50 + 5\underline{\chi} > 59.5) = 2P(\underline{\chi} > 1.9) = 0.057$$

(Pearson and Hartley, 1970, Table 1).

The probability that the actual number of homozygous plants deviates more than 10% from its expected values is thus shown to depend strongly on the population size.

Example 7.2 Assume that seeds, obtained by harvesting a number of plants in bulk, represent a population with genotypic composition (0.1, 0.1, 0.8) for locus A-a, *i.e.* $p = 0.85$. Next season N plants are grown. These consist of $\underline{n}_0$ plants with genotype aa, $\underline{n}_1$ plants with genotype Aa and $\underline{n}_2$ plants with genotype AA. The probability distribution for $\underline{n}_0$, $\underline{n}_1$ and $\underline{n}_2$ is given by the multinomial probability distribution function:

$$P(\underline{n}_0 = n_0;\ \underline{n}_1 = n_1;\ \underline{n}_2 = n_2 | \Sigma n_i = N) = \frac{N!}{n_0! n_1! n_2!} 0.1^{n_0} 0.1^{n_1} 0.8^{n_2}$$

For $N = 10$ the probability $P(\underline{n}_0 = 1; \underline{n}_1 = 0; \underline{n}_2 = 9)$, implying $p = 0.9$, is 0.1343. The probability $P(\underline{n}_0 = 0; \underline{n}_1 = 0; \underline{n}_2 = 10)$, implying $P = 0.95$, is also 0.1343. The probability of fixation is $P(\underline{n}_0 = 0; \underline{n}_1 = 0; \underline{n}_2 = 10) + P(\underline{n}_0 = 10; \underline{n}_1 = 0; \underline{n}_2 = 0) = 0.1074$.

For $N = 100$ the probability of fixation, *i.e.* $P(\underline{n}_0 = 0; \underline{n}_1 = 0; \underline{n}_2 = 100) + P(\underline{n}_0 = 100; \underline{n}_1 = 0; \underline{n}_2 = 0)$ is only 2.04×10^{-10}, and therefore effectively nil.

A remedy to cure loss of genetic variation is re-introduction of the original plant material or partial exchanges with other collections.

Some aspects of the random variation of allele frequencies, including fixation, are now illustrated for the most simple situation, namely a population with a constant size of $N = 2$ plants. We consider $\underline{p}$, the frequency of allele A of some locus A-a. There is no selection with regard to the trait(s) affected by this locus. The probability distribution of $\underline{p}$ will be derived for successive generations. The values which may be assumed by $\underline{p}$ are 0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ or 1. Fixation implies $\underline{p} = 0$ or $\underline{p} = 1$. We consider P_f, the **probability of fixation**: $P_f = P(\underline{p} = 0) + P(\underline{p} = 1)$. It will be shown that – for the described situation – P_f increases monotonously in the course of the generations.

The probability distribution to be derived is $P(\underline{p} = p)$, where p may assume the value 0, $\frac{1}{4}$, $\frac{1}{2}$, $\frac{3}{4}$ or 1. It is derived from the probability distribution $P(\underline{k} = k)$ of $\underline{k}$, *i.e.* the number of gametes with haplotype A among the four gametes giving rise, after random fusion of these gametes, to the next generation. The probability distribution $P(\underline{k} = k)$ of $\underline{k}$, instead of the probability distribution $P(\underline{p} = p)$ of $\underline{p}$, is considered because of the relation $\underline{p} = \frac{1}{4}\underline{k}$.

It is assumed that the frequency of allele A in population G_0, *i.e.* the initial population, is equal to $\frac{1}{2}$. Thus $p_0 = q_0 = \frac{1}{2}$. The probability distribution $P(\underline{p}_1 = p_1)$ of $\underline{p}_1$, the allele frequency in population G_1, follows from the probability distribution function for $\underline{k}$, *i.e.*

$$P(\underline{k} = k) = \binom{4}{k} \left(\tfrac{1}{2}\right)^k \left(\tfrac{1}{2}\right)^{4-k} = \binom{4}{k} \left(\tfrac{1}{2}\right)^4$$

Thus

k	$P(\underline{k} = k)$	$p_1 (= \frac{1}{4}k)$	$P(\underline{p}_1 = p_1)$
0	$\frac{1}{16}$	0	$\frac{1}{16}$
1	$\frac{4}{16}$	$\frac{1}{4}$	$\frac{4}{16}$
2	$\frac{6}{16}$	$\frac{1}{2}$	$\frac{6}{16}$
3	$\frac{4}{16}$	$\frac{3}{4}$	$\frac{4}{16}$
4	$\frac{1}{16}$	1	$\frac{1}{16}$

The probability distribution of $\underline{p}_1$ is depicted in Fig. 7.1.

Because $\mathrm{E}\underline{k} = 4 \times \frac{1}{2} = 2$ it follows that $\mathrm{E}\underline{p}_1 = \frac{1}{2} = p_0$. The probability of fixation in population G_1 is

$$P_{f,1} = 2\left(\tfrac{1}{16}\right) = 0.125$$

whereas

$$P(\underline{p}_1 \neq p_0) = P(\underline{p}_1 \neq \tfrac{1}{2}) = \frac{10}{16} = 0.625.$$

The probability distribution of $\underline{p}_2$, *i.e.* the frequency of allele A in the next generation (in population G_2) depends on the value assumed in population G_1

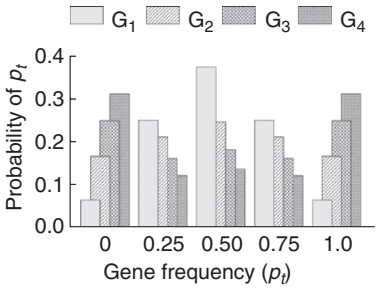


Fig. 7.1 The probability distribution of $\underline{p}_t$, the frequency of allele A in generation G_t ($t = 1, 2, 3,$ or 4) obtained by continued random mating starting in generation G_0 with allele frequency $p_0 = 0.5$. The population size is always $N = 2$ plants

by $\underline{p}_1$. Thus for each possible value for p_1 there exists a conditional probability distribution for $\underline{p}_2$, namely $P(\underline{p}_2 = p_2|p_1)$. The unconditional probability $P(\underline{p}_2 = p_2)$ is equal to the expected value of $P(\underline{p}_2 = p_2|p_1)$, calculated across all values possible for p_1. Thus

$$P(\underline{p}_2 = p_2) = \sum_{\forall p_1} P(\underline{p}_2 = p_2|p_1) \cdot P(\underline{p}_1 = p_1)$$

Because $\underline{p}_2 = \frac{1}{4}\underline{k}$, the probability distribution $P(\underline{p}_2 = p_2|p_1)$ is identical to the probability distribution $P(\underline{k} = k|p_1)$. Thus we calculate

$$P(\underline{k} = k) = \sum_{\forall p_1} \left[\binom{4}{k} {p_1}^k (1 - p_1)^{4-k} \cdot P(\underline{p}_1 = p_1) \right]$$

Each possible value for k implies a specific value for p_2. Thus, for each possible value for k, the above sum of products can be calculated as the matrix product of two vectors, *viz.* a row vector, consisting of the probabilities $\binom{4}{k} {p_1}^k (1 - p_1)^{4-k}$ as calculated for each of the five possible values for p_1, and a column vector, say $\mathbf{P}_1$, presenting the probability distribution $P(\underline{p}_1 = p_1)$ for each possible value for p_1.

For example, for $k = 0$, which implies $\underline{p}_2 = 0$, the appropriate row vector is

$$\binom{4}{0}\left(\frac{0}{4}\right)^0\left(\frac{4}{4}\right)^4; \binom{4}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^4; \binom{4}{0}\left(\frac{2}{4}\right)^0\left(\frac{2}{4}\right)^4;$$

$$\binom{4}{0}\left(\frac{3}{4}\right)^0\left(\frac{1}{4}\right)^4; \binom{4}{0}\left(\frac{4}{4}\right)^0\left(\frac{0}{4}\right)^4$$

i.e.

$$\left(1; \tfrac{81}{256}; \tfrac{16}{256}; \tfrac{1}{256}; 0\right)$$

Likewise one gets for $k = 2$ the following row vector

$$\binom{4}{2}\left(\frac{0}{4}\right)^2\left(\frac{4}{4}\right)^2;\binom{4}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^2;\binom{4}{2}\left(\frac{2}{4}\right)^2\left(\frac{2}{4}\right)^2;$$

$$\binom{4}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^2;\binom{4}{2}\left(\frac{4}{4}\right)^2\left(\frac{0}{4}\right)^2$$

i.e.

$$\left(0; \tfrac{54}{256}; \tfrac{96}{256}; \tfrac{54}{256}; 0\right)$$

The five row vectors constitute the so-called **transition matrix T**, *i.e.*

$$\begin{pmatrix} 1 & \frac{81}{256} & \frac{16}{256} & \frac{1}{256} & 0 \\ 0 & \frac{108}{256} & \frac{64}{256} & \frac{12}{256} & 0 \\ 0 & \frac{54}{256} & \frac{96}{256} & \frac{54}{256} & 0 \\ 0 & \frac{12}{256} & \frac{64}{256} & \frac{108}{256} & 0 \\ 0 & \frac{1}{256} & \frac{16}{256} & \frac{81}{256} & 1 \end{pmatrix}$$

The probability distribution $P(\underline{p}_2 = p_2)$, represented by the column vector $\mathbf{P_2}$, is obtained by multiplying $\mathbf{T}$ and the column vector $\mathbf{P_1}$:

$$\mathbf{P_2} = \mathbf{TP_1}$$

Likewise

$$\mathbf{P_3} = \mathbf{TP_2} = \mathbf{TTP_1}$$

N.B. Even $\mathbf{P_1}$ may be calculated from $\mathbf{P_1} = \mathbf{TP_0}$, where $\mathbf{P_0}' = (0, 0, 1, 0, 0)$. The probability that $\underline{p}_2$ is 0, *i.e.* $P(\underline{p}_2 = 0)$, is equal to the matrix product of the first row of $\mathbf{T}$ and the column vector $\mathbf{P_1}$:

$$\begin{aligned}\left(1\ \tfrac{81}{256}\ \tfrac{16}{256}\ \tfrac{1}{256}\ 0\right) \cdot \boldsymbol{P_1} &= \left(1 \times \tfrac{1}{16} + \tfrac{81}{256} \times \tfrac{4}{16} + \tfrac{16}{256} \times \tfrac{6}{16} + \tfrac{1}{256} \times \tfrac{4}{16}\right) \\ &= 0.1660\end{aligned}$$

Altogether the following probability distributions $P(\underline{p} = p)$ can be derived for the successive generations G_1, G_2, G_3 and G_4:

	p					
	0	1/4	1/2	3/4	1	P_f
G_1	0.0625	0.2500	0.3750	0.2500	0.0625	0.1250
G_2	0.1660	0.2109	0.2461	0.2109	0.1660	0.3320
G_3	0.2489	0.1604	0.1812	0.1604	0.2489	0.4978
G_4	0.3116	0.1205	0.1356	0.1205	0.3116	0.6232

Fig. 7.1 presents these probability distributions graphically.

For all generations $E\underline{p}_t = p_0 = \frac{1}{2}$. It appears that P_f, the probability of fixation, increases continuously. The probability that fixation has not yet

occurred, *i.e.* $P_{nf} = 1 - P_f$, amounts in these first four generations to 0.875, 0.668, 0.502 and 0.377 respectively. It decreases continuously. This decrease is further considered. To measure it, the parameter ψ is defined:

$$\psi = \frac{P_{nf,t}}{P_{nf,t-1}} = \frac{1 - P_{f,t}}{1 - P_{f,t-1}} \tag{7.1}$$

The parameter ψ indicates the value of P_{nf} relative to its value in the preceding generation. For the considered generations of the elaborated situation it assumes the following values:

$$\frac{0.668}{0.875} = 0.7634;\ \frac{0.502}{0.668} = 0.7515 : \frac{0.377}{0.502} = 0.7510$$

These values converge to 0.75.

It can be shown (see *e.g.* Li (1976, pp. 552–557)) that ψ converges to the appropriate value for

$$1 - \frac{1}{2N} \tag{7.2}$$

In the words of Li (1976, p. 552) the parameter ψ measures 'the decay of variability'. This decay is small for values near to 1. In Note 7.1 the loss of genetic variation due to random variation of the allele frequencies is compared with the reduction of the frequency of heterozygous plants due to inbreeding.

Note 7.1 The parameter ψ is similar to the parameter λ representing the frequency of heterozygous plants relative to this frequency in the preceding generation, see Equation (3.3). A population size of $N = 1$ implies necessarily selfing. In the case of continued selfing the expected number of loci with a heterozygous single-locus genotype measure is halved each generation (Section 3.2.1). Indeed, at this population size the probability that fixation with regard to a certain locus has not yet occurred is halved each generation.

The stable value of ψ is thus given by

$$\psi = \frac{P_{nf,t}}{P_{nf,t-1}} = 1 - \frac{1}{2N} \tag{7.3}$$

Equation (7.3) yields for the elaborated example $1 - \frac{1}{4} = \frac{3}{4}$. This value is already closely approximated by the ratio of the P_{nf} values for generations G_4 and G_3. The part of $P_{nf,t-1}$ which applies to generation G_t is $\left(1 - \frac{1}{2N}\right)$. Thus

$$P_{nf,t} = \left(1 - \frac{1}{2N}\right) P_{nf,t-1} = P_{nf,t-1} - \frac{1}{2N} \cdot P_{nf,t-1} \tag{7.4}$$

implying

$$1 - P_{f,t} = (1 - P_{f,t-1}) - \frac{1}{2N} \cdot (1 - P_{f,t-1})$$

or

$$P_{f,t} - P_{f,t-1} = \frac{1}{2N} \cdot (1 - P_{f,t-1}) = \frac{1}{2N} \cdot P_{nf,t-1} \tag{7.5}$$

For a population consisting out of $N = 2$ plants, the random variation of the allele frequencies might imply that the frequencies of some allele A amount in

successive generations to $p_0 = \frac{1}{2}$, $p_1 = \frac{1}{4}$, $p_2 = \frac{1}{2}$, $p_3 = \frac{1}{2}$, $p_4 = p_5 = p_6 = \ldots = p_\infty = 1$. The fixation occurring from generation 3 to 4 means that from then onward the genetic variation for this locus is lost. Indeed, in populations consisting of a restricted number of plants the allele frequencies vary from one generation to the next until fixation occurs. The random variation of the allele frequencies is called **random genetic drift**.

P_f increases steadily. This implies that loss of alleles, belonging to loci controlling traits that are not subject to selection, is inevitable. The expected number of generations until fixation occurs is considered in Note 7.2.

Note 7.2 If a population with initial allele frequencies (p_0, q_0) is reproduced generation after generation on the basis of N plants, the expected number of generations until fixation occurs is

$$T = -4\mathrm{N}[p_0 \ln(p_0) + q_0 \ln(q_0)]$$

(Ewens, 1969, p. 58). This expression attains a maximum value at $q_0 = p_0 = \frac{1}{2}$. Then $T = -4N \ln(\frac{1}{2}) = 2.77N$; *i.e.* 5.5 generations for $N = 2$ and 27.7 generations for $N = 10$. For $q_0 = 0.95$ the formula yields $T = 0.79N$ and for $q_0 = 0.995$ it yields $T = 0.126N$. For this last situation fixation is expected to occur in one generation in a population with size $N = 8$.

The population becomes thus genetically uniform (in homozygous condition!) for an ever increasing number of loci. Notwithstanding the presence of random mating the population genetic, and consequently the quantitative genetic, effect is the same as the effect of continued inbreeding. A population consisting of a small number plants will thus 'suffer' from the small population size. This applies especially to traits with quantitative variation: the mean value for the considered trait will change in a way similar to that occurring with continued inbreeding (see Example 7.3).

When the population size varies from one generation to the next, the ratio of the probabilities that fixation has not yet occurred in the considered populations of generations t and $t-1$ may be rewritten as $P_{nf,t} = \psi_t P_{nf,t-1}$, where

Example 7.3 Omolo and Russell (1971) checked whether the maize variety 'Krug' could be maintained by means of open pollination of a population consisting of fewer than the usual number of 500 plants. They compared the kernel yield of populations maintained from 1962 up to 1966 on the basis of 500, 200, 80, 32 or 13 plants. In 1967 seed multiplication on the basis of 150 plants occurred, followed in 1968 by a yield trial. The results are presented in Table 7.1.

It appears that loss of genetic diversity, *i.e.* fixation of random alleles, caused a non-negligible yield reduction.

Table 7.1 The reduction of kernel yield occurring when maintaining the maize variety Krug by means of open pollination of N plants in the growing seasons of 1962 up to 1966, followed by multiplication in 1967 on the basis of 150 plants. (source: Omolo and Russell, 1971)

Maintenance population size	Kernel yield (kg/ha)	Reduction of kernel yield (kg/ha)
$\propto$ (check)	5350	
500	5150	200
200	5020	330
80	4290	1060
32	3970	1380
13	4330	1020

$$\psi_t = \frac{P_{nf,t}}{P_{nf,t-1}} = 1 - \frac{1}{2N_t}$$

The probability that fixation has not yet occurred across T generations can then be calculated according to

$$\Psi = \prod_{t=1}^{T} \psi_t = \prod_{t=1}^{T} \left(1 - \frac{1}{2N_t}\right)$$

If for each generation the population size is such that $\psi_t \approx 1$, then also $\Psi \approx 1$. However, if $\psi_t \approx 0$ for at least one generation/population then also $\Psi \approx 0$. This implies that continued maintenance, intended to occur on the basis of many plants but failing at least once, leads to a drastic decrease of P_{nf}: smaller population sizes are the most critical ones with regard to the decrease of P_{nf} (see Example 7.4).

Example 7.4 For three successive generations the sizes of some population are $N_1 = 500$, $N_2 = 6$ and $N_3 = 500$. Thus

$$\Psi = \left(1 - \frac{1}{1000}\right)\left(1 - \frac{1}{12}\right)\left(1 - \frac{1}{1000}\right) = 0.9148$$

This path-way of maintenance yields the same decrease of P_{nf} as three successive generations consisting of 17.1 plants, *viz.*

$$\left(1 - \frac{1}{34.2}\right)^3 = 0.9148.$$

Thus one may say that the **effective population size** amounts to 17.1 plants.

For the study described in Example 7.3 the decrease of P_{nf} between 1961 and 1968 can be derived from

$$\Psi = \left(1 - \frac{1}{64}\right)^5 \left(1 - \frac{1}{300}\right) = 0.9212$$

Smaller population sizes are the most critical ones with regard to the decrease of P_{nf}.

7.2 The Effect of the Mode of Reproduction on the Probability of Fixation

The effect of the mode of reproduction on the probability of fixation is illustrated in Example 7.5.

Example 7.5 The probability of fixation, P_f, is considered for three different modes of reproduction of a population consisting of four plants. The considered population is assumed to consist of four plants, *viz.* one plant with genotype aa, two plants with genotype Aa and one plant with genotype AA. The genotypic composition of the next generation is then expected to be

		Genotype		
		aa	Aa	AA
f:	After selfing	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{3}{8}$
	After panmixis	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	After outbreeding:	$\frac{5}{24}$	$\frac{14}{24}$	$\frac{5}{24}$

In accordance with Section 3.1 outbreeding is here assumed to imply random interplant pollination where self-fertilization is excluded (as in self-incompatible cross-fertilizing crops). Check for yourself that the foregoing genotypic compositions are indeed to be expected at the described situation).

The probability of fixation due to the small population size amounts to $2\left(\frac{3}{8}\right)^4 = 0.0396$ after selfing, to $2\left(\frac{1}{4}\right)^4 = 0.0078$ after panmixis and to $2\left(\frac{5}{24}\right)^4 = 0.0038$ after outbreeding. This shows that P_f depends clearly on the mode of reproduction. For outbreeding it is minimal.

According to Equation (7.5) the increase of P_f is a simple function of N. A more general expression is

$$P_{f,t} - P_{f,t-1} = \frac{1}{2N_e} P_{nf,t-1} \tag{7.6}$$

where N_e is the **effective population size**, *i.e.* the effective number of reproducing plants. The latter quantity is calculated from the actual number of reproducing plants. It is the number such that the increase of P_f calculated on the basis of Equation (7.6) is equal to the increase of P_f calculated from the actual numbers of plants. In Example 7.4 it is, for instance, shown that successive population sizes of 500, 6 and 500 plants yield the same increase of P_f as three generations with a constant (effective) size of 17.1 plants.

Li (1976, pp. 559–562) presents for diverse situations formulae for calculating N_e from the actual number(s) of plants. Three situations are considered:

- Random mating:
$$N_e = N \tag{7.7}$$
- Random mating where each parental plants contributes two gametes to constitute the next generation:
$$N_e = 2N - 1 \tag{7.8}$$
- Dioecy, where N_f represents the number of female parents and N_m the number of male parents:
$$N_e = \frac{4N_f N_m}{N_f + N_m} \tag{7.9}$$

Example 7.6 considers the maximum value of N_e for a given total number of female and male plants.

Example 7.6 Equation (7.9) applies to dioecious crops, maintained on the basis of $N = N_f + N_m$ plants. As $N_f = N - N_m$, the maximum value for N_e can be calculated by determining the derivative of N_e to N_m:

$$\frac{\mathrm{d}}{\mathrm{d}N_m}\left(\frac{4N_m(N - N_m)}{N}\right) = \frac{4N - 8N_m}{N} = 4 - \frac{8N_m}{N}$$

The second derivative of N_e to N_m is negative (it is $\frac{-8}{N}$). Thus N_e is maximal for $N_m = \frac{1}{2}N = N_f$, which yields $N_e = N$. For $N_m = 5$ and $N_f = 25$ Equation (7.9) yields $N_e = 16.7$, whereas the same population size with $N_m = N_f = 15$ yields $N_e = 30$.

It is generally desired that N_e is not less than about 30 to 50: for $N_e = 30$, Equation (7.3) yields $\psi = 0.9833$; for $N_e = 50$ it yields $\psi = 0.99$. An effective population size of less than 30 plants is considered too small: *e.g.* $N_e = 10$ yields $\psi = 0.95$. These minimal values for N_e are primarily based on the consideration that the accumulated reduction of P_{nf}, due to continued maintenance of a population with a small population size, should be restricted. The minimum does not assure complete absence of 'damage' (Example 7.3).

Equation (7.9) may also be applied to situations other than dioecy. In the case of HS-family selection a selected family may consist of n plants. These

descend from $N_f = 1$ maternal parent and N_m paternal parents, where N_m is unknown. Thus

$$N_e = \frac{4N_m}{N_m + 1}$$

For $N_m = 1$ we get $N_e = 2$, and for $N_m \to \infty$ we get $N_e = 4$. (In fact $1 \leq N_m \leq \min(n, N)$.) The effective number of parents of a single HS-family is thus at least two and at most four.

With regard to the possibility of fixation of alleles of loci controlling traits not subjected to selection, one should, in the case of family selection, select such numbers of families that the value of N_e is acceptable. This should be reconciled with the wish to apply the highest possible intensity of selection. The problems involved when searching a compromise have been considered by Vencovsky and Godoi (1976).

When applying continued family selection, one should realize that the effective number of ancestors may be smaller than supposed. Thus 100 families in generation t may descend from 100 plants belonging to only 25 families in generation $t - 1$. These 25 families may have been obtained from 25 plants belonging to only 10 families in generation $t - 2$; *etc.* It will be clear that such a pedigree may lead to strong shifts in the allele frequencies of loci controlling traits that are not under conscious selective pressure. The associated probability of fixation tends to be higher in the case of family selection than in the case of mass selection. Further, it will tend to be higher when selecting among families which are evaluated in reproductive isolation, than when selecting among non-separated families. It will also be higher when selecting before pollen distribution than when selecting after pollen distribution. The effective number of parents, grandparents, great grandparents, *etc.* of the plants occurring in some population is generally unknown. It depends on the previous breeding history:

- Presence or absence of selection
- Presence or absence of a few widely diverging pedigrees originating from successful ancestors (combined with the extinction of other pedigrees)
- Selection before or after pollen distribution
- Presence or absence of separation of the families

All this inhibits expression of the reduction of P_{nf} in exact and simple formulae. One should, nevertheless, be aware of the process of a gradual loss of genetic diversity. This applies not only to continued maintenance of entries belonging to a collection of accessions of a cross-fertilizing crop, but also to the long-term maintenance of landraces of self-fertilizing crops.

Chapter 8
Components of the Phenotypic Value of Traits with Quantitative Variation

Many of the important traits of horticultural or agricultural crops display quantitative variation. The phenotypic values observed for such a trait tend to depend both on the quality of the growing conditions as well as on the (complex) genotype with regard to loci affecting the trait. The goal of horticulturists and agronomists is the manipulation of the growing conditions in such a way that the performance of the crop better obeys the goals of the growers and consumers. The goal of breeders is improvement, by means of selection, of the (average) genotypic value concerning the trait. For breeders it is, therefore, important to have some understanding of the degree in which the phenotypic expression of traits with quantitative variation is due to the genetic make-up. Breeders should select the candidates with the most attractive genotypic values, not those with the most attractive phenotypic values. The partitioning of the phenotypic values of the candidates into components, including components of the genotypic value, is therefore a topic to be considered seriously.

8.1 Introduction

In the context of this book, genetic variation with regard to a certain trait is of prime interest, both with regard to genetic analysis or in plant breeding. The variation may be such that only two distinct phenotypic classes occur, *e.g.* male plants versus female plants. Otherwise it may also be such that one can easily distinguish several different levels of expression, *e.g.* for the number of ears produced by different wheat plants (this is called quasi-continuous variation). In this chapter attention is mainly given to traits with a truly **continuous variation** of expression, *e.g.* for the grain yield of separate wheat plants or for the length of their longest culm.

A characteristic feature of a trait showing **quantitative variation** is the great range in expression. Even in absence of genetic variation, like in a clone, a pure line or an F_1-hybrid, there is a wide range of phenotypic values. In a genetically heterogeneous population, the variation is such that it is impossible to classify plants according to their genotype simply on the basis of their phenotypic values.

With regard to traits with **qualitative variation** the former is reasonably possible (however, dominance is a disturbing factor). This allows determination of the frequency of plants with a certain genotype. Classification of plants (and counting the number of plants in each class) is often applied with regard to traits like flower colour (white or blue in flax) or with regard to

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 119–172.

the presence or absence of a band at a certain position (in a lane of bands in a gel characterizing an individual plant). In the genetic analysis of such traits one studies segregation data, *i.e.* the numbers of plants in the various discrete phenotypic classes. The expression of traits with qualitative variation is mainly controlled by so-called **major genes**.

N.B. The locus controlling presence or absence of a band at a certain position in a lane of bands is responsible for a qualitative trait. If different bar codes, *i.e.* different patterns of bands being present or absent, can be shown to be associated with different levels of expression of a trait with quantitative variation one may call the polymorphism (a certain band is present or absent) a marker. Such an association is due to linkage of the locus controlling the **marker** phenotypes, *i.e.* presence or absence of a band at a certain position in the lane of bands, with one or more loci affecting the trait with the quantitative variation. Because marker assisted selection is based on such associations, the phenomenon of linkage is given proper attention in this book; notwithstanding the 'proof' (see Chapter 1) that linkage plays a minor role in the inheritance of polygenic traits.

Quantitative variation is due to two causes, which may act simultaneously:

1. Variation in the quality of the growing conditions and
2. Genetic variation

Variation in the quality of the growing conditions

Whenever the genotype only partly controls the phenotypic expression, variation in the quality of the growing conditions induces variation in phenotypic expression. The size of the phenotypic variation within genetically homogeneous plant material reflects the balance between the strength of the genetic control of the expression and the size of the effects of variation in the quality of the growing conditions. Different genotypes may, with the same variation in the quality of the growing conditions, show different phenotypic variation (see Example 8.9).

Genetic variation

The expression of traits with quantitative variation can be affected genetically by a large number of loci. Within a common genetic background, different single-locus genotypes may give rise to small differences in expression, but differences in expression of different complex genotypes, *i.e.* the aggregate genotype with regard to all relevant **polygenic loci** together, may be large. (In recent years the term **quantitative trait loci** (QTL) (Thoday, 1976) has become popular). Not all quantitative variation is due to many loci. For example, a yield component like number of seeds per plant may be expected to be affected by a smaller number of loci than grain yield itself.

In Chapter 1 it was emphasized that characters can show qualitative variation as well as quantitative variation. Quantitative variation is often expressed

for characters of great biological and economic importance. Some examples include

1. Plant height: tallness is desired in flax (*Linum usitatissimum* L.); a reduced height is desired in cereals such as rye, wheat and rice (*Oryza sativa* L.).
2. Yield of some chemical compound (per plant or per unit area): sugar, oil, protein, lysine, vitamins, drugs.
3. Yield of some botanical component
   - Dry seeds (in cereals, bean, oil flax)
   - Fresh fruits (apple (*Malus* spp.), peach (*Prunus persica* L.), strawberry (*Fragaria ananassa* Duch.), tomato (*Lycopersicon esculentum* Mill.), paprika (*Capsicum annuum* L.), pumpkin (*Cucurbita maxima* Duch. ex Lam))
   - Tubers (potato (*Solanum tuberosum* L.), sweet potato (*Ipomoea batatas* (L.) Lam.))
   - Roots (carrots (*Daucus carota* L.)).
     The yield of seeds, fruits and tubers reflects the fertility component of fitness (Section 6.1). Indeed, fitness is an important quantitative trait.
4. Yield of (nearly) the whole plant: timber, silage maize, forage grasses.
5. Earliness, *i.e.* date of flowering or date of maturity. Some national lists of varieties classify varieties according to their earliness (for example potato, maize, Brussels sprouts, radish (*Raphanus sativus* L.)).
6. Partial resistance against diseases or pests or tolerance against stress (drought, heat, frost).

Quantitative genetic theory (or biometrical genetics) aims to describe the inheritance of quantitative variation by means of as few parameters as possible. The items of interest are the **effects** of genotypes. Thus we may distinguish the population genetical effect of inbreeding, *viz.* reduction of the frequency of heterozygous plants, from its possible quantitative genetic effect, *i.e.* the phenotypic expression of plants with a more homozygous genotype.

The basis for quantitative genetic theory, aiming to describe the inheritance of quantitative characters by the smallest acceptable number of parameters, has been laid by Fisher (1918), Wright (1921) and Haldane (1932). They defined important parameters, such as additive genetic effect, degree of dominance and genetic correlation. Procedures to estimate these parameters for certain traits of certain crops (and their actual estimates) followed later. The founders of this work were, in animal breeding Lush (1945), Lerner (1950, 1958) and Henderson (1953) and, in plant breeding, Comstock and Robinson (1948), Mather (1949), Hayman (1954), Jinks (1954), Griffing (1956) and Finlay and Wilkinson (1963).

Quantitative genetic theory is based on the effects of so-called **Mendelian genes**, *i.e.* genes located on the chromosomes. It dates, therefore, from after the appreciation (since 1900) of Mendel's explanation of the inheritance of qualitative variation for a number of traits in peas. Before 1900 there was

already extensive research into the inheritance of traits with quantitative variation. Notably Galton, a cousin of Charles Darwin, and Pearson tried to gain understanding by comparing parents and their offspring. They established that tall fathers tend to produce sons who are indeed tall, but generally not as tall as their fathers. This phenomenon was called **regression**, a term that nowadays occupies a central position in statistics. Around 1910 the Mendelian basis of quantitative characters had already been shown. The study of Nilsson-Ehle (1909) is well known, he explained variation, *i.e.* segregation, for kernel colour of wheat and oats on the basis of three polygenic loci. Other classical studies are those by East (1910, 1916) on the inheritance of the corolla length of flowers of *Nicotiana longiflora* Cav.

Manuals that contributed greatly to the spreading of knowledge of quantitative genetic theory are those by Falconer (1989) or Falconer and MacKay (1996), with an emphasize on cross-fertilizing species (domesticated animals), and Mather and Jinks (1977, 1982) or Kearsey and Pooni (1996), emphasizing self-fertilizing crops.

Continuous variation occurs despite the fact that genetic information is transmitted by means of discrete units, the genes. This continuous variation is due to the overlap of the frequency distributions of the phenotypic values for different genotypes. Nilsson-Ehle (1909) was able, through careful observation, to associate very narrow ranges of expression for the intensity of grain colour of wheat with certain genotypes (at superficial observation continuous variation seemed to exist).

Figure 8.1 illustrates how observations for some trait, for each of the three genotypes for locus *B-b* affecting the trait, could be distributed in a sample taken from an F_2-population. Compared to the genetic variation, there is a small effect of variation in growing conditions. On the basis of the phenotypic

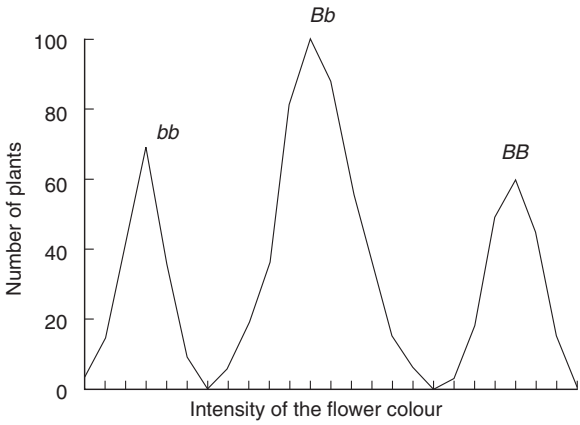


Fig. 8.1 The numbers of plants, in an F_2-population, with specified intensities of the colour of the flowers. The population segregates for locus *B-b* affecting flower colour intensity. The ranges of the phenotypic values for the three genotypes *bb*, *Bb* and *BB* just fail to overlap

value of a plant one can correctly assign a genotype to it. Locus *B*-*b* controls, in this case, qualitative variation. The genetic control of the trait can then be understood from the segregation ratio.

One can also use a statistical tool to determine whether or not a trait with quantitative variation is affected by a locus with major genes. In the latter case the locus induces the frequency distribution to be multimodal. A locus with major genes is then indicated if the null hypothesis assuming a unimodal distribution, *i.e.* H_0: 'no major genes segregating', is rejected when tested against the alternative hypothesis H_a: 'major genes segregating' (see Schut, 1998).

The mere demonstration of the presence of a locus with major gene effects does, of course, not indicate the identity of the locus. It is, however, possible to identify an individual locus affecting the phenotypic values for a trait with quantitative variation by means of molecular markers. In that context such loci are often designated as QTLs (quantitative trait loci) rather than as polygenes. QTLs may not just be identified, their effects can also be ascertained (see Section 12.3.1, dealing with marker-assisted selection). All this might imply that the distinction between loci with major genes and polygenic loci (or the corresponding distinction between traits with quantitative variation and traits with qualitative variation) will become outdated.

If the effect of variation in growing conditions is large compared to the effect of genetic variation, the ranges of expression for plants with genotype *bb* or *Bb* or *BB* overlap (Fig. 8.2). Then it is impossible to assign unambiguously a genotype to each plant on the basis of its phenotypic value. Segregation ratios cannot be established. This complicates the elucidation

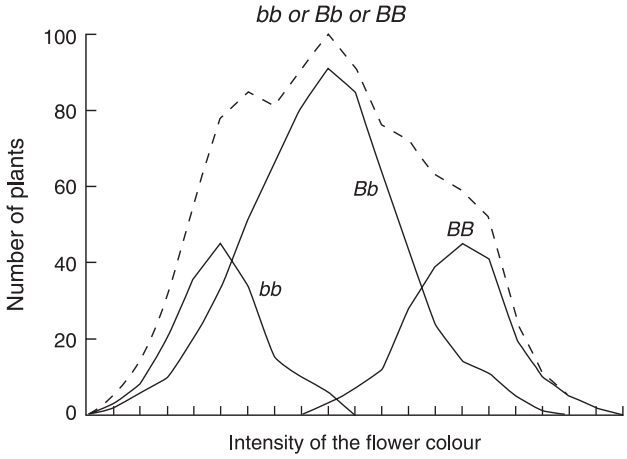


Fig. 8.2 The numbers of plants, in an F_2-population, with specified intensities of the colour of the flowers. The population segregates for locus *B*-*b* affecting flower colour intensity. The ranges for the phenotypic values for the three genotypes *bb*, *Bb* and *BB* overlap to a great extent

of the genetic control underlying quantitative variation. Quantitative genetic analysis consists, in this case, of interpreting estimates of statistical parameters in quantitative genetical terms. This is based on population genetic assumptions and inferences:

(a) If the mean phenotypic value of the offspring of parents P_1 and P_2 does not differ significantly from the mid-parent phenotypic value, the genetic control of the involved trait is assumed to be additive (see Example 9.2 for details).
(b) The estimate of the regression of HS-family mean phenotypic values on their maternal plant phenotypic values is taken to be an estimate of the heritability in the narrow sense of the considered trait (see Section 11.2.2 for details).

The shape of the frequency distribution of the phenotypic values for a trait with quantitative variation tends often towards the shape of a normal distribution (see Fig. 8.2). This is mainly due to a normal distribution of the contributions of the environmental conditions to the phenotypic value. In genetically homogeneous plant material a normal distribution is entirely due to a normal distribution of the environmental conditions. Examples 8.13 and 8.15 show that segregating populations may also tend to show a normal distribution for phenotypic values in the absence of variation of environmental conditions.

The size of the phenotypic (or genotypic) quantitative variation may be measured by different yardsticks:

1. The **range**, *i.e.* the absolute value of the difference between the lowest (smallest) and the highest (largest) phenotypic value encountered.
   This yardstick should only be used as a rough descriptor of the variation because the value obtained for the range depends on the sample size.
2. The **standard deviation** or its square, the **variance**.
   These two popular yardsticks are scale dependent and should thus always be used with an indication of the scale of measurement. For example, when expressed as standard deviation the variation of plant height measured in centimetres is 2.54 times as high as when measured in inches; when expressed as variance this factor is $(2.54)^2 = 6.4516$.
3. The **coefficient of phenotypic variation** ($\boldsymbol{\nu c_p}$), *i.e.* the ratio of the standard deviation of the phenotypic values (σ_p) and its expectation ($\mathrm{E}\underline{p}$) of the phenotypic values; thus: $\boldsymbol{\nu c_p} := \frac{\boldsymbol{\sigma_p}}{\boldsymbol{E_p}}$. This yardstick is scale independent.
   It allows a meaningful comparison of the variation of several traits of plants belonging to the same population, as well as a comparison of the variation for the same trait as expressed by different populations (of the same or different crops). This is illustrated in Example 8.1.

The size of the phenotypic variation for a character displaying quantitative variation depends on:

Example 8.1 Table 8.1 presents the range for culm length, *i.e.* plant height, for the genetically homogeneous spring wheat variety Peko, as well as for two genetically heterogeneous populations of winter rye.

Table 8.1 Mean phenotypic value ($\bar{p}$) and range of phenotypic values (w) for culm length and grain yield of plants belonging to the pure-line spring wheat variety Peko (data of Wageningen, The Netherlands, 1971; plants grown in a $15 \times 25\,\text{cm}^2$ rectangular pattern of plant positions) and of diploid and tetraploid winter rye plants (data of Wageningen, growing season 1977–1978; plants grown in a regular triangular pattern of plant positions with an interplant distance of 15 cm) N: sample size

	Culm length (cm)			Grain yield (decigram)		
	N	$\bar{p}$	w	N	$\bar{p}$	w
spring wheat:	1,099	93.4	43			
winter rye: $2n = 2x$:	5,111	158.8	143	5,107	102.2	315
$2n = 4x$:	4,473	179.7	164	4,471	89.9	345

Table 8.2 presents, for the same plant material, as well as a maize population, estimates of the phenotypic variance and the coefficient of phenotypic variation.

Table 8.2 Estimated variance (s^2) and coefficient of phenotypic variation ($\nu\hat{c}_p$) for plant height, grain yield and length and area of the fourth leaf from the top of spring wheat (Table 8.1), diploid and tetraploid winter rye (Table 8.1) and maize plants (data from Wageningen, The Netherlands, 1973; 1049 plants grown in a 40×67.5 cm rectangular pattern of plant positions)

					Fourth leaf from the top			
	Plant height (cm)		Grain yield (g)		Length (cm)		Area (cm^2)	
	s^2	$\nu\hat{c}_p$	s^2	$\nu\hat{c}_p$	s^2	$\nu\hat{c}_p$	s^2	$\nu\hat{c}_p$
Spring wheat	36	0.06						
Winter rye: $2n = 2x$	156.3	0.08	1,296	0.35				
$2n = 4x$	372.5	0.11	3,249	0.64				
Maize:	285.6	0.12	252,000	0.47	42.3	0.09	8,208	0.17

One may conclude that within the populations the variation for grain yield is higher than that for plant height. The variation for plant height in the maize population appeared to be twice as large in the maize population as in the pure line spring wheat variety.

1. The particular crop and the trait under consideration.
   The size of the phenotypic variation may also be associated with the level of expression of the trait. Thus the variation in flowering date of an early flowering pure line may tend to be smaller than the variation in a late flowering date of an early flowering pure line may tend to be smaller than the variation in a late flowering line. The phenomenon is also illustrated by Example 8.9: short pure lines of maize tend to have a smaller phenotypic variation for plant height than tall single cross hybrid varieties.

2. The size of the genetic variation.
   It may seem a paradox but this variation depends on the environmental conditions. The effect of plant density on the genetic variance is illustrated in Example 8.8.
3. The size of the variation in growing conditions.
   Early in this section it was already indicated that different genotypes may differ in their responses to variation in growing conditions. The latter variation is, nevertheless, mostly measured by the phenotypic variation, for the trait of interest, among the plants constituting a genetically homogeneous population. It is only rarely measured directly by measuring the variation for physical growth factors, *e.g.* soil temperature or oxygen content of the soil.

In this book attention is focussed on

- The mean genotypic value, designated by $E\underline{\mathcal{G}}$ or by μ_g
- The genetic variance, designated by $\text{var}(\underline{\mathcal{G}})$ or by ${\sigma_g}^2$.

Breeders manipulate these parameters in such a way that the mean/expected genotypic value is changed in the desired direction. The manipulation may involve the mode of reproduction, especially when producing hybrid varieties by crossing pure lines. The large influence of the inbreeding coefficient will appear. When applying selection the genetic variance is exploited, in fact it is reduced, in order to attain the breeding goal.

In the case of a normal distribution of the genotypic values this distribution is completely specified by the parameters μ_g and σ_g. If accurate estimates of these parameters are available, one can derive properties of the population for the trait under study (see, for example, Section 11.1 with regard to selection intensity). Section 8.3.2 provides a genetic explanation for the occurrence of the frequently encountered (approximately) normal distribution.

Normality of the observed distribution does not necessarily imply the presence of many segregating loci. Even in the absence of variation in growing conditions, it is, even for three or four segregating loci, already necessary that a rather large number of plants are observed in order to prove the significance of departures from normality. According to Thoday and Thompson (1976) the sample size required would amount to 500 to 1,000 plants.

Instead of the symmetric shape of the normal distribution of the phenotypic values, one may observe an asymmetric, skew distribution. Indeed, for traits such as date of flowering or yield, a deviation from normality is often observed. For date of flowering this may be due to variation in the daily temperatures. The distribution for yield often shows positive skewness, which, according to Spitters (1979, p. 91) is due to interplant competition. In the absence of competition, *i.e.* at a very low plant density, the distribution is normal or practically normal.

In the case of negative skewness there is a long tail at the left-hand side of the distribution (see Example 8.14). Then the expected phenotypic value is

smaller than the **median** phenotypic value, *i.e.* the value such that 50% of the observed phenotypic values is smaller than this value and 50% is larger. With positive skewness there is a long tail at the right. Then the expected phenotypic value is larger than the median. For asymmetric distributions the median is often preferred as a measure for the central value, because in contrast to the expectation the median is not affected by outliers.

The skewness of the distribution of grain yield of individual plants of small cereals grown at high plant density follows from the strong correlation between grain yield and number of ears (this correlation was estimated to be 0.90 for winter rye, grown at the rather low plant density of 51.3 plants/m^2 (Bos, 1981, p. 16)). At high plant density the values tend to have a Poisson distribution. The positive skewness can often be eliminated by some **transformation**, *e.g.* a **logarithmic transformation** or the **square root transformation**.

As general features of traits with quantitative variation we may note:

1. Presence of continuous phenotypic variation.
   This may be due to continuous variation in the quality of the growing conditions.
2. An approximate normal distribution.
   This can be explained from a polygenic genetic basis (Section 8.3.2), and/or a normal probability distribution of the quality of the growing conditions.
3. Occurrence of inbreeding depression at a positive value of $\mathscr{F}$ (inbreeding coefficient) and of heterosis at $\mathscr{F} < 0$.
   Especially in cross-fertilizing crops the mean phenotypic value of most quantitative traits is negatively affected by inbreeding and positively by outbreeding.
4. The phenotypic values for different quantitatively varying traits are correlated.
   This is discussed and illustrated in Example 8.2. The correlation implies that selection with regard to one trait may give rise to changes in the performance for other traits (Chapter 12).

Example 8.2 A well-known positive correlation in cereals is that between grain yield and plant height. This positive correlation has not prevented the development of high yielding, short-statured wheat varieties replacing the former lower yielding, taller varieties. This correlation is in part due to variation in competitive ability: at high plant density highly competitive plants produce long culms and many tillers, whereas plants with a poor competitive ability produce short culms and many tillers, whereas plants with a poor competitive ability produce short culms and few tillers.

Bos (1981, p. 94 and 124) estimated this coefficient of correlation for winter rye populations grown in the growing season 1977–78. He obtained for a diploid population $r = 0.31$ ($N = 102$) and for an autotetraploid population $r = 0.53$ ($N = 4,471$).

Yield is a trait of prime importance and generally displays quantitative variation. It is determined not only by the pattern of reactions with regard to external conditions (such as presence or absence of pathogens, pests and drought, the temperature, the actual photo period, the amount of fertilizers, *etc.*), but also by the internal control of the distribution of the products of photosynthesis (and their reallocation at grain filling and maturation). An aim is often to increase yield by improvement of the yield components and by improved resistance to biotic and abiotic factors reducing the yield. The notion of yield components is somewhat developed in Example 8.3.

Example 8.3 Yield components receive a lot of attention, especially in cereals. The grain yield (Y) is the product of $X_1 :=$ number of ears per plant; $X_2 :=$ number of spikelets per ear; $X_3 :=$ number of grains per spikelet; and $X_4 :=$ single-grain weight.

In contrast to Y and its components, the harvest index (Y/biomass), is hardly affected by the plant density, *i.e.* by the strength of interplant competition.

The opinion that the quantitative variation in certain traits is determined (directly or indirectly) by many loci is supported by the results of some long-lasting selection experiments: after apparently successful selection, continued for 50 or more generations, the genetic variation was still not exhausted (Example 8.4).

Example 8.4 Dudley, Lambert and Alexander (1974) reported that after 70 generations of selection in maize the mean phenotypic values for high protein (HP), low protein (LP), high oil (HO) and low oil (LO) content amounted, in the populations obtained by continued selection, to 215%, 23%, 341% and 14%, respectively, of the means of the original population (with 10.9% protein and 4.7% oil).

Selection had not yet exhausted the genetic variation: a comparison of the last six generations of the HP, LP, HO and LO populations grown in 1970 and 1971 showed significant differences among the generations. Furthermore, significant genetic variation among half sib families of the sixty-fifth generation was established.

A correlated response to selection was only found for oil and protein content in the LP population, where the reduction in protein to 4.5% was accompanied by a significant reduction in oil content. As a result of increased oil fertility, protein content increased in both HO and LO.

Selection had a marked effect on kernel weight and appearance of the plant material: kernels of HP and HO were small and vitreous, with those of HP being the smaller. In contrast, kernels of LP and LO were larger and had a high content of soft starch. Kernels of LO were the largest.

In the breeding of self-fertilizing crops it is of utmost importance that the F_2 population (and so its predecessor, the F_1) consists of many plants. In this case it may contain one or more plants with a highly heterozygous genotype capable of generating homozygous offspring that perform in a superior way when grown in the absence of variation for competitive ability. The breeder is charged with the task of identifying, in such a large heterogeneous F_2 population, plants with the genotype with this capability. As a matter of fact it is virtually impossible to fulfil this task fully: mostly there is hardly a correlation between the yield of F_2 plants and the yield obtained from the corresponding F_3 lines (Example 8.5, Section 18.3). Chapter 17 summarizes retrospectively the causes for the low efficiency of selection.

Example 8.5 McGinnes and Shebeski (1968) estimated the correlation between F_2 plant yield and F_3 line yield for wheat to amount to only 0.13. Similar research has been reported by DePauw and Shebeski (1973), Hamblin and Donald (1974) and Whan, Rathjen and Knight (1981) and Whan, Knight and Rathjen (1982).

Inefficiency of selection results from

1. Non-identical reproduction.
2. Variation in the quality of the growing conditions, *e.g.* variation in soil fertility.
3. Competition.
4. Inaccuracy of the observations underlying the selection. This applies especially to visual assessment of the candidates.

Non-identical reproduction as a cause for inefficient selection

Identical reproduction occurs when the genotype of the offspring obtained from some entry is identical to the genotype of its parent. It occurs at asexual reproduction of clones, at selfing of pure lines, and at re-production (by making the underlying crosses again) of single-cross hybrids. In this case the composition of a population is constant in successive generations.

A genetic cause for a disappointing response to selection is **non-identical reproduction** of the selected entries, *i.e.* single plants, lines or families. By this is meant that the genotypes of the entries selected on the basis of their phenotype (these entries constitute generation $G_t{}'$), are not identically reproduced and do, consequently, not reoccur unaltered in generation G_{t+1}. For example, in the F_2 many plants are heterozygous for many loci. This heterozygosity may give rise to heterosis. If so, then preferentially highly heterozygous F_2 plants will be selected. These will produce less heterozygous offspring whose performance is inferior when compared to their parents. This mechanism applies of course also to cross-fertilizing crops: excellent (*i.e.* possibly strongly

heterozygous) plants are likely to generate less heterozygous and consequently less excellent offspring.

Selection at a situation with identical reproduction occurs when selecting among clones, among completely homozygous plants of a self-fertilizing crop or among test hybrids when developing a single cross hybrid.

Variation in growing conditions as a cause for inefficient selection

Growing conditions always vary across the candidates. Therefore, when comparing entries, care should be taken to ensure that the growing conditions experienced by different candidates are equal (or taken into account). Only then can the candidates be ranked reliably according to their 'genetic quality'. Therefore Fisher (1935) advocated

1. Comparison of entries within **blocks**
   A block consists of a number of plots that offer, it is hoped, equal growing conditions. If this applies comparisons among entries, occurring within the same block, offer unbiased estimates of genetic differences. (In practice, however, growing conditions tend to vary within large blocks).
2. **Randomization**
   The candidates to be tested are assigned at random to the plots within each block. This removes correlation between the genotypic values of the candidates and quality of their growing conditions, *e.g.* the growth pattern of the direct neighbours.
3. **Replication**
   Replication allows not only estimation of the error variance, and consequently application of statistical tests, but it promotes also the accuracy of the estimation of the genotypic values of the tested candidates. Replicated testing of all candidates is often impossible, for example, because

   (a) Certain candidates can only be represented by a single plant (this applies to F_2 plants) or by a small number of plants (this applies to F_3 lines, *e.g.* of peas).
   (b) Because of limitations in the capacity for testing candidates, replicated testing of all candidates is prohibited.

Inability to apply replicated testing, as well as the notion that uniformity of the growing conditions within the blocks is an idealization, have stimulated interest in evaluation procedures employing incomplete block designs and/or non-replicated evaluation. These latter procedures make use of **standard plots** (Section 14.3.2) or **moving means** (Section 14.3.3). They are based on the fact that adjacent plots provide growing conditions that are more similar in quality than non-adjacent plots. (This does not include the quality of the growing conditions as determined by the strength of the competition exerted by candidates evaluated at directly adjacent plots (Chapter 15)).

Competition as a cause for inefficient selection

Competition reduces the efficiency of selection of genetically superior candidates from a genetically heterogeneous population of candidates. Candidates with a strong competitive ability, which are apt to be selected, may perform disappointingly when grown in the absence of variation in competitive ability (Chapter 15; Spitters, 1979, pp. 9–10).

Inaccuracy of the observations as a cause of inefficient selection

Inaccuracy of the observations underlying the selection contributes to the inefficiency of selection. It works out like random variation in the quality of the growing conditions. It occurs especially when evaluating candidates on the basis of visual assessment. This topic is elaborated in Chapter 14, notably Section 14.3.1.

In summary, one may say that the task of a breeder is very difficult because selection is on the basis of the phenotype of the candidates. The offspring of the selected candidates may perform differently to their parents. This is due to the fact that the parent and offspring have different genotypes (except in the case of identical reproduction) and/or due to different growing conditions. Therefore it is sometimes said that selection concerning quantitative variation is not so much a science but more an art.

Chapters 8 to 12 of this book aim to indicate how an answer can be obtained to the following questions:

1. What part of the observed phenotypic variation is due to genetic variation? In other words: how large is the heritability? The answer to this question indicates how efficient selection may be expected to be.
2. How large will the expected response to selection be when applying a certain selection intensity?
   The answer will, of course, depend on the efficiency of the selection and on the amount of genetic variation available.
3. How large is the probability that the genotypic value of a random plant, to be sampled from the F_∞ population still to be developed, exceeds the genotypic value of a standard variety?

8.2 Components of the Phenotypic Value

The expression observed for a quantitative trait of some candidate is mostly indicated by a numerical value, the **phenotypic value** (p). Example 8.6 shows that the decision about how to assign numerical values, *e.g.* the value $p = 0$, to a certain level of expression may be arbitrary.

Example 8.6 With regard to the reaction of a genotype to inoculation with a certain pathogen one may indicate 'not susceptible' by $p = 0$, and 'very susceptible' by $p = 10$. This is rather arbitrary because one could also follow the principle of assigning low values to undesired expressions and high values to desired expressions. Then 'very susceptible' would be coded as $p = 0$ and 'not susceptible' as $p = 10$ (This system is followed in the Dutch lists of varieties).

With regard to date of flowering p may indicate the number of days from sowing to flowering, or the number of days from May 1 to flowering, *etc.*

For traits like yield, plant height, protein content *etc.* there is a natural origin, *i.e.* the phenotypic value specified by $p = 0$. But then the scale of measurement still has to be chosen, *e.g.* yield in grams or kilograms, plant height in centimetres or inches, fruit size in gram or in centimetres.

The phenotypic value of an entry results from the interaction of the complex genotype of the observed entry and its growing conditions. It is useless to describe this dependency by $p = f(\mathcal{G}, e)$ because the function describing how the phenotypic value is determined by the (complex) genotype ($\mathcal{G}$) and by the growing conditions (e) is unknown. Quantitative genetic theory is not dedicated to clarifying the function relating phenotypic value to genotype and environment. Instead, quantitative genetic theory was developed from the side of the phenotypic values. On the basis of the phenotypic values observed for plants sharing a not further specified complex genotype, one assigns a genotypic value to the complex genotype. In Section 8.3 ways are developed to partition this genotypic value into contributions due to the single-locus genotype for each separate relevant locus.

The distinction, first made by Johansson (1909), between the genotype of a plant and its phenotype has been very fruitful. It showed that the relationship between genotype and phenotype varies: the presence of a certain allele does not always give rise to a phenotypically observable effect in comparison to the absence of that allele. Thus in the case of complete dominance of allele B over allele b the genotypes Bb and BB will give rise to identical phenotypes in the case of qualitative variation.

The phenotypic expression of a allele may also depend on the growing conditions or on plant-associated factors, *e.g.* age or sex. Sometimes only a portion of the plants with a certain genotype shows the phenotype that 'should be expressed'. This portion is called **penetrance**. The genetic background of this phenomenon is not considered further; it is only mentioned to show that a genotype may give rise to diverse phenotypes. Allard (1960, p. 66) gives an example.

In connection with the notions of 'phenotype' and 'genotype' the notions of phenotypic value (p) and **genotypic value** ($\mathcal{G}$) have been defined. The

parameter p represents the observation obtained from a single entry, *i.e.* a single plant or a single plot containing certain plant material. **Genotypic value** is defined as the expected phenotypic value of the considered genotype (gt) at the considered **macro-environmental conditions** (E). Thus:

$$\mathcal{G} = E(\underline{p}|gt, E)$$

The macro-environmental conditions are specified by the combination of site, growing season and applied cultivation regime (in Chapter 14 special attention is given to plant density).

The genotypic value of a certain genotype, grown under specified macro-environmental conditions, can be estimated by the arithmetic mean of the phenotypic values calculated across all n plants with the considered genotype and grown under the considered conditions:

$$\hat{\mathcal{G}} = \frac{\sum_{i=1}^{n} p_i}{n} = \bar{p}$$

If identical reproduction is impossible, each genotype is represented by only one plant ($n = 1$). In that case $\hat{\mathcal{G}} = p$. This estimate is of course very inaccurate (a way-out is suggested below). If, however, identical reproduction is possible, *e.g.* when dealing with a clone, a pure line or a single cross hybrid, n may be very large and accurate estimation of $\mathcal{G}$ is possible (see Example 8.7).

Example 8.7 The phenotypic value for plant height of some plant belonging to the spring wheat variety Peko, grown in 1971 at a $15 \times 25\,\mathrm{cm}^2$ pattern of plant positions, is 109 cm. The genotypic value of Peko, when grown at these macro-environmental conditions, was estimated to be 93.4 cm (Table 8.1).

In Example 9.1 it is shown that in the case of absence of dominance and epistasis the expected phenotypic (and genotypic) value of the plants belonging to the line obtained from some plant P_i is equal to the genotypic value of that plant. Thus:

$$\mathrm{E}\underline{p}_{\mathrm{L}(\mathrm{P}_i)} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{L}(\mathrm{P}_i)} = \mathcal{G}_{\mathrm{P}_i}$$

Likewise, Example 9.2 shows, for the same conditions, that the expected phenotypic value of the plants belonging to the full sib family obtained from some cross $\mathrm{P}_i \times \mathrm{P}_j$ is equal to the mean genotypic value of the two parental plants:

$$\mathrm{E}\underline{p}_{\mathrm{FS}(\mathrm{P}_i \times \mathrm{P}_j)} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}_{ij}} = \frac{1}{2}(\mathcal{G}_{\mathrm{P}_i} + \mathcal{G}_{\mathrm{P}_j})$$

If the full sib families FS_{ij}, FS_{ik} and FS_{jk} are obtained from plants P_i, P_j and P_k, and if a 'reasonable number' of plants of these families are grown and observed, one may obtain accurate estimates for $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}_{ij}}$, $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}_{ik}}$ and $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}_{jk}}$.

Then one may derive from the above equation estimates of the genotypic values of the parental plants. Van der Vossen (1974) applied progeny testing in order to be able to estimate the genotypic values of oil palm genotypes represented by a single tree.

The genotypic value of a genotype applies only to the specified macro-environmental growing conditions. This means that the genotypic value assigned to a genotype depends on the macro-environment. As a consequence, the variance of the genotypic values depends on the growing conditions. This is illustrated in Example 8.8.

Example 8.8 Spitters (1979, Tables 25, 27, 28 and 38) grew, in 1977, 12 different spring barley varieties at four different macro-environmental conditions:

1. as pure lines at a plant density of 80 (plants/m^2);
2. as mixtures also at a density of 80;
3. as mixtures at a plant density of only 3.2; and
4. as pure lines at commercial plant density (about 180 plants/m^2, the amount of seed was 110 kg/ha).

The yield and rank number of each variety under each of the four conditions are summarized in Table 8.3.

Table 8.3 Grain yield (in g/plant; for condition 4 in g/row) and rank (from 1 = lowest to 12 = highest) of 12 spring barley varieties grown in 1977 under four different conditions (see text) (source: Spitters, 1979, Tables 25, 27, 28, 38)

	Condition							
	1		2		3		4	
Variety	yield	rank	yield	rank	yield	rank	yield	rank
Varunda	5.3	6.5	5.1	5.5	41	4	150	5
Tamara	5.7	10	7.8	12	53	11	165	11.5
Belfor	5.3	6.5	5.4	9.5	57	12	161	10
Aramir	6.1	12	5.3	7.5	49	8	154	7
Camilla	5.0	5	5.4	9.5	50	9	165	11.5
G. Promise	4.5	1	4.9	4	40	2.5	132	4
Balder	4.8	4	5.1	5.5	42	5.5	156	8.5
WZ	5.5	8	4.8	3	51	10	151	6
Goudgerst	4.7	3	7.7	11	42	5.5	131	3
L98	6.0	11	3.5	2	40	2.5	106	1
Titan	4.6	2	1.6	1	37	1	109	2
Bigo	5.6	9	5.3	7.5	45	7	156	8.5
	$\bar{\mathcal{G}} = 5.26$		$\bar{\mathcal{G}} = 5.16$		$\bar{\mathcal{G}} = 45.6$			
			$s_g^2 = 2.65$		$s_g^2 = 39.0$			

It appears that the genotypic value depends on the plant density (compare conditions 1 and 4) and, for a certain plant density, on the presence

or absence of genetic variation for competitive ability (compare conditions 1 and 2). This dependency affects the genetic variance. Thus the variance of the genotypic values presented in Table 8 is 0.269 $(g/plant)^2$ at condition 1 and 2.43 $(g/plant)^2$ at condition 2.

Goudgerst had a relatively low genotypic value for grain yield when grown as a pure line but a relatively high genotypic value when grown in mixtures. For other genotypes grown as pure lines, plant density had an important impact on genotypic value, *e.g.* L98. The ranking of the varieties at low plant density differed strongly from the ranking at commercial plant density. Thus important effects of genotype × density interaction are evident.

According to our definition of the genotypic value, the quality of the macro-environmental conditions affects the genotypic value: the same genotype will thus have different genotypic values in different macro-environments. The ranking of a set of genotypes according to their genotypic values in one environment may thus differ from their ranking in another environment. Such **genotype × environment interaction** implies that one should not make statements such as 'the single-cross hybrid of inbred lines A and B shows mid-parent heterosis with regard to number of grains per ear', or 'variety P_1 yields better than variety $P_2{}'$ without specifying the macro-environmental conditions for which the statement is made. In Chapter 13 attention is given to the phenotypic values of genotypes in different macro-environments. That situation requires a somewhat different definition for the notion of genotypic value.

Here, as well as in all other chapters, except Chapter 13, the situation of absence of variation in macro-environmental conditions is considered. This implies that the genotypic values (and consequently their variance) are not affected by a change of macro-environment. Differences between populations, in fact differences between different generations of the same population, with regard to their expected genotypic values or their genetic variances are then not due to differences between the growing conditions prevailing in the different growing seasons.

The difference between the phenotypic value assigned to an entry (a plant or an entry grown as a plot) and the genotypic value assigned to the entry, is attributed to the complex of environmental conditions to which the considered entry is exposed. This difference is called **environmental deviation** (e). Thus

$$e = p - \mathcal{G}$$

When considering a number of entries sharing the same genotype we can write

$$\underline{e} = \underline{p} - \mathcal{G}$$

The expected value of the environmental deviation is, due to the definition of the genotypic value, necessarily equal to 0:

$$\mathrm{E}\underline{e} = \mathrm{E}(\underline{p} - \mathcal{G}) = (\mathrm{E}\underline{p}) - \mathcal{G} = \mathcal{G} - \mathcal{G} = 0$$

For a genetically homogeneous group of plants the expression

$$\underline{p} = \mathcal{G} + \underline{e}$$

implies

$$\mathrm{E}\underline{p} = \mathrm{E}(\mathcal{G} + \underline{e}) = \mathcal{G}$$

and

$$\mathrm{var}(\underline{p}) = \mathrm{var}(\underline{e})$$

For a genetically heterogeneous population of entries the expression

$$\underline{p} = \mathcal{G} + \underline{e} \tag{8.1}$$

implies

$$\mathrm{E}\underline{p} = \mathrm{E}(\underline{\mathcal{G}} + \underline{e}) = \mathrm{E}\underline{\mathcal{G}}$$

and

$$\mathrm{var}(\underline{p}) = \mathrm{var}(\underline{\mathcal{G}} + \underline{e}) = \mathrm{var}(\underline{\mathcal{G}}) + \mathrm{var}(\underline{e}) + 2\mathrm{cov}(\underline{\mathcal{G}}, e)$$

In the case of a random exposure of the genotypes of the entries to the **micro-environmental conditions** the random variables $\underline{\mathcal{G}}$ and $\underline{e}$ are independently distributed across the entries. This implies $\mathrm{cov}(\underline{\mathcal{G}}, \underline{e}) = 0$. Randomization thus induces absence of correlation of genotypic value and environmental deviation. It implies

$$\mathrm{var}(\underline{p}) = \mathrm{var}(\underline{\mathcal{G}}) + \mathrm{var}(\underline{e}) \tag{8.2}$$

In words: the **phenotypic variance** (variance of the phenotypic values) is equal to the **genetic variance** (variance of the genotypic values) plus the **environmental variance** (variance of the environmental deviations).

The simple model described by Equation (8.1), *i.e.* $\underline{p} = \underline{\mathcal{G}} + \underline{e}$, results from the way of defining the environmental deviation. Other models may also be considered as a basis for developing a quantitative genetic theory, *e.g.*:

1. $\underline{p} = \underline{\mathcal{G}} \cdot \underline{e}$
   This simplifies by logarithmic transformation, *i.e.* $\log(\underline{p}) = \log(\underline{\mathcal{G}}) + \log(\underline{e})$, into $\underline{p}' = \underline{\mathcal{G}} + \underline{e}'$.
2. $\underline{p} = \underline{c}(\mu + \underline{\mathcal{G}}) + \underline{e}$, (Spitters, 1979, p. 51, where μ is the population mean and $\underline{c}$ the genetically determined competitive ability, see Section 15.1).

A high value for the environmental variance, or for the (dimensionless!) environmental coefficient of variation ($\boldsymbol{\nu}\mathbf{c}_e = \frac{\boldsymbol{\sigma}_e}{\mathbf{E}_p}$), does not necessarily mean that the plants are exposed to very variable growing conditions. The environmental variance as such is a poor yardstick for measuring the variation in the growing conditions. If a genotype shows a large environmental variance, it could mean that it has a small capacity to buffer its phenotypic values against a relatively

small variation in the growing conditions. (**Canalization** is buffering of the phenotypic values in such a way that variation in growing conditions does not give rise to phenotypic variation: all tulip plants belonging to a certain clonal variety produce a flower with the same colour intensity, notwithstanding variation in micro-environmental conditions.) Indeed, the genotype determines how the phenotypic values of the plants with the considered genotype vary under some range of growing conditions. Some genotypes give rise to more stable phenotypes than others: they show, for the same variation in growing conditions, a smaller environmental variance than other genotypes. Such genotypes are said to posses a higher physiological homeostasis. (The latter is sometimes claimed to be associated with a higher heterozygosity. That would confer a higher average fitness value across various micro-environmental conditions as compared to more homozygous genotypes, see Section 13.2 for a more detailed discussion.)

Association, across different genotypes, of $\mathrm{E}\underline{p}$ and $\mathrm{var}(\underline{p})$ in such a way that the coefficient of phenotypic variation (vc_p) is constant is called a **scale effect**. Generally, a logarithmic transformation then leads to equal variances (Falconer, 1989, p. 294). The estimates for vc_p given in Table 8.4 are nearly constant; however, those for the inbred lines are the highest.

If some genetically uniform entry (a clone, a pure line or a single cross hybrid) is grown in different fields, the environmental variances with regard to some trait, as estimated for each separate field, indicate how the variation for the trait is affected by the variation in the growing conditions as offered by each field. Example 8.9 illustrates a relation between the average phenotypic value and the phenotypic variance. It also discusses the possible relationship with the degree of heterozygosity.

8.3 Components of the Genotypic Value

8.3.1 Introduction

The complex genotype affecting the phenotypic value of an entry for a trait with quantitative variation consists of the aggregate, across all relevant loci, of the single-locus genotype for each relevant locus. These relevant loci comprise segregating loci, contributing to the genetic variation in the considered population, as well as non-segregating loci (for which all plants in the population have the same (homozygous) genotype). It is often (sometimes implicitly) assumed that each segregating locus segregates for only two alleles. The situations where this restriction can be justified were indicated in Section 2.2.1.

Example 8.9 For the same field, plants of the potato variety Bintje were less buffered with regard to yield per plant against variation in the growing conditions than plants of the spring wheat variety Peko for plant height. The coefficients of environmental variation amounted to 0.25 and 0.06 (Table 8.2), respectively.

Van Cruchten (1973) measured the height (in centimetres; from the soil to the lowest branch of the male inflorescence) of maize plants. He did so for four inbred lines (W, X, Y and Z), for two single-cross hybrids (WX and YZ) and for the double-cross hybrid (WXYZ, produced by crossing the single-cross hybrids). He estimated for each entry $E\underline{p}$, $\text{var}(\underline{p})$ and vc_p (These parameters can, except for WXYZ, be interpreted as $\mathcal{G}$, $\text{var}(\underline{e})$ and $vc_{e\rangle}$. The results are summarized in Table 8.4.

Table 8.4 Estimates for $E\underline{p}$, $\text{var}(\underline{p})$ and vc_p for plant height (in centimetres) in maize

Material	$\bar{p}$	${s_p}^2$	$v\hat{c}_p$
W	103.8	185	0.13
X	121.1	256	0.13
Y	80.5	90.3	0.12
Z	111.6	285.6	0.15
WX	177.6	424.4	0.12
YZ	141.2	240.3	0.11
WXYZ	188.2	475.3	0.12

Across these seven entries the coefficient of correlation between $\bar{p}$ and s_p^2 amounted to 0.95. There is thus a very clear indication of occurrence of a scale effect. The values for ${s_p}^2$ reflect the balance of this positive relation and the negative relation between the inbreeding coefficient and the stability.

This latter relation is observed or assumed by some researchers. Falconer's question 'What then is the cause of some characters being more variable in inbreds than in hybrids?' (Falconer, 1989, p. 269) suggests a negative relation between inbreeding coefficient and stability. Also Allard and Bradshaw (1964) conclude that the size of $\text{var}(\underline{e})$ depends on the degree of heterozygosity of the genotype: 'In outbreeding species there is a good deal of work which indicates that buffering is conspicuously a property of a heterozygote ... In inbreeding species there is evidence that buffering can be a property of specific genotypes not associated with heterozygosity'. This topic is further discussed in Section 13.2.

In quantitative genetic theory developed for a locus represented by only two alleles, the three genotypes for some locus may be coded as follows:

1. The homozygous genotype with the lower genotypic value may be coded by A_2A_2
2. The heterozygous genotype by A_1A_2
3. The homozygous genotype with the higher genotypic value by A_1A_1

Falconer (1989, p. 112) used this coding. These codes do not reveal whether dominance occurs or, when it occurs, which of the two alleles is dominant.

In the present book locus B-b represents any locus affecting the expression for the considered quantitative trait. The coding of the genotypes is as follows:

1. The homozygous genotype giving rise to the lower genotypic value is coded bb
2. The heterozygous genotype is coded Bb
3. The homozygous genotype with the higher genotypic value is coded BB

With this coding system the notation reveals nothing about dominance. However, in Section 9.4.1 it is shown that, if dominance occurs, allele B tends to be the dominant allele. It is, indeed, shown that **unidirectional dominance** is to be expected, *i.e.* allele B is the dominant allele for most of the k relevant loci B_1-$b_1, \ldots, B_k$-b_k. This implies that for many traits the (population) genetic and the quantitative genetic implications of the codes coincide. This is not the case if **ambidirectional dominance** occurs, *i.e.* for some relevant loci allele B is dominant and for other relevant loci allele b. Ambidirectional dominance has been established for certain traits, *e.g.* in wheat for date of anthesis and for compactness of the ear.

Quantitative genetic analysis predominantly reveals effects emerging from segregating loci. The contribution to the phenotypic values due to the common complex genotype for all non-segregating loci, sometimes indicated as **genetic background**, is measured by an important quantitative genetic parameter, *viz.* m (Section 8.3.2).

One may generally state that k segregating loci, say B_1-$b_1, \ldots, B_k$-b_k, affect the variation for the considered trait. The value for k varies from trait to trait and for a given trait from population to population. An arbitrary locus from this set of loci is locus B_i-b_i. In short, we let locus B-b represent any of the segregating loci.

Different systems have been adopted for the partitioning of genotypic values in meaningful components. They aim at the derivation of simple expressions for expectations and variances of genotypic values in terms of their components. Section 8.3.2 deals with the socalled F_∞-metric for partitioning of the genotypic value. It applies well to situations where loci are represented by only two alleles. According to Section 2.2.1 this is common in populations of self-fertilizing crops. For situations with multiple allelism, which is to be expected in populations of cross-fertilizing crops, partitioning of the genotypic value in the additive genotypic value and the dominance deviation is appropriate, see Section 8.3.3. The latter components will also be written in terms of F_∞-metric parameters. Because of that, first attention is given to the F_∞-metric.

8.3.2 *Partitioning of Genotypic Values According to the* F_∞*-metric*

In the F_∞-metric the genotypic values for the three genotypes for locus B-b are partitioned in terms of the parameters m, a and d, where

$$\begin{aligned} m &:= \tfrac{1}{2}(\mathcal{G}_{bb} + \mathcal{G}_{BB}) \\ a &:= \tfrac{1}{2}(\mathcal{G}_{BB} - \mathcal{G}_{bb}) \\ d &:= \mathcal{G}_{Bb} - m \end{aligned}$$

These definitions allow the following partitioning of the genotypic values:

	Genotype		
	bb	Bb	BB
$\mathcal{G}$	$m-a$	$m+d$	$m+a$

Due to its definition, component m is called the **midparent value**. This parameter represents the contribution to the genotypic values due to the genetic background. In fact the F_∞-metric owes its name to the way of defining m for any number of segregating loci.

The parameter a describes the deviations of the genotypic value of the homozygous genotypes from the midparent value:

$$a = \mathcal{G}_{BB} - m = m - \mathcal{G}_{bb}$$

Because of the system of coding of the genotypes, the inequality $\mathcal{G}_{BB} > \mathcal{G}_{bb}$ applies. Thus $a \geq 0$.

The parameter d indicates the deviation of the genotypic value of the heterozygous genotype from the midparent value:

$$d = \mathcal{G}_{Bb} - m$$

If $d = 0$ then $\mathcal{G}_{Bb} = m = \frac{1}{2}(\mathcal{G}_{bb} + \mathcal{G}_{BB})$: the genotypic value of Bb is intermediate with regard to those of bb and BB. This absence of dominance implies **additivity** of allele effects. If $\mathcal{G}_{Bb} - \mathcal{G}_{bb} \neq \mathcal{G}_{BB} - \mathcal{G}_{Bb}$ the genotypic value of Bb is not intermediate. Then the effect of the second allele present in a genotype depends on the first allele. This phenomenon is sometimes called **intra-locus-interaction**, but it is more commonly called **dominance**. In the F_∞-metric it is, in the case of dominance, impossible to consider the genotypic value as the sum of the effects of the two alleles involved in the genotype. Because dominance is a common phenomenon one should, within the F_∞-metric system of partitioning of genotypic values, avoid the use of the word **allele-effect**. Within the alternative system for partitioning genotypic values, developed in Section 8.3.3, use of the term allele-effect is legitimate, even in the presence of dominance.

The **degree of dominance** follows from the comparison of a and d:

$d < -a$:	overdominance of b
$d = -$a:	complete dominance of b
$-$a $< d < 0$:	incomplete dominance of b
$d = 0$:	no dominance, *i.e.* additivity
$0 < d < a$:	incomplete dominance of B

$d = a$: complete dominance of B

$d > a$: overdominance of B (see Note 8.1)

Note 8.1 From about 1910 Shull and East formulated hypotheses to explain **heterosis**, the phenomenon that heterozygous plant material performs better than its homozygous parents. Because overdominance at the level of single-locus genotypes is a rare phenomenon (Section 6.2), an explanation of heterosis on the basis of single-locus overdominance is inappropriate. However, in Section 9.4.1 it will be explained that heterosis is to be expected at any degree of dominance provided that $d > 0$.

Example 8.10 illustrates how one may assign numerical values to the parameters m, a and d.

Example 8.10 For the following genotypic values

	Genotype		
	b_1b_1	B_1b_1	B_1B_1
$\mathcal{G}$	12	14	16

one can derive: $m = \frac{1}{2}(12 + 16) = 14, a_1 = \frac{1}{2}(16 - 12) = 2$ and $d_1 = 14 - 14 = 0$.
For

	Genotype		
	b_2b_2	B_2b_2	B_2B_2
$\mathcal{G}$	7	15	15

we get $m = \frac{1}{2}(7 + 15) = 11$, $a_2 = \frac{1}{2}(15 - 7) = 4, d_2 = 15 - 11 = 4$.

Example 8.11 shows that it may be difficult to decide about presence or absence of dominance.

Example 8.11 The size of tomatoes may be measured by their weight as well as by their diameter. The two different scales of measurement give rise to different genotypic values and to different degrees of dominance. This is illustrated by means of data on fruit size of tomato species and of their interspecific hybrid. MacArthur and Butler (1938) measured fruit size by determining fruit weight (w; in g) and obtained the following results:

	Fruit size (g)		
Cross	P_1	P_2	F_1
1	1.1	12.1	4.2
2	1.1	54.1	7.4
3	1.1	152.4	10.1
4	12.4	112.6	35.5

It may be concluded that, as measured by weight, small fruit size tends to be dominant.

When measuring fruit size by r, the radius of the spherical fruits, and approximating r (in cm) by $r = \left(\frac{0.75w}{\pi}\right)^{\frac{1}{3}}$ we get

	Fruit size (cm)		
Cross	P_1	P_2	F_1
1	0.640	1.424	1.001
2	0.640	2.346	1.209
3	0.640	3.314	1.341
4	1.436	2.996	2.039

According to this scale of measurement there is hardly any dominance for fruit size.

Yield is a complex trait. In its simplest form it is the product of number of fruits and single fruit weight. The genetic control of each of these two components may be expected to be more direct and more simple than the (indirect) genetic control of yield itself. Tables 9.3 and 9.4 present for each of these components examples of intermediate phenotypic values of the offspring, compared to the parents, whereas heterosis appears to occur with regard to yield.

Now the partitioning of genotypic values according to the F_∞-metric is extended to complex genotypes consisting of single-locus genotypes for each of the K segregating polygenic loci B_1-$b_1, \ldots, B_K$-b_K.

First the situation of $K = 2$ is considered. The genotypic value of some complex genotype for loci B_1-b_1 and B_2-b_2, designated as $\mathcal{G}_{B1\text{-}b1,B_2\text{-}b_2}$, is assumed to consist of the sum of

- the genotypic value of the complex genotype for all non-segregating loci, say m;
- a contribution due to the genotype for locus B_1-b_1, say $\mathcal{G}'_{B1\text{-}b1}$;
- a contribution due to the genotype for locus B_2-b_2, say $\mathcal{G}'_{B2\text{-}b2}$ and
- the effect of interaction of the single-locus genotypes for loci B_1-b_1 and B_2-b_2, say $\boldsymbol{i}_{B1\text{-}b1,B2\text{-}b2}$.

Thus

$$\mathcal{G}_{B_1\text{-}b_1,B_2\text{-}b_2} = \boldsymbol{m} + \mathcal{G}'_{B_1\text{-}b_1} + \mathcal{G}'_{B_2\text{-}b_2} + \boldsymbol{i}_{B_1\text{-}b_1,B_2\text{-}b_2} \tag{8.3}$$

If $\boldsymbol{i}_{B1\text{-}b1,B_2\text{-}b2}$, say $\boldsymbol{i}$, is zero for each of the nine complex genotypes, the genotypic value of a complex genotype simply consists of $\boldsymbol{m} + \mathcal{G}'_{B1\text{-}b1} + \mathcal{G}'_{B2\text{-}b2}$. The contribution of the single-locus genotype for locus B_1-b_1 to the genotypic value of the complex genotype does then not depend on the genotype for locus B_2-b_2. The difference $\mathcal{G}_{B1b1..} - \mathcal{G}_{b1b1..}$ is then equal to $\mathcal{G}'_{B1b1} - \mathcal{G}'_{b1b1}$, whatever

the genotype for locus B_2-b_2 is. This may be called **additivity of single-locus genotype effects**.

If $\boldsymbol{i} \neq 0$ for one or more of the nine complex genotypes, **inter-locus-interaction**, more commonly called **epistasis**, is present. In that case one cannot specify single-locus genotype effects, and then one should not use the term **genotype-effect**. (Note 8.2 indicates that the meaning of the word epistasis depends on the context).

Note 8.2 For qualitative variation the term **epistasis** has a more specific meaning than for quantitative variation, where it indicates the presence of any form of inter-locus-interaction (which is also indicated as **non-allelic interaction**).

Example 8.12 illustrates (a) the partitioning of the genotypic values of complex genotypes in terms of the parameters m, a and d, and (b) how to conclude about the presence or the absence of epistasis.

Example 8.12 The scheme below provides the genotypic values for the nine complex genotypes possible for loci B_3-b_3 and B_4-b_4:

	b_3b_3	B_3b_3	B_3B_3
b_4b_4	11	13	13
B_4b_4	12	14	14
B_4B_4	12	14	14

It appears that epistasis is absent.

The value of m is calculated as the mean genotypic value across the four homozygous genotypes: $m = \frac{1}{4}(11 + 13 + 12 + 14) = 12.5$.

At both loci there is complete dominance: $a_3 = d_3 = 1; a_4 = d_4 = \frac{1}{2}$.

The next scheme provides the genotypic values for the nine complex genotypes for loci B_5-b_5 and B_6-b_6:

	b_5b_5	B_5b_5	B_5B_5
b_6b_6	11	11	11
B_6b_6	11	13	13
B_6B_6	11	13	13

It appears that $\mathcal{G}_{B5B5b6b6} - \mathcal{G}_{b5b5b6b6} = 0$, whereas $\mathcal{G}_{B5B5B6B6} - \mathcal{G}_{b5b5B6B6} = 2$. This means that the effect of genotype B_5B_5 in comparison to b_5b_5 depends on the genotype for locus B_6-b_6. Inter-locus-interaction of the two loci is demonstrated. Epistasis is present.

Epistasis occurs – of course – in the hypothetical situation where the marginal contribution of genotype BB, in comparison to genotype bb, to the genotypic value of complex genotypes is smaller as the total number of B alleles

present at the K-1 other loci is higher. This hypothesis, resembling the law of diminishing returns, was put forward by Rasmusson (1933). Physiological limits with regard to the expression of quantitative variation certainly induce the occurrence of epistasis, implying that it will become harder to realize further progress by selection as this physiological limit is more closely approximated.

Epistasis should generally be expected because the genotypic value for some trait is ultimately due to genotypes for loci controlling successive steps of a metabolic process: the homozygous genotype b_1b_1 for the mutant allele b_1 may block the process, influencing the effect of genotype B_2B_2 in comparison to genotype b_2b_2.

So far, the interaction of the single-locus genotypes for loci B_1-b_1 and B_2-b_2, was generally indicated by $i_{B1\text{-}b1,B2\text{-}b2}$. The interaction effects occurring within pairs of single-locus genotypes when considering the nine complex genotypes possible for $K = 2$ will be represented by logical symbols: *aa*, *ad*, *da* and *dd* (Kearsey and Pooni, 1996, p. 225).

- *aa* represents the effect of interaction of a homozygous genotype for locus B_1-b_1 and a homozygous genotype for locus B_2-b_2
- *ad* represents the effect of interaction of a homozygous genotype for locus B_1-b_1 and a heterozygous genotype for locus B_2-b_2
- *da* represents the effect of interaction of a heterozygous genotype for locus B_1-b_1 and a homozygous genotype for locus B_2-b_2
- *dd* represents the effect of interaction of a heterozygous genotype for locus B_1-b_1 and a heterozygous genotype for locus B_2-b_2

Table 8.5 presents the partitioning of the genotypic values for the nine complex genotypes possible for $K = 2$.

Partitioning of the genotypic value of a complex genotype requires in the case of occurrence of epistasis thus extra parameters. When two alleles segregate for each of the K loci 3^K different complex genotypes can be distinguished. To partition unambiguously the genotypic values of each of these 3^K genotypes in total 3^K parameters are required. One of these is m. This parameter occurs in the partitioning of each genotypic value. It functions as the origin. In the so-called F_∞-metric m is equal to the unweighted mean genotypic value across the 2^K complex homozygous genotypes. It is due to the complex genotype with regard to all non-segregating loci. The $3^K - 1$ other

Table 8.5 The partitioning of the genotypic values of the nine complex genotypes with regard to loci B_1-b_1 and B_2-b_2

		Genotype for locus B_1-b_1		
		b_1b_1	B_1b_1	B_1B_1
Genotype for locus B_2-b_2:	b_2b_2:	$m - a_1 - a_2 + aa$	$m + d_1 - a_2 - da$	$m + a_1 - a_2 - aa$
	B_2b_2	$m - a_1 + d_2 - ad$	$m + d_1 + d_2 + dd$	$m + a_1 + d_2 + ad$
	B_2B_2	$m - a_1 + a_2 - aa$	$m + d_1 + a_2 + da$	$m + a_1 + a_2 + aa$

m: Origin, the unweighted mean across the four homozygous genotypes.
a_1, d_1, a_2 and d_2: Parameters for main effects of single-locus genotypes.
aa, *ad*, *da* and *dd*: Parameters for effects of interaction within pairs of single-locus genotypes.

parameters designate main effects due to single-locus genotypes and effects of interaction within pairs, within triplets, within quartets, *etc.* of such single-locus genotypes.

For $K = 3$ loci the $3^3 - 1 = 26$ parameters for main effects and interaction effects are

- Per locus: a and d; in total $3 \times 2 = 6$ parameters
- Per pair of loci: aa, ad, da and dd; in total $3 \times 4 = 12$ parameters
- Per triplet of loci: aaa, aad, ada, daa, add, dad, dda and ddd; in total $1 \times 8 = 8$ parameters

The genotypic value of genotype $B_1b_1B_2B_2b_3b_3$ is thus partitioned as

$$m + d_1 + a_2 - a_3 + da_{12} - da_{13} - aa_{23} - daa_{123}.$$

Generally the $3^K - 1$ parameters for main effects and interaction effects are

- Per locus: 2; across K loci in total: $2K$
- Per pair of loci: 4; across $\binom{K}{2}$ pairs in total $2^2\binom{K}{2}$
- Per triplet of loci: 8; across $\binom{K}{3}$ triplets in total $2^3\binom{K}{3}$, *etc.*

Altogether this adds up to

$$\sum_{i=1}^{K} \binom{K}{i} 2^i = \left[\sum_{i=0}^{K} \binom{K}{i} 2^i\right] - 1$$

Because

$$\sum_{i=0}^{K} \binom{K}{i} x^i = (1+x)^K$$

the former sum is $3^K - 1$.

The number of parameters quickly becomes unmanageable for even small values for K: for $K = 3$ it is 26, but for $K = 7$ it is already 2186. Effects of interactions within groups of three or more single-locus genotypes are therefore mostly neglected, in which case there remain

$$2K + 2^2\binom{K}{2} = 2K + 2K(K-1) = 2K^2$$

parameters; *i.e.* 18 if $K = 3$ and 98 if $K = 7$.

With regard to further development of the quantitative genetic theory, a choice between two options has to be made:

1. Development of the quantitative genetic theory on the basis of a complete partitioning of the genotypic values, or on the basis of partitioning of the genotypic values while neglecting effects of interactions within groups of three or more single-locus genotypes. In the latter situation only main-effect parameters and parameters for the interaction within pairs of single-locus genotypes are considered. The major drawback of this option is the complexity of mathematical expressions for expectations and variances of genotypic values in terms of these parameters.
2. Development of the theory on the basis of the assumption that inter-locus interaction does not occur. The drawback is that such quantitative genetic theory cannot fully be justified in those cases where epistasis occurs. Then conclusions on the basis of applications of the theory will be false and decisions may be inappropriate.

In this book the second option is chosen. Thus *absence of epistasis is assumed throughout the book.* The number of parameters then amounts to only $2K+1$. In connection with the also generally applied assumption of absence of linkage (Chapter 1), the present assumption yields relatively simple algebraic derivations and expressions for $\mathrm{E}\underline{\mathcal{G}}$ and $\mathrm{var}(\underline{\mathcal{G}})$. The reader is referred to Mather and Jinks (1982) or Kearsey and Pooni (1996) for a development of the theory based on the assumption that epistasis is present. Note 8.3 considers some findings and opinions related to the choice between the two above options.

Note 8.3 Jana (1971), Jana and Seyffert (1971, 1972) and Forkman and Seyffert (1977) considered whether the assumption of absence of epistasis can be justified. They did so by spectrophotometric determination of the content of anthocyanins in fresh flowers of common stock, *Matthiola incana* (L.) R. Br. From this point of view the trait showed quantitative variation. The genotype for the one, two or three relevant segregating loci was, however, known in the studied plant material, whereas the genetic background was uniform for all plants.

Earlier studies, involving an analysis in terms of gene-frequency dependent gene and interaction effects, were reanalysed by Jana (1971) in terms of the F_∞-metric parameters *a*, *d*, *aa*, *ad*, *da* and *dd*. It was established systematically that the original analyses led to an underestimation of the contribution of interaction effects in comparison to the analysis on the basis of the F_∞-metric.

Forkman and Seyffert (1977) established the law of the diminishing returns: 'The phenotypic response to allelic substitutions follows the characteristics of a saturation curve.'

For breeders it is important to know whether epistasis occurs or not. They may be interested in the genetic control of the heterosis expressed by a single cross-hybrid. Is the heterosis due to pseudo-overdominance or is it due to epistasis? The former requires crossing-over with regard to tightly linked loci to obtain superior homozygous genotypes; the latter may be exploited by developing and selecting a homozygous genotype. With regard to epistasis, Gardner and Lonnquist (1966) made the following remark: 'Although epistasis does not appear to be an important source of genetic variation in open-pollinated varieties of corn, this does not mean that epistasis is unimportant in corn breeding. Epistasis may be very important indeed in the hybrid produced by crossing two inbred lines.'

It is, indeed, useful to distinguish the relative contribution of epistatic effects to the genotypic values, and the relative contribution of epistatic effects to the variance of these genotypic values. In this book, like those of Hallauer and Miranda (1981) or Falconer and MacKay (1996), it is taken for granted that the major part of the genotypic value of a complex genotype is due to the effects of single-locus genotypes.

The origin in the F_∞-metric is m, *i.e.* the contribution to the genotypic value due to the common genotype for all non-segregating loci. From Table 8.5 it can be understood that it is equal to the unweighted mean genotypic value across the 2^K complex homozygous genotypes with regard to all segregating loci. In the case of absence of linkage and absence of selection the frequency of each homozygous genotype will be $(\frac{1}{2})^K$ in F_∞. Then

$$m = E\underline{\mathcal{G}}_{F\infty} = E\underline{p}_{F\infty} \tag{8.4}$$

This implies that one may estimate m by $\overline{p}_{F\infty}$. In Section 11.2.3 the estimation of m is more extensively considered.

Because m is defined for homozygous genotypes the interpretation of m is obscure when dealing with cross-fertilizing crops. In the absence of dominance, the value of m applying to the plants of a FS-family can be estimated by the mid-parent value (see Example 9.2): all plants belonging to this family share the genetic background consisting of the homozygous complex genotype shared by the two parents. This value of m applies only to a restricted group of plants; another value of m will apply to the plants of another FS-family. The estimation of the value of m for populations consisting of mixtures of FS-families or HS-families is thus not straightforward.

At the end of this section it will be explained, by considering the F_2 generation of a self-fertilizing crop (which is identical to the offspring of a single-cross hybrid), why the probability distribution of the genotypic values for the quantitative variation of a trait tends to the **normal distribution**. For populations with different segregation ratios as well as for panmictic populations, irrespective of the allele frequencies of the segregating polygenic

loci, a similar explanation of the commonly observed tendency of a normal distribution can be developed.

The explanation can be understood by considering two models for the distribution of the genotypic values. Both models assume segregation for K unlinked, non-epistatic **isomeric** loci, *i.e.* loci with equal single-locus effects; thus $a_1 = a_2 = \ldots = a_K$ and $d_1 = d_2 = \ldots = d_K$, say a, respectively d.

- Model 1: Absence of dominance, $d = 0$
- Model 2: Presence of complete dominance: $d = a$

Model 1: Absence of dominance

In the absence of dominance the genotypic value of some genotype is a simple function of the number of B and b alleles in its complex genotype involving K relevant loci. The number of B alleles in the complex genotype is designated by $\underline{j}$ and the number of b alleles by $2K - \underline{j}$, where the random variable $\underline{j}$ may adopt any value in the range $0, 1, 2, \ldots, 2K$. The genotypic value of some random plant is:

$$\underline{\mathcal{G}} = m + (\underline{j} - K)a$$

The expected genotypic value and the genetic variance, *i.e.* the variance of the genotypic values of the plants, amount then to

$$\mathrm{E}\underline{\mathcal{G}} = m + (\mathrm{E}\underline{j} - K)a$$

and

$$\mathrm{var}(\underline{\mathcal{G}}) = a^2 \mathrm{var}(\underline{j})$$

The probability distribution for $\underline{j}$ in the F_2 population is in fact a binomial distribution, *i.e.*

$$P(\underline{j} = j) = \binom{2K}{j} \left(\frac{1}{2}\right)^j \left(\frac{1}{2}\right)^{2K-j} = \binom{2K}{j} \left(\frac{1}{4}\right)^K$$

with

$$\begin{aligned} \mathrm{E}\underline{j} &= 2K \cdot \tfrac{1}{2} = K \\ \mathrm{var}(\underline{j}) &= 2K \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{2}K \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{E}\underline{\mathcal{G}} &= m \\ \mathrm{var}(\underline{\mathcal{G}}) &= \tfrac{1}{2}Ka^2 \end{aligned}$$

The former is illustrated in Example 8.13.

Example 8.13 For $K = 4$ isomeric loci, $m = 10$, $a = 1$ and $d = 0$, the genotypic values and their probability distribution in an F_2 population are:

j	$\mathcal{G}$	$P(\underline{j} = j)$
0	6	0.0039
1	7	0.0313
2	8	0.1094
3	9	0.2188
4	10	0.2734
5	11	0.2188
6	12	0.1094
7	13	0.0313
8	14	0.0039

Then $\mathrm{E}\underline{\mathcal{G}} = 10 (= \mathrm{E}\underline{p})$ and $\mathrm{var}(\underline{\mathcal{G}}) = \frac{1}{2} \cdot 4 \cdot 1^2 = 2$.

Model 2: Presence of complete dominance

In the presence of complete dominance some complex genotype may consist of $\underline{k}$ loci with single-locus with genotype B; *i.e.* BB or Bb, and $(K - \underline{k})$ loci with single-locus genotype bb, where $\underline{k}$ may adopt any value in the range $0, 1, 2, \ldots, K$. The genotypic value of such genotype is then

$$\underline{\mathcal{G}} = m + (2\underline{k} - K)a$$

implying

$$\mathrm{E}\underline{\mathcal{G}} = m + (2\mathrm{E}\underline{k} - K)a$$

and

$$\mathrm{var}(\underline{\mathcal{G}}) = 4a^2 \mathrm{var}(\underline{k})$$

The probability distribution for $\underline{k}$ in an F_2 population is also in this case a binomial distribution, *viz.*

$$P(\underline{k} = k) = \binom{K}{k} \left(\frac{3}{4}\right)^k \left(\frac{1}{4}\right)^{K-k}$$

with

$$\mathrm{E}\underline{k} = \tfrac{3}{4}K$$

and

$$\mathrm{var}(\underline{k}) = \frac{3}{16}K$$

implying

$$\mathrm{E}\underline{\mathcal{G}} = m + (2 \cdot \tfrac{3}{4} \cdot K - K)a = m + \tfrac{1}{2}Ka$$
$$\mathrm{var}(\underline{\mathcal{G}}) = \tfrac{3}{4}Ka^2$$

Example 8.14 provides an illustration.

Example 8.14 For $K = 4$ isomeric loci, $m = 10$ and $a = d = 1$, the genotypic values and their probability distribution in an F_2 population are:

K	$\mathcal{G}$	$P(\underline{k} = k)$
0	6	0.0039
1	8	0.0469
2	10	0.2109
3	12	0.4219
4	14	0.3164

Then $\mathrm{E}\underline{\mathcal{G}} = 10+2 = 12(= \mathrm{E}\underline{p})$ and $\mathrm{var}(\underline{\mathcal{G}}) = \frac{3}{4} \cdot 4 \cdot 1^2 = 3$. Thus $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{F2}} \neq m$ in the presence of dominance. The probability distribution is skew; the modal genotypic value is 12.

The probability distribution presented in Example 8.14 is skewed. This is caused by the dominance in combination with a low value for K.

In the preceding two models the probability distributions for the genotypic values are given by the **binomial distribution**. For high values for K this distribution can be approximated by the normal distribution, because the central limit theorem states that for $K \to \infty$ the distribution of

$$\frac{\underline{j} - \mathrm{E}\underline{j}}{\sigma_j}$$

converges to the standard normal distribution $\underline{\chi}$, or N(0, 1). Thus

$$P(\underline{j} = j) = \mathrm{P}(j - \tfrac{1}{2} < \underline{j} < j + \tfrac{1}{2})$$

can be approximated by

$$P\left(\frac{j - \frac{1}{2} - \mathrm{E}\underline{j}}{\sigma_j} < \underline{\chi} < \frac{j + \frac{1}{2} - \mathrm{E}\underline{j}}{\sigma_j}\right)$$

The approximation is illustrated by Example 8.15.

Example 8.15 In Example 8.13, dealing with $K = 4$, $P(\underline{j} = 5)$ was calculated to be 0.2188. The approximation on the basis of the central limit theorem yields

$$P\left(\frac{4.5-4}{\sqrt{2}} < \underline{\chi} < \frac{5.5-4}{\sqrt{2}}\right) = P(0.354 < \underline{\chi} < 1.06) = 0.2186$$

Likewise, the distribution of the ratio $\frac{\underline{\mathcal{G}}-\mathrm{E}\underline{\mathcal{G}}}{\sigma_g}$ can be approximated by the standard normal distribution if $K \to \infty$. For model 1, assuming absence of dominance, this implies

$$\frac{\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}}{\sigma_g} = \frac{[m + (\underline{j} - k)a] - [m + (\mathrm{E}\underline{j} - k)a]}{a\sigma_j} = \frac{\underline{j} - \mathrm{E}\underline{j}}{\sigma_j} \simeq \underline{\chi}$$

The distribution of the genotypic values will thus be approximately normal, especially for higher values for K. The approximation is better as the polygenic trait is controlled by more segregating loci and/or in absence of dominance for a larger portion of the relevant loci.

8.3.3 Partitioning of Genotypic Values into their Additive Genotypic Value and their Dominance Deviation

In this book quantitative genetic theory is developed on the basis of the parameters partitioning genotypic values according to the F_∞-metric. For self-fertilizing crops the F_∞-metric is applied to partition the genotypic values of separate genotypes with the aim to derive simple expressions for $\mathrm{E}\underline{\mathcal{G}}$ and $\mathrm{var}(\underline{\mathcal{G}})$, *i.e.* the expected genotypic value and the variance of the genotypic value of the genotypes in the studied population. For cross-fertilizing crops the genotypic values may also be partitioned by the parameters of the F_∞-metric. However, an alternative system for partitioning has found general application. In this system each genotypic value is partitioned into the sum of the so-called **additive genotypic value**, here designated by the symbol γ, and the so-called **dominance deviation**, here designated by δ. Then $\mathrm{E}\underline{\mathcal{G}}$ and $\mathrm{var}(\underline{\mathcal{G}})$ may be expressed in terms of γ and δ. The components γ and δ as well as their variances will be derived in the present section.

Compared to the parameters a and d of the F_∞-metric, the components γ and δ have an important drawback: they are **frequency-dependent** (see Note 8.4). Thus, for a given genotype, their values change if the frequency of that genotype changes. They change if the locus affects a trait subjected to selection! The components γ and δ, which will be described in terms of a and

d, are thus functions of the allele frequencies. Notwithstanding this drawback, attention is given to the development of quantitative genetic theory of cross-fertilizing crops on the basis of the components γ and δ. Application of this partitioning in the case of multiple allelism, which should be anticipated for cross-fertilizing crops, is straightforward. Multiple allelism is to be expected in populations of cross-fertilizing crops. Presence of only two alleles for a certain locus is then a special case, which occurs – for example – in the generations tracing back to a single cross hybrid.

Note 8.4 Frequency-dependent components of the genotypic value describing epistasis have also been elaborated (Cockerham, 1954; Kempthorne, 1957; Weber, 1978). The partitioning of the genotypic values occurs in a way similar to the so-called least squares method of estimation in linear regression. Thus the variance of interaction components is minimized, implying that the additive genetic variance is maximized. The relative size of the so-called **interaction variance** leads then to an underestimation of the relative importance of the contribution of the epistatic component to the genotypic value (see also Note 8.3).

The partitioning gives rise to the important concepts of **breeding value** (Section 8.3.4), a quantity closely related to the additive genotypic value, and that of **additive genetic variance**, which is the variance of the additive genotypic values. The latter is an important yardstick for the perspectives of further improvement of the expected genotypic value by means of selection.

The partitioning of a genotypic value is into the **additive genotypic value** (γ) and the **dominance deviation** (δ). (For the simple case of two alleles these components of $\mathcal{G}$ will also be expressed in terms of the F_∞-metric parameters a and d). In this section the components of the genotypic value and of the genotypic variance will be considered for only one segregating locus. The conditions required for a straightforward extension of the derived expressions to the case of K segregating loci are discussed in Section 10.1.

Multiple alleles, random mating

First the partitioning of the genotypic values of the genotypes occurring with regard to the multiple allelic locus B_1-B_2-$\cdots$-B_n, with allele frequencies $p_1, p_2, \cdots, p_n$, is considered.

In the present section the genotypic value $\mathcal{G}_{ij}$ of some genotype B_iB_j is partitioned according to the commonly used linear model for data in a two-way table. Absence of reciprocal differences is assumed. This implies that it is irrelevant whether allele B_i entered the genotype via an egg or via a pollen grain. This assumption gives rise to the following linear model for $\mathcal{G}_{ij}$:

$$\mathcal{G}_{ij} = \mu + \alpha_i + \alpha_j + \delta_{ij}; i, j = 1, \ldots, n$$

where

$$\begin{aligned}\mu &= \mathrm{E}\underline{\mathcal{G}} = \text{ the expected genotypic value}\\ \alpha_i &= \text{the main effect of allele } B_i\\ \alpha_j &= \text{the main effect of allele } B_j\\ \delta_{ij} &= \text{the effect of intra-locus interaction of alleles } B_i \text{ and } B_j.\end{aligned}$$

In the present context the main effects are called **allele effects** (or '**average effects**'; or **additive effects**) and the intra-locus interaction effects are called: **dominance deviations**.

Some of the derivations following hereafter simplify when considering

$$\mathcal{G}_{ij}{}' = \mathcal{G}_{ij} - \mu$$

where $\mathcal{G}_{ij}{}'$ represents the so-called reduced genotypic value. For this reason μ is first derived.

The genotypic composition of the population due to a single round of panmictic reproduction follows from the two-way table below. The vertical margins of the table present the haplotypic composition of the eggs; the horizontal margins present the haplotypic composition of the pollen; the central part provides the genotypic composition of the obtained population.

		Haplotypic composition of the pollen							
		B_1		B_2		$\ldots$	B_n		
Haplotypic composition of the eggs	B_1	${p_1}^2$	B_1B_1	p_1p_2	B_1B_2		p_1p_n	B_1B_n	p_1
	B_2	p_2p_1	B_2B_1	${p_2}^2$	B_2B_2		p_2p_n	B_2B_n	p_2
	.								.
	B_n	p_np_1	B_nB_1	p_np_2	B_nB_2		${p_n}^2$	B_nB_n	p_n
		p_1		p_2		$\cdots$	p_n		1

Application of the representation of the genotypic composition used in Section 2.2.2, for $i = 1, \ldots, n$ and $j = i, \ldots, n$:

	Genotype				
	B_1B_1	$\ldots$	B_iB_j	$\ldots$	B_nB_n
f	${p_1}^2$		$2p_ip_j$		${p_n}^2$
$\mathcal{G}$	$\mathcal{G}_{11}$		$\mathcal{G}_{ij}$		$\mathcal{G}_{nn}$

yields the following expression for the expected genotypic value

$$\mu = \mathrm{E}\underline{\mathcal{G}} = {p_1}^2\mathcal{G}_{11} + \ldots + 2p_ip_j\mathcal{G}_{ij} + \ldots + {p_n}^2\mathcal{G}_{nn}$$

When deriving $\mathrm{E}\underline{\mathcal{G}}^2$ in a similar way, one may calculate the variance of the genotypic values in the following way:

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{E}\underline{\mathcal{G}}^2 - \mu^2$$

(The concepts 'expected genotypic value' and 'genotypic variance' are extensively discussed in Chapter 9 and 10, respectively). With regard to the reduced genotypic values we get:

$$\mathrm{E}\underline{\mathcal{G}}' = \mathrm{E}(\underline{\mathcal{G}} - \mu) = 0$$
$$\mathrm{var}(\underline{\mathcal{G}}') = \mathrm{var}(\underline{\mathcal{G}}) = \mathrm{E}\underline{\mathcal{G}}'^2 - (\mathrm{E}\underline{\mathcal{G}}')^2 = \mathrm{E}\underline{\mathcal{G}}'^2$$

The main effect of allele B_i is defined to be equal to the (conditional) expectation of the reduced genotypic value of plants containing allele B_i. Thus

$$\alpha_i = \mathrm{E}\left(\underline{\mathcal{G}}'_{ij}|B_i\right) = p_1\mathcal{G}_{i1}{}' + p_2\mathcal{G}_{i2}{}' + \cdots + p_n\mathcal{G}_{in}{}' = \sum_{j=1}^{n} p_j\mathcal{G}_{ij}{}' = \sum_{j=1}^{n} p_j\mathcal{G}_{ji}{}' \tag{8.5}$$

The **breeding value** (bv) of genotype B_iB_j is now defined as the sum of the effects of the alleles present in the genotype. Thus

$$bv_{ij} := \alpha_i + \alpha_j$$

The **additive genotypic value** (γ) of genotype B_iB_j is defined as: $\mathrm{E}\underline{\mathcal{G}}$ plus its **breeding value**. Thus

$$\gamma_{ij} := \mu + bv_{ij} = \mu + \alpha_i + \alpha_j \tag{8.6}$$

The expected value of the main effect of an allele, calculated across all alleles belonging to the involved locus, is calculated as follows:

$$\mathrm{E}\underline{\alpha} = p_1\alpha_1 + \cdots + p_n\alpha_n = p_1\left(\sum_{j=1}^{n} p_j\mathcal{G}_{1j}{}'\right) + \cdots + p_n\left(\sum_{j=1}^{n} p_j\mathcal{G}'_{nj}\right)$$

Thus

$$\mathrm{E}\underline{\alpha} = p_1p_1\mathcal{G}_{11}{}' + p_1p_2\mathcal{G}_{12}{}' + \ldots + p_np_{n-1}\mathcal{G}_{nn-1}{}' + p_np_n\mathcal{G}_{nn}{}' = \mathrm{E}\underline{\mathcal{G}}' = 0 \tag{8.7}$$

This implies $\mathrm{E}\underline{\gamma} = \mu$.

The **dominance deviation** of a genotype is defined to be equal to the difference between its genotypic value and its additive genotypic value. The dominance deviation of genotype B_iB_j is thus:

$$\delta_{ij} := \mathcal{G}_{ij} - \gamma_{ij} = \mathcal{G}_{ij} - (\mathrm{E}\underline{\mathcal{G}} + \alpha_i + \alpha_j) = \mathcal{G}_{ij}{}' - \alpha_i - \alpha_j \tag{8.8}$$

The expected value of $\underline{\delta}$ across all genotypes for the considered locus is equal to

$$\mathrm{E}\underline{\delta} = \mathrm{E}[\underline{\mathcal{G}} - (\mathrm{E}\underline{\mathcal{G}} + \underline{\alpha} + \underline{\alpha})] = 0$$

Altogether the pursued partitioning of the genotypic value of genotype B_iB_j is

$$\mathcal{G}_{ij} = \gamma_{ij} + \delta_{ij}$$

In general

$$\underline{\mathcal{G}} = \underline{\gamma} + \underline{\delta} \tag{8.9}$$

Example 8.16 illustrates the present partitioning for locus $B - b - \beta$.

Example 8.16 A population with the Hardy–Weinberg genotypic composition with regard to locus B-b-β, where $p_B = \frac{1}{2}, p_b = \frac{1}{4}$ and $p_\beta = \frac{1}{4}$, is considered.

	Genotype					
	BB	bb	$\beta\beta$	Bb	$B\beta$	$b\beta$
f	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$
$\mathcal{G}$	10	8	6	10	9	7

Thus

$$\mu = \frac{1}{4} \times 10 + \cdots + \frac{1}{8} \times 7 = 9, \mathrm{E}\underline{\mathcal{G}}^2 = \frac{1}{4} \times 10^2 + \cdots + \frac{1}{8} \times 7^2 = 82.625, \text{and } \sigma_g^2 = 82.625 - 9^2 = 1.625$$

The two-way table below describes the origin of the population: the horizontal margins and the vertical margins present the haplotypic compositions of the gametes underlying the genotypes, the central part presents the genotypes and their reduced genotypic values $\mathcal{G}' = \mathcal{G} - \mu = \mathcal{G} - 9$.

		Haplotypic composition of the pollen						
		B		b		β		
Haplotypic composition	B	BB	1	Bb	1	$B\beta$	0	$\frac{1}{2}$
of the eggs:	b	Bb	1	bb	-1	$b\beta$	-2	$\frac{1}{4}$
	β	$B\beta$	0	$b\beta$	-2	$\beta\beta$	-3	$\frac{1}{4}$
		$\frac{1}{2}$		$\frac{1}{4}$		$\frac{1}{4}$		1

The main effects of alleles B, b and β are calculated from this table in the following way:

$$\begin{aligned}
\alpha_B &= \tfrac{1}{2} \times 1 + \tfrac{1}{4} \times 1 + \tfrac{1}{4} \times 0 = \tfrac{3}{4} \\
\alpha_b &= \tfrac{1}{2} \times 1 + \tfrac{1}{4} \times (-1) + \tfrac{1}{4} \times (-2) = -\tfrac{1}{4} \\
\alpha_\beta &= \tfrac{1}{2} \times 0 + \tfrac{1}{4} \times (-2) + \tfrac{1}{4} \times (-3) = -1\tfrac{1}{4} \\
\text{Check } \mathrm{E}\underline{\alpha} &= \tfrac{1}{2} \times \tfrac{3}{4} + \tfrac{1}{4} \times (-\tfrac{1}{4}) + \tfrac{1}{4} \times (-1\tfrac{1}{4}) = 0
\end{aligned}$$

After having determined the allele effects one can partition the genotypic values:

	Genotype					
	BB	bb	$\beta\beta$	Bb	$B\beta$	$b\beta$
f	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$
$\mathcal{G}$	10	8	6	10	9	7
γ	10.5	8.5	6.5	9.5	8.5	7.5
δ	-0.5	-0.5	-0.5	0.5	0.5	-0.5

The variance of the additive genotypic values is called **additive genetic variance**, usually designated by σ_a^2. It is equal to

$$\mathrm{var}(\underline{\gamma}) = \mathrm{var}(\mathrm{E}\underline{\mathcal{G}} + \underline{\alpha} + \underline{\alpha}) = 2\mathrm{var}(\underline{\alpha}) = 2\mathrm{E}\underline{\alpha}^2 \tag{8.10}$$

(Because of random fusion of female and male gametes the effects of the maternal and paternal alleles are uncorrelated. Their covariance is then zero.) The additive genetic variance, *i.e.* the variance of the additive genotypic values, is thus twice the variance of the main effects of the alleles.

The variance of the dominance deviations, usually called **dominance variance** and designated by σ_d^2, is equal to $\mathrm{E}\underline{\delta}^2$.

The variance of the genotypic values, usually called **genetic variance** and designated by σ_g^2, is

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\gamma} + \underline{\delta}) = \mathrm{var}(\underline{\gamma}) + \mathrm{var}(\underline{\delta}) + 2\mathrm{cov}(\underline{\gamma}, \underline{\delta}).$$

In Note 8.5 it is shown that $\mathrm{cov}(\underline{\gamma}, \underline{\delta}) = 0$. This implies

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\gamma}) + \mathrm{var}(\underline{\delta}) \tag{8.11}$$

Note 8.5 The covariance of the additive genotypic value and the dominance deviation can be shown to be zero:

$$\mathrm{cov}(\underline{\gamma}, \underline{\delta}) = \mathrm{cov}(\underline{\gamma} - \mu, \underline{\mathcal{G}} - \underline{\gamma}) = \mathrm{E}[(\underline{\gamma} - \mu) \cdot (\underline{\mathcal{G}} - \underline{\gamma})]$$

as

$$[\mathrm{E}(\underline{\gamma} - \mu)] \cdot [\mathrm{E}(\underline{\mathcal{G}} - \underline{\gamma})] = 0$$

Thus

$$\begin{aligned}\mathrm{cov}(\underline{\gamma}, \underline{\delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\alpha_i + \alpha_j)\left(\mathcal{G}_{ij}{}' - \alpha_i - \alpha_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \alpha_i \mathcal{G}_{ij}{}' + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \alpha_j \mathcal{G}_{ij}{}' - \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j (\alpha_i + \alpha_j)^2\end{aligned}$$

As

$$\alpha_i + \alpha_j = \gamma_{ij} - \mu = \gamma_{ij} - \mathrm{E}\underline{\gamma}$$

the last term is equal to

$$\mathrm{E}(\underline{\gamma} - \mathrm{E}\underline{\gamma})^2 = \mathrm{var}(\underline{\gamma})$$

Thus

$$\begin{aligned}\mathrm{cov}(\underline{\gamma},\underline{\delta}) &= \sum_{i=1}^{n} p_i\alpha_i\left(\sum_{j=1}^{n} p_j\mathcal{G}_{ij}{}'\right) + \sum_{j=1}^{n} p_j\alpha_j\left(\sum_{i=1}^{n} p_i\mathcal{G}_{ij}{}'\right) - \mathrm{var}(\underline{\gamma}) \\ &= \sum_{i=1}^{n} p_i{\alpha_i}^2 + \sum_{j=1}^{n} p_j{\alpha_j}^2 - \mathrm{var}(\underline{\gamma}) = 2\mathrm{E}\underline{\alpha}^2 - \mathrm{var}(\underline{\gamma}) = 0.\end{aligned}$$

Example 8.17 illustrates the calculation of the genetic variance and its components for the situation of Example 8.16.

Example 8.17 For the population described in Example 8.16, the additive genotypic variance amounts to:

$$\mathrm{var}(\underline{\gamma}) = \tfrac{1}{4}\times(10.5)^2 + \cdots + \tfrac{1}{8}\times(7.5)^2 - 9^2 = 1.375$$

This is indeed equal to

$$\begin{aligned}2\mathrm{E}(\underline{\alpha})^2 &= 2\left[\tfrac{1}{2}\times(\tfrac{3}{4})^2 + \tfrac{1}{4}\times(-\tfrac{1}{4})^2 + \tfrac{1}{4}\times\left(-1\frac{1}{4}\right)^2\right] \\ &= 2\times 0.6875 = 1.375.\end{aligned}$$

As

$$\mathrm{E}\underline{\delta} = \tfrac{1}{4}\times(-0.5) + \cdots + \tfrac{1}{8}\times(-0.5) = 0$$

the dominance variance is equal to:

$$\mathrm{var}(\underline{\delta}) = \tfrac{1}{4}\times(-0.5)^2 + \cdots + \tfrac{1}{8}\times(-0.5)^2 = 0.25.$$

It is thus confirmed that $\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\gamma}) + \mathrm{var}(\underline{\delta})$. This follows also from the fact that the covariance of $\underline{\gamma}$ and $\underline{\delta}$, *i.e.*

$$\begin{aligned}\mathrm{cov}(\underline{\gamma},\underline{\delta}) &= \mathrm{E}(\underline{\gamma}\cdot\underline{\delta}) = \tfrac{1}{4}\times 10.5\times(-0.5) + \tfrac{1}{16} \\ &\quad\times 8.5\times(-0.5) + \cdots + \tfrac{1}{8}\times 7.5\times(-0.5)\end{aligned}$$

is equal to 0.

The partitioning developed here may seem rather abstract. In practice, however, the additive genotypic value can be estimated rather easily. Consider, for example, the result of open pollination of a plant with genotype B_iB_j

		Haplotypic composition of the pollen							
		B_1		B_2		...	B_n		Expected genotypic
		p_1		p_2			p_n		value of the offspring
Haplotype	B_i	p_1	B_iB_1	p_2	B_iB_2		p_n	B_iB_n	$\mu + \alpha_i$
of the egg:	B_j	p_1	B_jB_1	p_2	B_jB_2		p_n	B_jB_n	$\mu + \alpha_j$

The expected genotypic value of the offspring due to open pollination of a plant with genotype B_iB_j is thus equal to

$$\mathrm{E}(\underline{\mathcal{G}}|B_iB_j) = \mu + \tfrac{1}{2}\alpha_i + \tfrac{1}{2}\alpha_j = \tfrac{1}{2}\mu + \tfrac{1}{2}\gamma_{ij}$$

This implies that

$$\gamma_{ij} = 2\mathrm{E}(\underline{\mathcal{G}}|B_iB_j) - \mu,$$

i.e. that

$$\gamma_{ij} - \mu = \alpha_i + \alpha_j = 2[\mathrm{E}(\underline{\mathcal{G}}|B_iB_j) - \mu] \tag{8.12}$$

Earlier in this section, the latter quantity was defined as the breeding value of genotype B_iB_j (see also Section 8.3.4).

An unbiased estimate of γ_{ij}, *i.e.* the additive genotypic value of an open pollinated plant with genotype B_iB_j, is thus twice the mean phenotypic value of its offspring minus the mean phenotypic value of all plants in the (offspring) population:

$$\hat{\gamma}_{ij} = 2\overline{p}_{HS_{ij}} - \overline{p}$$

The difference between an unbiased estimate of the genotypic value of this plant and the unbiased estimate of its additive genotypic value is an unbiased estimate of its dominance deviation δ_{ij}:

$$\hat{\delta}_{ij} = \hat{\mathcal{G}} - \hat{\gamma}_{ij}$$

The difference between the expected genotypic values of the plants belonging to the HS-families obtained after open pollination of two different plants, with genotypes B_iB_j and B_kB_l, is equal to half the difference between the additive genotypic values of these plants:

$$\mathrm{E}(\underline{\mathcal{G}}|B_iB_j) - \mathrm{E}(\underline{\mathcal{G}}|B_kB_l) = \tfrac{1}{2}(\gamma_{ij} - \gamma_{kl})$$

As $\mathrm{cov}(\underline{\gamma}, \underline{\delta}) = 0$ (see Note 8.5), the covariance of the genotypic value of an open pollinated (maternal) plant ($\underline{\mathcal{G}}_M$) and the expected genotypic value of the members of the HS-family produced by this plant ($\underline{\mathcal{G}}_{HS|M}$) is

$$\mathrm{cov}(\underline{\mathcal{G}}_M, \underline{\mathcal{G}}_{HS|M}) = \mathrm{cov}(\underline{\gamma} + \underline{\delta}, \tfrac{1}{2}\mu + \tfrac{1}{2}\underline{\gamma}) = \tfrac{1}{2}\mathrm{var}(\underline{\gamma}) = \tfrac{1}{2}\sigma_a^2 \tag{8.13}$$

Two alleles, random mating

Early in this section it was said that, in the simple case of two alleles per segregating locus, the additive genotypic value (γ) and the dominance deviation (δ) can be expressed in terms of the F_∞-metric parameters a and d. This will now be elaborated.

Locus B-b, with allele frequencies p and q, is considered for a population with the Hardy–Weinberg genotypic composition. This population originates from random combination of female and male gametes according to the following scheme:

		Haplotypic composition of the pollen		
		b	B	
Haplotypic composition	b	q^2bb	$qpBb$	q
of the eggs:	B	$pqBb$	p^2BB	p
		q	p	1

Thus

	Genotype		
	bb	Bb	BB
f	q^2	$2pq$	p^2
$\mathcal{G}$	$m-a$	$m+d$	$m+a$

The expected genotypic value is

$$\begin{aligned} \mathrm{E}\underline{\mathcal{G}} &= q^2(m-a) + 2pq(m+d) + p^2(m+a) \\ &= m + (p^2 - q^2)a + 2pqd = m + (p-q)a + 2pqd \end{aligned} \tag{8.14}$$

The effects of alleles b and B are

$$\begin{aligned} \alpha_b &= q(m-a) + p(m+d) - [m + (p-q)a + 2pqd] \\ &= -qa + pd - (p-q)a - 2pqd = -pa + (p - 2pq)d \\ &= -p[a - (p-q)d] \end{aligned} \tag{8.15}$$

and

$$\begin{aligned} \alpha_B &= q(m+d) + p(m+a) - [m + (p-q)a + 2pqd] \\ &= qd + pa - pa + qa - 2pqd = qa + (q - 2pq)d \\ &= q[a - (p-q)d] \end{aligned} \tag{8.16}$$

Half the difference between the additive genotypic values of the homozygous genotypes BB and bb amounts to

$$\tfrac{1}{2}(\gamma_{BB} - \gamma_{bb}) = \alpha_B - \alpha_b = (q+p)[a - (p-q)d] = a - (p-q)d \tag{8.17}$$

For panmictic populations this expression indicates the so-called 'average effect of an allele substitution', *viz.* substitution of allele b by allele B. It

is designated by α_{RM}. It occurs in many relevant mathematical expressions derived in quantitative genetic theory applying to the situation of $n = 2$ alleles representing the considered locus.

As $\alpha_b = -p\alpha_{\mathrm{RM}}$ and $\alpha_B = q\alpha_{\mathrm{RM}}$, the following partitioning of the genotypic values is obtained:

	Genotype		
	bb	Bb	BB
f	q^2	$2pq$	p^2
j	0	1	2
$\mathcal{G}$	$m - a$	$m + d$	$m + a$
γ	$\mu - 2p\alpha_{\mathrm{RM}}$	$\mu - (p - q)\alpha_{\mathrm{RM}}$	$\mu + 2q\alpha_{\mathrm{RM}}$
δ	$m - a - [\mu - 2p\alpha_{\mathrm{RM}}]$	$m + d - [\mu - (p - q)\alpha_{\mathrm{RM}}]$	$m + a - [\mu + 2q\alpha_{\mathrm{RM}}]$

It appears that $\underline{\gamma}$ is equal to $\mu + (\underline{j} - 2p)\alpha_{\mathrm{RM}}$, *i.e.*

$$\underline{bv} = \underline{\gamma} - \mu = (\underline{j} - 2p)\alpha_{\mathrm{RM}} = (\underline{j} - 2p)[a - (p - q)d] \tag{8.18}$$

This implies that $\mathrm{var}(\underline{bv}) = \mathrm{var}(\underline{\gamma}) = {\sigma_a}^2$.

Note 8.7 shows that $\mathrm{var}(\underline{j}) = 2pq$ in the case of random mating. The additive genetic variance amounts thus to

$$\mathrm{var}(\underline{\gamma}) = \alpha_{\mathrm{RM}}^2 \mathrm{var}(\underline{j}) = 2pq\alpha_{\mathrm{RM}}^2$$

The partitioning is illustrated in Example 8.18.

Example 8.18 The following panmictic population is considered:

	Genotype		
	bb	Bb	BB
f	0.36	0.48	0.16
$\mathcal{G}$	11.5	13.5	13.5

Thus $p = 0.4, q = 0.6, m = 12.5, a = d = 1$, *i.e.* complete dominance.

$$\begin{aligned} \mu &= 0.36 \times 11.5 + 0.48 \times 13.5 + 0.16 \times 13.5 = 12.78 \\ \mathrm{var}(\underline{\mathcal{G}}) &= 0.36(11.5)^2 + 0.64(13.5)^2 - (12.78)^2 = 0.9216 \end{aligned}$$

Because

$$\alpha_{\mathrm{RM}} = a - (p - q)d = 1 - (0.4 - 0.6) \times 1 = 1.2$$

it follows that

$$\begin{aligned} \alpha_b &= -p\alpha_{\mathrm{RM}} = -0.4 \times 1.2 = -0.48 \\ \alpha_B &= q\alpha_{\mathrm{RM}} = 0.6 \times 1.2 = 0.72 \end{aligned}$$

The genotypic values are then partitioned in the following way:

	Genotype		
	bb	*Bb*	*BB*
f	0.36	0.48	0.16
$\mathcal{G}$	11.5	13.5	13.5
γ	$12.78 + 2 \times (-0.48) = 11.82$	$12.78 - 0.48$ $+0.72 = 13.02$	$12.78 + 2$ $\times 0.72 = 14.22$
δ	$11.5 - 11.82 = -0.32$	$13.5 - 13.02$ $= 0.48$	$13.5 - 14.22$ $= -0.72$

Thus

$$\mathrm{var}(\underline{\gamma}) = 0.36(11.82)^2 + 0.48(13.02)^2 + 0.16(14.22)^2 - (12.78)^2 = 0.6912$$

which is equal to

$$2pq\alpha_{\mathrm{RM}}^2 = 2(0.4)(0.6)(1.2)^2$$

Two alleles, inbreeding

Section 2.1.1 specified situations where only two alleles per locus segregate. This is especially to be expected in the case of continued selfing starting in an F_1. In Note 8.6 it is derived that the allele effects, expressed in terms of the F_∞-metric parameters a and d, are then follows:

$$\alpha_b = -p\left[a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right] \tag{8.19}$$

$$\alpha_B = q\left[a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right] \tag{8.20}$$

Note 8.6 An inbred population may be described as follows:

	Genotype		
	bb	*Bb*	*BB*
f	$q^2 + pq\mathscr{F}$	$2pq(1-\mathscr{F})$	$p^2 + pq\mathscr{F}$
$\mathcal{G}$	$m - a$	$m + d$	$m + a$
γ	$\mu + 2\alpha_b$	$\mu + \alpha_b + \alpha_B$	$\mu + 2\alpha_B$

where

$$\mu = m + (-q^2 - pq\mathscr{F} + p^2 + pq\mathscr{F})a + 2pq(1-\mathscr{F})d = m + (p-q)a + 2pq(1-\mathscr{F})d$$

The additive genotypic values are fitted to the genotypic values in such a way, that the expected value of the square of the deviations is minimal. Thus:

$$\begin{aligned}\mathrm{E}(\underline{\mathcal{G}}-\underline{\gamma})^2 &= (q^2+pq\mathscr{F})(m-a-\mu-2\alpha_b)^2+2pq(1-\mathscr{F})\\ &\quad\times(m+d-\mu-\alpha_b-\alpha_B)^2+(p^2+pq\mathscr{F})(m+a-\mu-2\alpha_B)^2\end{aligned}$$

is minimal for the values assigned to α_b and α_B. The derivatives of $\mathrm{E}(\underline{\mathcal{G}}-\underline{\gamma})^2$ to α_b and α_B are then zero, *i.e.*

$$-4(q^2+pq\mathscr{F})(m-a-\mu-2\alpha_b)-4pq(1-\mathscr{F})(m+d-\mu-\alpha_b-\alpha_B)=0,$$

and

$$-4pq(1-\mathscr{F})(m+d-\mu-\alpha_b-\alpha_B)-4(p^2+pq\mathscr{F})(m+a-\mu-2\alpha_B)=0$$

or

$$\begin{aligned}&8(q^2+pq\mathscr{F})\alpha_b+4pq(1-\mathscr{F})(\alpha_b+\alpha_B)\\ &\quad=4(q^2+pq\mathscr{F})(m-a-\mu)+4pq(1-\mathscr{F})(m+d-\mu),\end{aligned}\tag{a}$$

and

$$\begin{aligned}&4pq(1-\mathscr{F})(\alpha_b+\alpha_B)+8(p^2+pq\mathscr{F})\alpha_B\\ &\quad=4pq(1-\mathscr{F})(m+d-\mu)+4(p^2+pq\mathscr{F})(m+a-\mu)\end{aligned}\tag{b}$$

Summation of equations (a) and (b) yields on the right hand side:

$$\begin{aligned}&4[(q^2+pq\mathscr{F})(m-a-\mu)+8pq(1-\mathscr{F})(m+d-\mu)\\ &\quad+4(p^2+pq\mathscr{F})(m+a-\mu)]=4[\mu-\mu]=0,\end{aligned}$$

and on the left hand side:

$$8\alpha_b[q^2+pq\mathscr{F}+pq(1-\mathscr{F})]+8\alpha_B[pq(1-\mathscr{F})+p^2+pq\mathscr{F}]=8(q\alpha_b+p\alpha_B)$$

This implies

$$\mathrm{E}\underline{\alpha}=q\alpha_b+p\alpha_B=0$$

Division of equations (a) and (b) by $4q$ and $4p$, respectively, yields

$$\begin{aligned}&\alpha_b[2q+2p\mathscr{F}+p(1-\mathscr{F})]+\alpha_B p(1-\mathscr{F})\\ &\quad=(q+p\mathscr{F})(m-a-\mu)+p(1-\mathscr{F})(m+d-\mu),\end{aligned}$$

and

$$\begin{aligned}&\alpha_b q(1-\mathscr{F}) + \alpha_B[q(1-\mathscr{F}) + 2p + 2q\mathscr{F}]\\ &\quad = q(1-\mathscr{F})(m+d-\mu) + (p+q\mathscr{F})(m+a-\mu)\end{aligned}$$

As

$$2q + p\mathscr{F} + p = 1 + q + (1-q)\mathscr{F} = 1 + \mathscr{F} + (1-\mathscr{F})q,$$

and

$$q + 2p + q\mathscr{F} = 1 + p + (1-p)\mathscr{F} = 1 + \mathscr{F} + (1-\mathscr{F})p,$$

these equations can be rewritten as:

$$\begin{aligned}&\alpha_b(1+\mathscr{F}) + (1-\mathscr{F})(q\alpha_b + p\alpha_B)\\ &\quad = (q + p\mathscr{F} + p - p\mathscr{F})m - (q + p\mathscr{F})a\\ &\qquad + p(1-\mathscr{F})d - [m + (p-q)a + 2pq(1-\mathscr{F})d],\end{aligned}$$

and

$$\begin{aligned}&\alpha_B(1+\mathscr{F}) + (1-\mathscr{F})(q\alpha_b + p\alpha_B)\\ &\quad = (q - p\mathscr{F} + p + p\mathscr{F})m + (p + q\mathscr{F})a\\ &\qquad + q(1-\mathscr{F})d - [m + (p-q)a + 2pq(1-\mathscr{F})d],\end{aligned}$$

i.e. as

$$\begin{aligned}\alpha_b(1+\mathscr{F}) &= -(q + p\mathscr{F} + p - q)a + p(1-\mathscr{F})(1-2q)d\\ &= -p(1+\mathscr{F})a + p(p-q)(1-\mathscr{F})d,\end{aligned}$$

and

$$\begin{aligned}\alpha_B(1+\mathscr{F}) &= (p + q\mathscr{F} - p + q)a + q(1-\mathscr{F})(1-2p)d\\ &= q(1+\mathscr{F})a - q(p-q)(1-\mathscr{F})d,\end{aligned}$$

respectively.

The allele effects giving the minimum value of $\mathrm{E}(\underline{G} - \underline{\gamma})^2$ are thus:

$$\alpha_b = -p\left[a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right] \text{ and } \alpha_B = q\left[a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right].$$

This still implies that

$$\mathrm{E}\underline{\alpha} = q\alpha_b + p\alpha_B = 0$$

For an inbred population the 'average effect of the gene substitution' ($\alpha_{\mathscr{F}}$) amounts to

$$\alpha_{\mathscr{F}} = \alpha_B - \alpha_b = a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d \tag{8.21}$$

We have now arrived at the situation where the inbred population can be described as follows:

	Genotype		
	bb	Bb	BB
f	$q^2 + pq\mathscr{F}$	$2pq(1-\mathscr{F})$	$p^2 + pq\mathscr{F}$
j	0	1	2
$\mathcal{G}$	$m-a$	$m+d$	$m+a$
γ	$\mu + 2\alpha_b + 0(\alpha_B - \alpha_b)$	$\mu + 2\alpha_b + 1(\alpha_B - \alpha_b)$	$\mu + 2\alpha_b + 2(\alpha_B - \alpha_b)$

This scheme shows that

$$\underline{\gamma} = \mu + 2\alpha_b + \underline{j}\alpha_{\mathscr{F}}$$

In Note 8.7 it is derived that

$$\mathrm{var}(\underline{j}) = 2pq(1+\mathscr{F})$$

thus

$$\mathrm{var}(\underline{\gamma}) = \sigma^2_{a\mathscr{F}} = 2pq(1+\mathscr{F})\alpha^2_{\mathscr{F}}$$

As

$$\begin{aligned}
\alpha_{\mathscr{F}} &= \left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d + a - \left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)a \\
&= \left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)\alpha_{\mathrm{RM}} + \frac{(1+\mathscr{F})a - (1-\mathscr{F})a}{1+\mathscr{F}} \\
&= \left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)\alpha_{\mathrm{RM}} + \left(\frac{2\mathscr{F}}{1+\mathscr{F}}\right)a \\
&= \alpha_{\mathrm{RM}} + \frac{1}{1+\mathscr{F}}\left(2\mathscr{F}a + (1-\mathscr{F})\alpha_{\mathrm{RM}} - (1+\mathscr{F})\alpha_{\mathrm{RM}}\right) \\
&= \alpha_{\mathrm{RM}} + \frac{2\mathscr{F}}{1+\mathscr{F}}(a - \alpha_{\mathrm{RM}}) = \alpha_{\mathrm{RM}} + \frac{2\mathscr{F}}{1+\mathscr{F}}(p-q)d
\end{aligned}$$

it follows that

$$\alpha_F = \alpha_{\mathrm{RM}}$$

if $\mathscr{F} = 0$, if $d = 0$, or if $p = q = \frac{1}{2}$.

The equation

$$\sigma_{a\mathscr{F}}{}^2 = (1+\mathscr{F})\sigma_a{}^2$$

applies thus only if $p = q = \frac{1}{2}$.

In Note 8.7 it is shown that $\mathrm{cov}(\underline{\gamma}, \underline{\delta}) = 0$ also applies in the case of inbreeding. The partitioning

$$\underline{\mathcal{G}} = \underline{\gamma} + \underline{\delta}$$

implies then

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\gamma}) + \mathrm{var}(\underline{\delta})$$

Expressions for $\mathrm{var}(\underline{\mathcal{G}}), \mathrm{var}(\underline{\gamma})$ and $\mathrm{var}(\underline{\delta})$ in terms of the parameters a and d are also derived in Note 8.7. This gives

$$var(\underline{\gamma}) = 2pq(1+\mathscr{F})\left[a-(p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right]^2 \text{ and} \tag{8.22}$$

and

$$var(\underline{\delta}) = 4pq\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d^2\left[\mathscr{F}+pq(1-\mathscr{F})^2\right] \tag{8.23}$$

Note 8.7 The following scheme allows the determination of a few important quantitative genetic parameters:

	Genotype		
	bb	Bb	BB
f	$f_0 = q^2 + pq\mathscr{F}$	$f_1 = 2pq(1-\mathscr{F})$	$f_2 = p^2 + pq\mathscr{F}$
$\mathcal{G}$	$m - a$	$m + d$	$m + a$
j	0	1	2
γ	$\mu + 2\alpha_b$	$\mu + 2\alpha_b + \alpha_{\mathscr{F}}$	$\mu + 2\alpha_b + 2\alpha_{\mathscr{F}}$
δ	$\mathcal{G}_{bb} - \mu - 2\alpha_b$	$\mathcal{G}_{Bb} - \mu - 2\alpha_b - \alpha_{\mathscr{F}}$	$\mathcal{G}_{BB} - \mu - 2\alpha_b - 2\alpha_{\mathscr{F}}$

The scheme shows that

$$\underline{\gamma} = \mu + 2\alpha_b + \underline{j}\alpha_{\mathscr{F}}$$

and that

$$\underline{\delta} = \mathcal{G} - \mu - 2\alpha_b + \underline{j}\alpha_{\mathscr{F}}$$

Thus

$$\mathrm{cov}(\underline{\gamma}, \underline{\delta}) = \mathrm{cov}(\underline{j}\alpha_{\mathscr{F}}, \underline{\mathcal{G}} - \underline{j}\alpha_{\mathscr{F}}) = -{\alpha_{\mathscr{F}}}^2 \mathrm{var}(\underline{j}) + \alpha_{\mathscr{F}} \mathrm{cov}(\underline{j}, \underline{\mathcal{G}})$$

The quantity $\mathrm{cov}(\underline{\gamma}, \underline{\delta})$ is obtained via derivations of $\mathrm{var}(\underline{j})$ and $\mathrm{cov}(\underline{j}, \underline{\mathcal{G}})$:

$$\begin{aligned}
\mathrm{var}(\underline{j}) &= \mathrm{E}\underline{j}^2 - (\mathrm{E}\underline{j})^2 = f_1 + 4f_2 - (f_1 + 2f_2)^2 \\
&= 2p + 2f_2 - (2p)^2 = 2f_2 + 2p(1-2p) = 2f_2 - 2p(p-q) \\
&= 2p^2 + 2pq\mathscr{F} - 2p^2 + 2pq = 2pq(1+\mathscr{F}) \\
\mathrm{cov}(\underline{j}, \underline{\mathcal{G}}) &= \mathrm{E}(\underline{j}\cdot\underline{\mathcal{G}}) - (\mathrm{E}\underline{j})\mu = f_1(m+d) + 2f_2(m+a) \\
&\quad - [2p][m + (f_2 - f_0)a + f_1 d] \\
&= (f_1 + 2f_2)m + f_1 d + 2f_2 a - [2p][m + (f_2 - f_0)a + f_1 d]
\end{aligned}$$

$$= 2pm + f_1 d + 2f_2 a - 2pm - 2p(p^2 + pq\mathscr{F} - q^2 - pq\mathscr{F})a - 2pf_1 d$$

$$= (1-2p)f_1 d + [2f_2 - 2p(p-q)]a$$

$$= -2pq(1-\mathscr{F})(p-q)d + [2p^2 + 2pq\mathscr{F} - 2p^2 + 2pq]a$$

$$= 2pq(1+\mathscr{F})a - 2pq(p-q)(1-\mathscr{F})d$$

$$= 2pq(1+\mathscr{F})\left[a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right] = 2pq(1+\mathscr{F})\alpha_{\mathscr{F}}$$

Thus:

$$\mathrm{cov}(\underline{\gamma}, \underline{\delta}) = -2pq(1+\mathscr{F})\alpha_{\mathscr{F}}{}^2 + 2pq(1+\mathscr{F})\alpha_{\mathscr{F}}{}^2 = 0$$

Now expressions for $\mathrm{var}(\underline{\mathcal{G}})$, $\mathrm{var}(\underline{\gamma})$ and $\mathrm{var}(\underline{\delta})$ as applying to inbred populations will be derived. The expression for $\mathrm{var}(\underline{\delta})$ is obtained by subtracting $\mathrm{var}(\underline{\gamma})$ from $\mathrm{var}(\underline{\mathcal{G}})$.

As $\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\mathcal{G}} - m) = \mathrm{E}(\underline{\mathcal{G}} - m)^2 - [\mathrm{E}(\underline{\mathcal{G}} - m)]^2$, $\mathrm{var}(\underline{\mathcal{G}})$ is derived from the following scheme:

	Genotype		
	bb	Bb	BB
$\mathcal{G} - m$	$-a$	d	a
f	$q^2 + pq\mathscr{F}$	$2pq(1-\mathscr{F})$	$p^2 + pq\mathscr{F}$

Thus:

$$\begin{aligned}
\mathrm{var}(\underline{\mathcal{G}}) &= (q^2 + pq\mathscr{F})a^2 + 2pq(1-\mathscr{F})d^2 + (p^2 + pq\mathscr{F})a^2 \\
&\quad -[(p-q)a + 2pq(1-\mathscr{F})d]^2 \\
&= 2pqa^2 + 2pq\mathscr{F}a^2 + 2pq(1-\mathscr{F})d^2 - 4pq(1-\mathscr{F}) \\
&\quad (p-q)ad - 4p^2q^2(1-\mathscr{F})^2 d^2 \\
&= 2pq[(1+\mathscr{F})a^2 + (1-\mathscr{F})d^2 - 2(1-\mathscr{F})(p-q)ad - 2pq(1-\mathscr{F})^2 d^2] \\
&= 2pq(1+\mathscr{F})\left[a^2 - 2\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)(p-q)ad\right] \\
&\quad +2pq\left[(1-\mathscr{F})d^2 - 2pq(1-\mathscr{F})^2 d^2\right] \\
&= 2pq(1+\mathscr{F})\left[a - \left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)(p-q)d\right]^2 \\
&\quad -2pqd^2\left[(p-q)^2\frac{(1-\mathscr{F})^2}{1+\mathscr{F}} + 2pq(1-\mathscr{F})^2 - (1-\mathscr{F})\right]
\end{aligned}$$

The first term in this expression was shown to be equal to $\text{var}(\underline{\gamma})$. As $\text{var}(\underline{\delta}) = \text{var}(\underline{\mathcal{G}}) - \text{var}(\underline{\gamma})$, it follows that

$$\begin{aligned}
var(\underline{\delta}) &= -2pq\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d^2\left[(1-\mathscr{F})(1-4pq)+2pq(1-\mathscr{F}^2)-(1+\mathscr{F})\right] \\
&= -2pq\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d^2\left(1-4pq-\mathscr{F}+4pq\mathscr{F}+2pq-2pq\mathscr{F}^2-1-\mathscr{F}\right) \\
&= -2pq\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d^2\left(-2pq-2\mathscr{F}+4pq\mathscr{F}-2pq\mathscr{F}^2\right) \\
&= 4pq\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d^2\left[\mathscr{F}+pq(1-2\mathscr{F}+\mathscr{F}^2)\right] \\
&= 4pq\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)\left[\mathscr{F}+pq(1-\mathscr{F})^2\right]d^2
\end{aligned}$$

Example 8.19 shows the partitioning of $\underline{\mathcal{G}}$ in the case of an inbred population.

Example 8.19 Selfing of the population described in Example 8.18 yields the following population:

	Genotype		
	bb	Bb	BB
f	0.48	0.24	0.28
$\mathcal{G}$	11.5	13.5	13.5

Thus $p = 0.4, q = 0.6, \mathscr{F} = 0.5, m = 12.5$ and $a = d = 1$.

$$\mu = m + (p-q)a + 2pq(1-\mathscr{F})d = 12.5 - 0.2 + 0.24 = 12.54$$

The latter is of course equal to $0.48 \times 11.5 + 0.52 \times 13.5$.

$$\text{var}(\underline{\mathcal{G}}) = 0.48 \times 11.5^2 + 0.52 \times 13.5^2 - (12.54)^2 = 0.9984$$

$$\alpha_{\mathscr{F}} = \left[a - (p-q)\left(\frac{1-\mathscr{F}}{1+\mathscr{F}}\right)d\right] = 1 + 0.2 \times \left(\frac{0.5}{1.5}\right) = 1.0667$$

Thus

$$\alpha_b = -p\alpha_{\mathscr{F}} = -0.4 \times 1.0667 = -0.4267$$
$$\alpha_B = q\alpha_{\mathscr{F}} = 0.6 \times 1.0667 = 0.64$$

This yields

	Genotype		
	bb	Bb	BB
f	0.48	0.24	0.28
$\mathcal{G}$	11.5	13.5	13.5
γ	$12.54+$ $2(-0.4267) = 11.6866$	$12.54 - 0.4267$ $+0.64 = 12.7533$	$12.54+$ $2(0.64) = 13.82$
δ	-0.1866	0.7467	-0.32

Where

$$\mathrm{E}\underline{\gamma} = 0.48 \times 11.6866 + 0.24 \times 12.7533 + 0.28 \times 13.82 = 12.54 = \mu$$
$$\mathrm{var}(\underline{\gamma}) = 0.8193$$
$$\mathrm{E}\underline{\delta} = 0$$
$$\mathrm{var}(\underline{\delta}) = 0.1791$$

Thus

$$\mathrm{var}(\underline{\gamma}) + \mathrm{var}(\underline{\delta}) = 0.8193 + 0.1791 = 0.9984 = \mathrm{var}(\underline{\mathcal{G}})$$

Up to now we have considered the components of the genotypic value (and the components of the genotypic variance) for only one segregating locus. The conditions for extending Equations (8.22) and (8.23) to the case of K segregating loci are discussed in Section 10.1. In actual situations the number of relevant loci and the number of alleles at each of these loci are unknown. The present derivations, see also Kempthorne (1957), can thus not directly be applied. However, the partitioning $\underline{\mathcal{G}} = \underline{\gamma} + \underline{\delta}$ is of practical interest because of the relation between the additive genotypic value (Equation (8.6)) and the so-called **breeding value** (Equation (8.12)). This relation is more extensively considered in Section 8.3.4.

8.3.4 Breeding Value: A Concept Dealing with Cross-fertilizing Crops

In the previous section the concept of breeding value was introduced as a rather abstract quantity applying in the case of random mating (see Equation (8.12) for its definition). The practical implications of this quantity for the estimation of the prospects of successful selection are, however, great. For this reason some more aspects of the concept are considered in this section, whereas Section 11.3 gives attention to its application.

Breeders aim to select plants producing superior progenies. This is relatively easy in the case of identical reproduction as the breeder should then simply identify candidates with superior genotypes. The present section gives attention to the much more demanding task of the identification among the candidate of plants producing superior offspring after cross-fertilization, *e.g.* identification of inbred lines producing, after crossing, heterotic hybrids. The best approach is to select among the candidate plants on the basis of the performance of their offspring. This occurs in the case of **progeny testing** (Section 6.3.6). The latter requires maintenance of the parental plants, so that these are still present after the evaluation of their offspring. Such maintenance is possible:

- Vegetatively, either spontaneously for perennial crops or artificially by vegetative reproduction (by means of tissue culture, for instance)
- Sexually, as a (pure) line (this is of relevance when developing a hybrid variety)

The present section is dedicated to the situation where the offspring is obtained by crossing of candidates with a so-called **tester population**. The progenies are HS-families.

Mostly the tester population coincides with the population to which the candidates belong. Then the allele frequencies of the tester population are designated by p and q. Open pollination, as in the case of a **polycross**, is the simplest way of producing the offspring.

The tester population may also be a different population. This is called inter-population testing (see Section 11.3). Then its allele frequencies are designated p' and q'. The aggregate of all test-crosses is then equal to a bulk cross (Section 2.2.1). This situation applies to **top-crossing** as well as to **reciprocal recurrent selection** (Section 11.3). Top-crossing involves pollination of a set of (pure) lines, which have been emasculated, by haplotypically diverse pollen. This pollen may have been produced by a single-cross hybrid (SC-hybrid) or by a genetically heterogeneous population. (In the case of **early testing**, young lines are involved in the top-cross (Section 11.5.2).) Both polycross and top-cross can contribute to the development of a synthetic variety (Section 9.4.3).

Assume that I candidates are crossed with the tester population. The progeny test involves then I HS-families. HS-families performing (far) better than average descend from parents to be selected. Because all candidates have been pollinated by the same tester population the superiority of a HS-family is assumed to be due to its maternal parent. Thus twice the superiority of a HS-family over the mean performance across all HS-families measures the superiority of its maternal parent. Indeed the genetic superiority of a candidate (possibly a single plant) appears from its offspring. The **breeding value** (bv) of some (maternal) parent is therefore defined as:

$$\underline{bv} := 2(\underline{\mathcal{G}}_{\mathrm{HS}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}}) \tag{8.24}$$

In the former section, breeding value was defined as the sum of the main effects of the alleles (Equation (8.12)):

$$\gamma_{ij} - \mu = \alpha_i + \alpha_j = 2[\mathrm{E}(\underline{\mathcal{G}}|B_iB_j) - \mu]$$

The present definition is at the level of expression of quantitative variation in the trait. The quantity $\underline{\mathcal{G}}_{\mathrm{HS}}$ in Equation (8.24), *i.e.* the genotypic value of the HS-family obtained from the parent, is equivalent to the expected genotypic value of the plants representing the HS-family. The quantity $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}}$, *i.e.* the expected genotypic value of the HS-families, is at intrapopulation testing, equivalent to $\mu = \mathrm{E}\underline{\mathcal{G}}$ (see below). The present definition will now be elaborated in terms of quantitative genetic parameters for a single locus, *i.e.* locus $B - b$. Table 8.6 presents for this locus the result of pollination of the plants belonging to some population by the tester population.

The genotypic composition of the aggregate of all HS-families is equal to the result of bulk crossing, *viz.* $(qq', pq' + p'q, pp')$ (Equation (2.1)). Thus

$$\mathrm{E}\underline{\mathcal{G}} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}} = m + (pp' - qq')a + (pq' + p'q)d \tag{8.25}$$

Equation (8.18) provides the breeding values for interpopulation testing. The derivation of the breeding values for interpopulation testing, see Table 8.6, is illustrated for genotype BB. Thus:

$$\begin{gathered} bv_2 = 2[\{m + p'a + q'd\} - \{m - (pp' - qq')a - (pq' + p'q)d\}] \\ = 2[(p' - pp' + qq')a + (q' - pq' - p'q)d] = 2[(p'q + qq')a + (qq' - p'q)d] \\ = 2q[a - (p' - q')d] = (2 - 2p)[a - (p' - q')d] \end{gathered}$$

The part

$$a - (p' - q')d$$

is a function of the allele frequencies in the tester population. In the case of interpopulation progeny testing it will be designated by α' and in the case of intrapopulation progeny testing by α. Thus

$$\alpha' = a - (p' - q')d \tag{8.26a}$$

$$\alpha = a - (p - q)d \tag{8.26b}$$

Table 8.6 The expected genotypic value, *i.e.* $\mathcal{G}_{\mathrm{HS}}$, of the HS-family obtained when pollinating maternal plants by a tester population. The derivation of the breeding values (bv) of the parental plants is explained in the text

	Parental population			Genotypic composition of the HS-families			
gt	f	$\mathcal{G}$	bv	bb	Bb	BB	$\mathcal{G}_{HS}$
bb	f_0	$m - a$	$(0 - 2p)\alpha'$	q'	p'	0	$m - q'a + p'd$
Bb	f_1	$m + d$	$(1 - 2p)\alpha'$	$\frac{1}{2}q'$	$\frac{1}{2}$	$\frac{1}{2}p'$	$m + \frac{1}{2}(p' - q')a + \frac{1}{2}d$
BB	f_2	$m + a$	$(2 - 2p)\alpha'$	0	q'	p'	$m + p'a + q'd$

The latter equation was in Equation (8.17) presented as the average effect of a gene substitution.

The breeding values presented in Table 8.6 for genotypes bb and Bb can be derived in a similar way. General expressions for the breeding value of a candidate with a genotype containing $\underline{j}B$ alleles are thus

$$\underline{bv}_j = (\underline{j} - 2p)\alpha' \tag{8.27a}$$

$$\underline{bv}_j = (\underline{j} - 2p)\alpha \tag{8.27b}$$

Note 8.8 presents a few additional remarks about the topics allele effect and average effect of a gene substitution.

Note 8.8 The breeding value of a genotype for locus $B - b$ depends not only on the allele frequencies p' and q' in the tester population, but also on the allele frequencies p and q in the population of plants to be tested. The allele frequencies p and q change in the case of selection then the breeding values will change as well. Thus, just like the additive genotypic value and the dominance deviation, the breeding value is also a frequency-dependent parameter.

The breeding value of genotype bb is due to 2 b alleles. Thus the so-called **average effect of a single *b* allele**, say ${\alpha_b}'$, is

$${\alpha_b}' = \tfrac{1}{2}bv_0 = -p\alpha'$$

Likewise ${\alpha_B}'$, *i.e.* **the average effect of a single *B* allele**, is

$${\alpha_B}' = \tfrac{1}{2}bv_0 = q\alpha'$$

The difference of the average effects of alleles B and b is

$${\alpha_B}' - {\alpha_b}' = q\alpha' + p\alpha' = \alpha'$$

For this reason α' is sometimes called: **the average effect of a gene substitution**.

The quantities ${\alpha_b}'$ and ${\alpha_B}'$ allow partitioning of the breeding values of the genotypes in terms of the effects of the involved alleles:

	Genotype		
	bb	Bb	BB
bv	$2{\alpha_b}'$	${\alpha_b}' + {\alpha_B}'$	$2{\alpha_B}'$

In Section 8.3.3 the parameters ${\alpha_b}'$ and ${\alpha_B}'$ were called **allele effects**. They are only meaningful in the context of abstract quantitative genetic theory. These effects are frequency-dependent. They change when selection is applied.

As $\mathrm{E}\underline{j} = 2p$ (Note 8.7), it follows from Equation (8.27a) that

$$\mathrm{E}\underline{bv} = \mathrm{E}(\underline{j} - 2p)\alpha' = 0$$

This follows also from the definition of the breeding value (Equation (8.24)):

$$\mathrm{E}\underline{bv} = 2\mathrm{E}(\underline{\mathcal{G}}_{\mathrm{HS}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}}) = 0$$

As

$$\underline{bv} = \underline{\gamma} - \mu$$

(Equation (8.18)), it also follows that

$$\mathrm{var}(\underline{b\nu}) = \mathrm{var}(\underline{\gamma}) = {\alpha_{\mathrm{RM}}}^2\mathrm{var}(\underline{j}) = 2pq{\alpha_{\mathrm{RM}}}^2 = {\sigma_a}^2 \tag{8.28}$$

From Equation (8.24) it is further derived that:

$$\mathrm{var}(\underline{b\nu}) = 4\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) = {\sigma_a}^2 \tag{8.29}$$

Example 8.20 provides an illustration of the calculation of a few of the introduced parameters.

Example 8.20 We consider once more Example 8.12. In the case of intrapopulation testing Equation (8.26b) yields for locus B_3-b_3, with $a = d = 1$ (complete dominance), at $p = 0.4$, $q = 0.6$:

$$\alpha = 1 - (0.4 - 0.6)1 = 1.2$$

The allele effects, see Equations (8.15) and (8.16), amount then to:

$$\alpha_0 = -0.4(1.2) = -0.48,$$

and

$$\alpha_1 = 0.6(1.2) = 0.72;$$

and the breeding value, see Equations (8.6) and (8.27b), to:

$$b\nu_0 = 2(-0.48) = -0.96 = (0 - 0.8)(1.2),$$
$$b\nu_1 = -0.48 + 0.72 = 0.24 = (1 - 0.8)(1.2),$$

and

$$b\nu_2 = 2(0.72) = 1.44 = (2 - 0.8)(1.2).$$

It appears that genotype BB has the highest breeding value.

One may further calculate:

$$\mathrm{E}\underline{b\nu} = 0.36(-0.96) + 0.48(0.24) + 0.16(1.44) = 0.0,$$

and

$$\mathrm{var}(\underline{b\nu}) = \mathrm{E}(\underline{b\nu})^2 = 0.36(-0.96)^2 + 0.48(0.24)^2 + 0.16(1.44)^2 = 0.6912.$$

Chapter 9
Effects of the Mode of Reproduction on the Expected Genotypic Value

In section 8.1 it was emphasized that this book focusses attention on the mean genotypic value as well as on the genetic variance. Breeders manipulate these parameters in such a way that the mean genotypic value is changed in the desired direction. The manipulation may involve the mode of reproduction. For this reason this chapter considers the influence of the coefficient of inbreeding on the mean genotypic value. The important quantitative genetic phenomena heterosis and inbreeding depression indicate that the effect of the mode of reproduction on the mean genotypic value is considerable. The relation between the inbreeding coefficient and the mean genotypic value is therefore considered for both random mating and inbreeding.

9.1 Introduction

In Note 8.6 the following equation was derived for some inbred population with regard to the expected genotypic value of the genotypes for some segregating locus B-b:

$$\mathrm{E}\underline{\mathcal{G}} = m + (p - q)a + 2pq(1 - \mathscr{F})d \tag{9.1}$$

The equation shows that $\mathrm{E}\underline{\mathcal{G}}$ can be changed by

1. changing p and q, *i.e.* by selection and
2. changing the inbreeding coefficient, $\mathscr{F}$.

In this chapter attention is focussed on the effects of $\mathscr{F}$, *i.e.* of the mode of reproduction, on $\mathrm{E}\mathcal{G}$.

In the case of the absence of epistasis the genotypic value of any complex genotype can be written as a sum of contributions due to the single-locus genotypes for the relevant loci (Chapter 1, Section 8.3.2). Consequently, the expected genotypic value with regard to complex genotypes is equal to the sum, across the K relevant loci, of the expected contributions due to the single-locus genotypes

$$\mathrm{E}\underline{\mathcal{G}} = m + \sum_{i=1}^{K}(p_i - q_i)a_i + 2(1 - \mathscr{F})\sum_{i=1}^{K} p_i q_i d_i \tag{9.2}$$

The presence or absence of linkage of the involved loci is irrelevant with regard to this expression.

According to Equation (9.2), the absence of inbreeding depression and/or heterosis indicates absence of directional dominance (Section 9.4.1). In the

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 173–203.

absence of (directional) dominance, Equation (9.2) simplifies. Certain useful applications of the equation can then be justified (Examples 9.1 to 9.3).

Example 9.1 The expected genotypic value of the line obtained by selfing some plant P_i, say $E\underline{\mathcal{G}}_{L(P_i)}$, is derived. Loci for which P_i, is homozygous do not segregate. Only the K relevant loci, heterozygous in P_i, need attention. For each of these loci the line segregates with genotypic composition ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$). The aggregate contributions of these loci to $\mathcal{G}_{P_i}$ and $E\underline{\mathcal{G}}_{L(P_i)}$ are

$$\sum_{i=1}^{K} d_i \text{ and } \frac{1}{2}\sum_{i=1}^{K} d_i,$$

respectively.

In the case of absence of dominance at each of the K loci or absence of directional dominance (both cases imply $d_1 = d_2 = \ldots = d_K = 0$), we get

$$\mathcal{G}_{Pi} = E\underline{\mathcal{G}}_{L(Pi)}$$

In this situation, the mean phenotypic value of the plants representing the line is an unbiased estimate for $\mathcal{G}_{Pi}$.

Example 9.2 The expected genotypic value of the FS-family obtained by crossing plants P_i and P_j, say: $E\underline{\mathcal{G}}_{FSij}$, is considered. This is done for all loci affecting the considered trait.

Loci for which P_i and P_j have the same homozygous genotype do not segregate in the FS-family. Their contribution to $\mathcal{G}_{Pi}, \mathcal{G}_{Pj}$ and $E\underline{\mathcal{G}}_{FSij}$ is represented by the common parameter m.

Now

- let loci B_1-$b_1, \ldots, B_I$-b_I indicate the I loci for which both P_i and P_j are heterozygous,
- let loci B_{I+1}-$b_{I+1}, \ldots, B_{I+J}$-b_{I+J} indicate the J loci for which one parent has the heterozygous genotype and the other parent the homozygous genotype with the lower genotypic value,
- let loci B_{I+J+1}-$b_{I+J+1}, \ldots, B_{I+J+K}$-$b_{I+J+K}$ indicate the K loci for which one parent has the heterozygous genotype and the other parent the homozygous genotype with the higher genotypic value and
- let loci $B_{I+J+K+1}$-$b_{I+J+K+1}, \ldots, B_{I+J+K+L}$-$b_{I+J+K+L}$ indicate the L loci for which the parents have different homozygous genotypes.

The expected genotypic value of the FS-family amounts then to

$$\mathrm{E}\underline{\mathcal{G}}_{FS_{ij}} = m + \tfrac{1}{2}\sum_{i=1}^{I} d_i + \tfrac{1}{2}\sum_{i=I+1}^{I+J}(-a_i + d_i) + \tfrac{1}{2}\sum_{i=I+J+1}^{I+J+K}(a_i + d_i) + \sum_{i=I+J+K+1}^{I+J+K+L} d_i$$

The mean of the genotypic values of the parents, *i.e.* the mid-parent genotypic value, is

$$\tfrac{1}{2}(\mathcal{G}_{P_i} + \mathcal{G}_{P_j}) = \tfrac{1}{2}\left[2m + 2\sum_{i=1}^{I} d_i + \sum_{i=I+1}^{I+J}(-a_i + d_i) + \sum_{i=I+J+1}^{I+J+K}(a_i + d_i)\right]$$

For the case of absence of dominance, *i.e.* for $d_i = 0$ for each segregating locus, it is thus derived that

$$\mathrm{E}\underline{\mathcal{G}}_{FS_{ij}} = \tfrac{1}{2}(\mathcal{G}_{P_i} + \mathcal{G}_{P_j}) = m - \tfrac{1}{2}\sum_{i=I+1}^{I+J} a_i + \tfrac{1}{2}\sum_{i=I+J+1}^{I+J+K} a_i \qquad (9.3)$$

If a set of plants is crossed pairwise, the average phenotypic values of the obtained FS-families can be used to get unbiased estimates of the genotypic values of individual parental plants on the basis of Equation (9.3), provided epistasis and dominance do not occur.

Example 9.3 In the framework of a quantitative genetic analysis of some trait of a self-fertilizing crop, the F_1 is sometimes backcrossed (BC) with both of its parents. These parents may have a different homozygous genotype for K loci. Now

- let loci B_1-$b_1, \ldots, B_I$-b_I indicate the I loci for which P_1 has the homozygous genotype with the higher genotypic value and P_2 the homozygous genotype with the lower genotypic value and
- let loci B_{I+1}-$b_{I+1}, \ldots, B_{I+J}$-b_{I+J} indicate the $J(= K - I)$ remaining loci for which P_1 has the homozygous genotype with the lower genotypic value and P_2 the homozygous genotype with the higher genotypic value.

The expected genotypic value of BC_1, the family resulting from the cross between F_1 and P_1, is

$$\mathrm{E}\underline{\mathcal{G}}_{BC_1} = m + \tfrac{1}{2}\sum_{i=1}^{I}(a_i + d_i) + \tfrac{1}{2}\sum_{i=I+1}^{I+J}(-a_i + d_i)$$

The expected genotypic value of BC_2, the family resulting from the cross between F_1 and P_2, is

$$E\underline{\mathcal{G}}_{BC_2} = m + \tfrac{1}{2}\sum_{i=1}^{I}(-a_i + d_i) + \tfrac{1}{2}\sum_{i=I+1}^{I+J}(a_i + d_i)$$

The average of the expected genotypic values of BC_1 and BC_2 is

$$E\underline{\mathcal{G}}_{BC} = m + \tfrac{1}{2}\sum_{i=1}^{I} d_i + \tfrac{1}{2}\sum_{i=I+1}^{I+J} d_i = m + \tfrac{1}{2}\sum_{i=1}^{K} d_i \tag{9.4}$$

9.2 Random Mating

A single round with panmictic reproduction implies for each locus $\mathscr{F} = 0$. With continued panmixis the genotypic composition with regard to single-locus genotypes will be constant from then on. Equation (9.1) simplifies for continued random mating to:

$$E\underline{\mathcal{G}} = m + (p - q)a + 2pqd \tag{9.5}$$

This equation expresses the contribution of any segregating locus to the expected genotypic value with regard to complex genotypes. In the case of absence of epistasis, that value is equal to the sum, across the K relevant loci, of the contributions due to the single-locus genotypes:

$$E\underline{\mathcal{G}} = m + \sum_{i=1}^{K}(p_i - q_i)a_i + 2\sum_{i=1}^{K} p_i q_i d_i \tag{9.6}$$

Thus, notwithstanding the fact that the genotypic composition with regard to complex genotypes will continue to change from generation to generation, until linkage equilibrium is attained, the expected genotypic value will be constant from G_1, the very first generation obtained by random mating. This is illustrated in Example 9.4. According to this result continued reproduction by means of random mating of plant material descending from a hybrid variety affects the expected genotypic value only when comparing the hybrid, say G_0, and G_1. Only in the presence of selection and/or epistasis will the expected genotypic value continue to change from generation to generation.

The effect of selection on the expected genotypic value appears from the relationship between $E\underline{\mathcal{G}}$ and the allele frequency p of the considered locus. When studying this relationship, or preferably that between

$$E\underline{\mathcal{G}} - m = (p - q)a + 2pqd$$

Example 9.4 Loci B_3-b_3 and B_4-b_4 (see Example 8.12) are considered for allele frequencies $p_3 = 0.4$ and $p_4 = 0.8$. The genotypic values of the complex genotypes and the single-locus genotype frequencies are:

	b_3b_3	B_3b_3	B_3B_3	$f_{B_4\text{-}b_4}$
b_4b_4	11	13	13	0.04
B_4b_4	12	14	14	0.32
B_4B_4	12	14	14	0.64
$f_{B_3\text{-}b_3}$	0.36	0.48	0.16	1.00

Epistasis is absent, whereas $m = 12.5, a_3 = d_3 = 1, a_4 = d_4 = 0.5$.

According to Equation (9.6) the expected genotypic value is

$$\begin{aligned}\mathrm{E}\underline{\mathcal{G}} &= 12.5 + (0.4 - 0.6) \times 1 + (0.8 - 0.2) \times 0.5 + 2 \times 0.4 \times 0.6 \times 1 \\ &\quad + 2 \times 0.8 \times 0.2 \times 0.5 = 13.24.\end{aligned}$$

This result can also be obtained directly from the above scheme, assuming that the population is in linkage equilibrium (which is in fact not known):

$$\mathrm{E}\underline{\mathcal{G}} = 0.36 \times 0.04 \times 11 + \ldots + 0.16 \times 0.64 \times 14 = 13.24$$

and p, one may distinguish

1. Loci with $d < -a$
2. Loci with $-a \leq d < 0$
3. Loci with $d = 0$
4. Loci with $0 < d \leq a$
5. Loci with $d > a$

For any locus with $d = 0, \mathrm{E}\underline{\mathcal{G}} - m$ is a linear function of p:

$$\mathrm{E}\underline{\mathcal{G}} - m = (2p - 1)a = -a + 2ap \tag{9.7}$$

For such loci the expected genotypic value is higher as the allele frequency is higher.

For loci with $d \neq 0$ the quantity $\mathrm{E}\underline{\mathcal{G}} - m$ is a quadratic function of p:

$$\begin{aligned}\mathrm{E}\underline{\mathcal{G}} - m &= (2p - 1)a + 2p(1 - p)d = -a + 2p(a + d) - 2p^2 d \\ &= -a - 2d\left[p^2 - \frac{p(a+d)}{d}\right] = -a - 2d\left[p - \frac{a+d}{2d}\right]^2 + 2d\left[\frac{a+d}{2d}\right]^2 \\ &= -a + \frac{(a+d)^2}{2d} - 2d\left[p - \frac{a+d}{2d}\right]^2 \\ &= -a + \frac{(a+d)^2}{2d} - 2d\left[p - \frac{a+d}{2d}\right]^2 \end{aligned} \tag{9.8}$$

The expected genotypic value has then a minimum or a maximum as a function of p when the first derivative is zero, *i.e.* when

$$-4d\left[p - \frac{a+d}{2d}\right] = 0,$$

thus for

$$p = \frac{a+d}{2d} \tag{9.9}$$

This value of the allele frequency will be indicated by the symbol p_m, the optimum frequency of allele B.

The second derivative, *i.e.* $-4d$, is negative for $d > 0$ (in which case the expected genotypic value has a maximum); it is positive for $d < 0$ (in which case the expected genotypic value has a minimum). Whether or not the maximum or the minimum value can be obtained depends on whether or not p_m is in the range of possible values for p, *i.e.* $0 \leq p \leq 1$. This latter condition requires that

$$0 \leq \frac{a+d}{2d} \leq 1$$

or

1. It requires for $d > 0$ that $d \geq a$, *i.e.* (over)dominance of allele B relative to allele b. With complete dominance ($d = a$) the expected genotypic value attains its maximum at $p_m = 1$, at $d > a$ the maximum is attained at $0 < p_m < 1$.
2. It requires for $d < 0$ that $d \leq -a$, *i.e.* (over)dominance of allele b relative to allele B. With complete dominance ($d = -a$) the expected genotypic value attains its minimum at $p_m = 0$, at $d < -a$ the minimum is attained at $0 < p_m < 1$.

According to Equation (9.8) the maximum or minimum value of $\mathrm{E}\underline{\mathcal{G}} - m$ amounts to

$$-a + \frac{(a+d)^2}{2d} = \frac{a^2+d^2}{2d} \tag{9.10}$$

Example 9.5 illustrates for several loci (all with $a = 2$, but varying with regard to the degree of dominance), the relationship between the allele frequency and the expected genotypic value.

Example 9.5 We consider loci B_1-$b_1, \ldots, B_5$-b_5, with $a_1 = a_2 = \ldots = a_5 = 2$ and $d_1 = -3, d_2 = -1, d_3 = 0, d_4 = 1$ and $d_5 = 3$.

According to Equation (9.9) the value of $\mathrm{E}\underline{\mathcal{G}} - m$ is for locus B_1-b_1 minimal for $p_m = \frac{1}{6} = 0.167$. It amounts then (see Equation (9.10)) to -2.17, see Figure 9.1(i).

Figure 9.1(ii) illustrates the relationship between $\mathrm{E}\underline{\mathcal{G}} - m$ for locus B_2-b_2. For locus $B_3 - b_3$ the relationship is linear. It is given by Equation (9.7) and

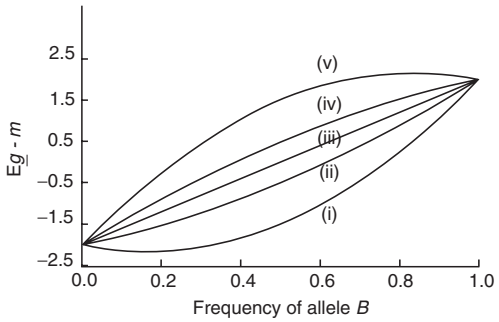


Fig. 9.1 The relation between the frequency of allele B and the expected genotypic value relative to m, *i.e.* $\mathrm{E}\underline{\mathcal{G}} - m$, for loci B_1-$b_1, \ldots, B_5$-b_5, with $a_1 = a_2 = \ldots = a_5 = 2$ and $d_1 = -3, d_2 = -1, d_3 = 0, d_4 = 1$ and $d_5 = 3$

illustrated by Figure 9.1(iii). Locus B_4-b_4 illustrates the situation for a locus with incomplete dominance of allele B: see Figure 9.1(iv). Locus B_5-b_5 is a locus with overdominance of allele B.

For this locus the maximum value of $\mathrm{E}\underline{\mathcal{G}} - m$ amounts to 2.17 (at $p_m = \frac{5}{6} = 0.833$), see Fig. 9.1(v).

9.3 Self-Fertilization

In self-fertilizing crops the frequencies of complex and single-locus genotypes change from generation to generation until complete homozygosity is attained. Consequently the expected genotypic value changes over the generations. This process is considered for the generations obtained by continued selfing of plant material descending from a cross between two pure lines. In the case of absence of selection the allele frequencies stay constant at $p = q = \frac{1}{2}$ for each segregating locus. Equation (9.2) simplifies then into

$$\mathrm{E}\underline{\mathcal{G}} = m + \frac{1}{2}(1 - \mathscr{F}) \sum_{i=1}^{K} d_i \tag{9.11}$$

Table 9.1 presents $\mathrm{E}\underline{\mathcal{G}}$ for a number of interesting generations.

Using the expressions for $\mathrm{E}\underline{\mathcal{G}}$ in Table 9.1, one may predict on the basis of estimates of m and $\sum_{i=1}^{K} d_i$, the expected genotypic value of any generation. This is illustrated in Example 9.6.

Table 9.1 The expected genotypic value ($E\underline{\mathcal{G}}$) of successive generations of a self-fertilizing crop. The inbreeding coefficients ($\mathscr{F}_t$) are derived from Table 3.1b

Generation (t)	*Population*	$\mathscr{F}_t$	$E\underline{\mathcal{G}}$
0	F_1	-1	$m + \sum_{i=1}^{K} d_i$
1	F_2	0	$m + \frac{1}{2}\sum_{i=1}^{K} d_i$
2	F_3	$\frac{1}{2}$	$m + \frac{1}{4}\sum_{i=1}^{K} d_i$
3	F_4	$\frac{3}{4}$	$m + \frac{1}{8}\sum_{i=1}^{K} d_i$
4	F_5	$\frac{7}{8}$	$m + \frac{1}{16}\sum_{i=1}^{K} d_i$
5	F_6	$\frac{15}{16}$	$m + \frac{1}{32}\sum_{i=1}^{K} d_i$
6	F_7	$\frac{31}{32}$	$m + \frac{1}{64}\sum_{i=1}^{K} d_i$
7	F_8	$\frac{63}{64}$	$m + \frac{1}{128}\sum_{i=1}^{K} d_i$
.			
∞	F_∞	1	m

Example 9.6 The famous maize breeder, Jones, collected data for ear length, plant height and grain yield of 2 pure lines, their single cross hybrid and later generations obtained by selfing of random plants (Jones, 1924, 1939). The data for ear length and plant height were obtained in 1923, those for grain yield are means across tests during up to six seasons. Table 9.2 presents summaries of these observations.

Table 9.2 The observed mean phenotypic values and their predictions for ear length (in cm), plant height (in inches) and grain yield (in bu/acre) of a number of generations of maize (source: Jones, 1924, pp. 413–417, 1939)

	Observations			Predictions		
Generation	Ear length	Plant height	Grain yield	Ear length	Plant height	Grain yield
P_1	8.4	67.9	19.5			
P_2	10.7	58.3	19.6			
F_1	16.2	94.6	101.2			
F_2	14.1	82.0	69.1	12.9	78.9	60.4
F_3	14.7	77.6	42.7	11.2	71.0	40.0
F_4	12.1	76.8	44.1	10.4	67.0	29.8
F_5	9.4	67.4	22.5	10.0	65.1	24.7
F_6	9.9	63.1	27.3	9.8	64.1	22.1
F_7	11.0	59.6	24.5	9.6	63.6	20.8
F_8	10.7	58.8	27.2	9.6	63.3	20.2

Assuming absence of epistasis one can estimate m and $\sum_{i=1}^{K} d_i$ in the following way:

- $\hat{m} = \frac{1}{2}(\overline{p}_{\mathrm{P1}} + \overline{p}_{\mathrm{P2}})$, see Section 11.2.3,
- $\sum_{i=1}^{K} \hat{d}_i = \overline{p}_{F1} - \hat{m}$, see Table 9.1.

This yields

	Ear length	Plant height	Grain yield
$\hat{m}$	9.55	63.1	19.55
$\sum_{i=1}^{K} \hat{d}_i$	6.65	31.5	81.65

Using these estimates, derived from $\mathrm{P}_1, \mathrm{P}_2$ and F_1, one may predict for any later generation the expected genotypic value on the basis of expressions for $\mathrm{E}\underline{\mathcal{G}}$ presented in Table 9.1. The predictions are presented in Table 9.2.

Some predictions deviate clearly from their observed value. This may be due to

- Genotype × season interaction, especially when considering ear length or plant height
- Unconscious selection
- Epistasis.

The expected genotypic value of the F_2 appears to be equal to the average of the expected genotypic values of backcross families BC_1 and BC_2, see Equation (9.4). This identity applies only in the absence of epistasis. This condition provides a possibility to test the hypothesis that epistasis does not occur. In the present context this hypothesis states

$$\mathrm{E}\left[\underline{\overline{p}}_{F_2} - \tfrac{1}{2}\left(\underline{\overline{p}}_{BC_1} + \underline{\overline{p}}_{BC_1}\right)\right] = 0$$

The test of this hypothesis and other similar tests are called **scaling tests**. They are applied in quantitative genetic studies and provide a simple way of deciding how reliable predictions may be if they assume a model without interaction.

In Chapter 3 some attention was given to inbreeding procedures yielding complete homozygosity sooner than obtained by continued self-fertilization of plants grown under normal growing conditions, namely the single-seed descent method (SSD; Section 6.1) as well as the production of doubled haploid lines (DH; Section 3.1). In a population genetic sense the SSD-method consists in fact of continued self-fertilization. Table 9.1 presents thus the expected genotypic value of the plant material obtained by the SSD-method.

In the case of unlinked loci the haplotypic frequencies do not change from generation to generation (Section 3.2.3). This means that the haplotypic composition of the gametes produced by some F_1 genotype reflects the genotypic composition of the F_∞ population obtained from it by continued self-fertilization. Doubling of the number of chromosomes of the haploid plants generated from the gametes produced by the F_1 yields thus a population with the genotypic composition of the F_∞ population.

Both the SSD- and the DH-method yield thus a homozygous population of which the expected genotypic value is equal to $E\underline{\mathcal{G}} = m$.

A breeding programme of a self-fertilizing crop may consist of crossing two pure lines followed by selection in the segregating generations. Multiple heterozygous plants may then produce offspring with an attractive recombinant genotype. As the frequency of multiple heterozygous plants decreases very fast in the case of continued selfing, this approach may soon reach a deadlock due to the lack of ample opportunities for recombination.

Errors in the selection are then irreparable. If the breeder crosses genotype $B_iB_ib_jb_j$ with $b_ib_iB_jB_j$ and selects accidentally, possibly due to a low heritability, in F_2 or any later generation, not a single plant with genotype $B_i \cdot B_j$; then (s)he has eliminated the possibility of obtaining genotype $B_iB_iB_jB_j$ in any forthcoming generation.

The breeder of a self-fertilizing crop should, therefore

1. Provide opportunities to allow suitable recombinants to be formed. (Example 9.7 shows that continued crossing and selection increase the probability of generating the best possible genotype.)

Example 9.7 Assume that a breeder has four phenotypically equivalent pure lines at his disposal. The lines differ genotypically. (This may appear from the F_2 s of a diallel cross.) Assume further that the quantitative variation in the considered trait is controlled by 10 loci and that the complex genotypes of the four pure lines are:

Pure line	Genotype									
A	B_1B_1	b_2b_2	b_3b_3	B_4B_4	b_5b_5	B_6B_6	b_7b_7	b_8b_8	b_9b_9	$B_{10}B_{10}$
B	b_1b_1	B_2B_2	b_3b_3	B_4B_4	b_5b_5	b_6b_6	B_7B_7	b_8b_8	b_9b_9	$B_{10}B_{10}$
C	B_1B_1	b_2b_2	B_3B_3	b_4b_4	b_5b_5	B_6B_6	b_7b_7	b_8b_8	b_9b_9	$B_{10}B_{10}$
D	b_1b_1	b_2b_2	B_3B_3	B_4B_4	B_5B_5	b_6b_6	b_7b_7	b_8b_8	b_9b_9	$B_{10}B_{10}$

One may conclude that these four lines represent a restricted source of genetic diversity: as for loci 8, 9 and 10 there is no genetic variation. The best obtainable genotype is $B_1B_1B_2B_2B_3B_3B_4B_4B_5B_5B_6B_6B_7B_7b_8b_8b_9b_9B_{10}B_{10}$. If the breeder only has available lines A, B and C, the best possible genotype is $B_1B_1B_2B_2B_3B_3B_4B_4b_5b_5B_6B_6B_7B_7b_8b_8b_9b_9B_{10}B_{10}$.

Emerson and Smith (1950) aimed to increase the number of grain rows per ear of maize. They started with seven inbred lines of maize, all producing

ears with 12 rows. By continued crossing and selection they developed lines with 22 rows. This result was obtained after establishing that the seven initial inbred lines differed genetically for the studied trait.

2. Maintain desirable combinations intact
3. Select attractive types at an early stage

The opportunities for successful breeding are amplified by starting the selection not in plant material resulting from a single cross, but in plant material resulting from a three-way cross, *i.e.* $F_1 \times \mathbf{P}_3$, or from a multiple cross (Bos, 1987). Lists of varieties show that many varieties of self-fertilizing crops have indeed been developed from complex crosses.

Selfing of plants of cross-fertilizing crops yields mostly poor-performing offspring. This is due to a homozygous genotype, at one or more loci, for undesirable (often recessive) alleles. (Maize breeders may be prepared to observe this phenomenon and, therefore, incorrectly consider vigorous S_1 plants to be the product of contamination.)

Elimination of such undesirable alleles may give rise to much better performing homozygous plant material. Indeed, inbreeding combined with selection may yield attractive homozygous plant material (see Example 9.8).

Example 9.8 Genter (1982) started a selection programme with the single-cross hybrid of the contrasting maize inbred lines Va17 and Va29. F_2 plants were crossed in pairs. The FS-families obtained, constituting population C_0, were tested in replicated trials. Crossing of the best families yielded population C_1. From then on the 'best' plants from one row were crossed with the 'best' plants from the other row. This was continued until C_9. The yield increased from 60% of the original single-cross hybrid up to 104%, *i.e.* 5% per cycle. The general combining ability (see Section 11.5.2) of families belonging to C_4 and C_5 with six testers was better than that of the original hybrid. The same applied to C_8 families. In this generation selfings were made. Some of the lines obtained yielded better than FS-families obtained from the same plants.

The existence of self-fertilizing crops that perform well and which may have evolved from cross-fertilizing predecessors, form a convincing example. Inbred lines that perform well have been developed for more-or-less cross-fertilizing crops, such as cucumber, sunflower (*Helianthus annuus* L.), onion (*Allium cepa* L.) and cotton (*Gossypium hirsutum* L.), or for even obligatory cross-fertilizing crops such as Brussels sprouts (*Brassica oleracea* var. *gemmifera* DC.; Kearsey, 1984). Development of plant material containing *B*-alleles at many loci may be pursued by mild forms of inbreeding, allowing some recombination, combined with selection.

Certain cucurbits are monoecious. This promotes outcrossing. Nevertheless, Genter (1967) reported that selfing hardly ever resulted in **inbreeding depression**, a phenomenon treated in Section 9.4. He supposed that in the

past often just a single plant was harvested to obtain seed for the next generation. Thus continued HS-mating, a mild form of inbreeding, combined with a mild selection, may have given rise to well-performing inbred lines of this group of cross-fertilizing crops. Also Jensen (1970) advocated for self-fertilizing crops the combination of continued selection and repeated crossing. According to him, important shortcomings of conventional cereal breeding procedures are

- the segregating population, obtained by crossing only two homozygous parental lines, affords insufficient genetic variation and
- after the first cross and segregation the probability of further recombination decreases rapidly.

9.4 Inbreeding Depression and Heterosis

9.4.1 Introduction

Inbreeding depression and heterosis are phenomena which may occur at positive and negative values of the inbreeding coefficient ($\mathscr{F}$) of the considered plant material, respectively. These phenomena may occur if $\mathscr{F}$ deviates from 0. Their size appears from the difference between the expected genotypic value ($\mathrm{E}\underline{\mathcal{G}}$) at the value for $\mathscr{F}$ in force and the expected genotypic value of the same plant material at $\mathscr{F} = 0$($\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}$). For self-fertilizing crops the latter is for $p = q = \frac{1}{2}$ equal to $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{F2}}$; for cross-fertilizing crops it is equal to the expected genotypic value of the population with the Hardy–Weinberg genotypic composition corresponding to the actual gene frequencies. The inbreeding depression or heterosis amounts thus to:

$$\mathrm{E}\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}$$

According to Equations (9.2) and (9.6) this yields

$$\left[m + \sum_{i=1}^{K}(p_i - q_i)a_i + 2(1-\mathscr{F})\sum_{i=1}^{K} p_i q_i d_i\right] - \left[m + \sum_{i=1}^{K}(p_i - q_i)a_i + 2\sum_{i=1}^{K} p_i q_i d_i\right]$$
$$= -2\mathscr{F}\sum_{i=1}^{K} p_i q_i d_i \tag{9.12}$$

If $\mathrm{E}\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} = 0$ at $\mathscr{F} \neq 0$ there is a strong indication of absence of dominance at the relevant loci. If $\mathrm{E}\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} \neq 0$ at $\mathscr{F} > 0$, **inbreeding depression** occurs, whereas $\mathrm{E}\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} \neq 0$ at $\mathscr{F} < 0$ implies the presence of **heterosis**.

At $\mathscr{F} \neq 0$ the frequency of heterozygous plants is $2pq(1-\mathscr{F})$, at $\mathscr{F} = 0$ it is $2pq$. The difference is $-2\mathscr{F}pq$, *i.e.* there is a deficit of heterozygous plants at $\mathscr{F} > 0$ and an excess at $\mathscr{F} < 0$. Considered in this way inbreeding depression

and heterosis are due to a deficit or an excess of heterozygous plants, measured in comparison with the Hardy–Weinberg frequency.

It has been observed that continued selfing is very often associated with a decreasing average phenotypic value (Hayes, Immer and Smith, 1955, pp. 76–79; Allard, 1960, pp. 213–219); Falconer, 1989, pp. 248–249). This applies especially to cross-fertilizing crops. Thus there is a general tendency for $\Sigma p_i q_i d_i$ to be positive, implying that $d > 0$ for most loci or for many of the most important loci. This **unidirectional dominance** of the alleles giving, in homozygous genotypes, rise to higher genotypic values has already been mentioned in Section 8.3.1.

There is an obvious reason to measure both inbreeding depression and heterosis in comparison to the performance of the corresponding population with the Hardy–Weinberg genotypic composition. In a cross-fertilizing crop, such as maize, heterosis is relevant if the outbred plant material performs better than conventional open-pollinating varieties. (Likewise, heterosis of self-fertilizing crops is measured by comparing the performance of F_1 hybrids to the performance of conventional pure line varieties.) Measuring heterosis in a cross-fertilizing crop in comparison to the performance of pure lines would not be of practical interest. Superiority of an F_1 hybrid over its homozygous parents is called **hybrid vigour**. In self-fertilizing crops hybrid vigour is less conspicuous than in cross-fertilizing crops and is hardly exploited. The F_2 and later generations may show **transgression**. This means that the segregating population contains plants with a genotypic value outside the range of the genotypic values expressed by the homozygous parents. If transgression does not occur one may conclude that the population did either not comprise enough plants in relation to the number of segregating loci to give rise to such genotypes, or that the involved parents represented already the genotypes with the extreme genotypic values.

Equation (9.12) shows that among the segregating loci only loci with $d_i \neq 0$ contribute to inbreeding depression or heterosis. Thus only such loci get attention in Section 9.4. Furthermore, the equation also shows that these two phenomena are linearly related to $\mathscr{F}$ and that they are affected by

1. The allele frequencies of the relevant loci
2. The number of relevant loci.

The effect of the allele frequencies
For $p = q = \frac{1}{2}$, which applies to plant material derived from an F_1, Equation (9.12) simplifies to

$$E\underline{\mathcal{G}} - E\underline{\mathcal{G}}_{\mathrm{RM}} = -\frac{1}{2}\mathscr{F}\sum_{i=1}^{K} d_i \tag{9.13}$$

For other values for p_i and q_i the product $p_i q_i$ is less than $\frac{1}{4}$, causing the absolute value of $E\underline{\mathcal{G}} - E\underline{\mathcal{G}}_{\mathrm{RM}}$ to be less than $\left|-\frac{1}{2}\mathscr{F}\sum_{i=1}^{K} d_i\right|$. Inbreeding depression and heterosis are consequently most pronounced at $p = q = \frac{1}{2}$.

The effect of the number of loci

For a smaller number of segregating loci, *i.e.* a smaller value for parameter K in Equation (9.12), the inbreeding depression or heterosis will be smaller than for a higher number of segregating loci. It is, indeed, not a good idea to develop a hybrid variety from related pure lines. In self-fertilizing crops fixation of alleles giving rise to homozygous genotypes with high genotypic values is pursued. Thus, for such crops inbreeding depression and heterosis are understandably smaller than for cross-fertilizing crops. This may also explain why the recently started selection from cross-fertilizing crops for inbred lines that perform well has been rather successful. Due to this, seed representing single-cross hybrids of maize can economically be produced.

At $\mathscr{F} = 1$ the inbreeding depression will be at its maximum, *viz.* $-2\sum_{i=1}^{K} p_i q_i d_i$. For $p_i = \frac{1}{2}$ for all relevant loci this amounts to $-\frac{1}{2}\sum_{i=1}^{K} d_i$. At $\mathscr{F} = -1$, implying $p_i = \frac{1}{2}$ for all relevant loci, heterosis will be at its maximum, *viz.* $\frac{1}{2}\sum_{i=1}^{K} d_i$. These extreme values for $\mathscr{F}$ are approached with a rate depending on the mode of reproduction.

With regard to the extreme values for inbreeding depression or heterosis, one should also take into consideration K, the number of relevant loci. Equation (3.23) indicates that the probability that a plant is completely homozygous is $\left(\frac{1+\mathscr{F}_t}{2}\right)^K$. This probability is smaller as K is larger. In the process of inbreeding it will amount to 0.99 or more, sooner when K is small than when K is large. Thus at low values for K the maximum inbreeding depression is reached relatively quickly. According to Allard (1960, Fig. 18.1), Jones established the maximum inbreeding depression for plant height in maize as early as in the S_5 population; for yield, in contrast, it had not yet occurred by S_{20}.

According to Equation (9.12) $\mathrm{E}\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}$ depends linearly on $\mathscr{F}$. Crow and Kimura (1970, p. 79–80) derived that $\mathrm{E}\underline{\mathcal{G}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}$ is a quadratic function of $\mathscr{F}$ in the occurrence of epistasis. A non-linear relation between the observed inbreeding depression and $\mathscr{F}$ may thus be due to epistasis (see Example 9.9).

Example 9.9 Hallauer and Sears (1973) studied the effect of continued selfing, in the absence of selection, on the mean phenotypic value ($\overline{p}$), in the various generations, for 10 different traits of maize. Propagation by single-seed descent was applied at a plant density of 2.9 (plants/m^2) in $S_0, \ldots, S_3$ or 3.87 in $S_4, \ldots, S_7$. The lines were evaluated in 1969 and 1970 at five locations and at a density of 4.14 (plants/m^2).

The linear relation between $\overline{p}$ and $\mathscr{F}$ across the eight generations was significant for each of the ten studied traits; at least 92% of the variation for a trait could be explained by the variation for $\mathscr{F}$. For yield (y, in kg/ha) the relation was $\hat{y} = 6548 - 4494\mathscr{F}$, at a coefficient of correlation estimated to be 0.998.

The quadratic relation between $\overline{p}$ and $\mathscr{F}$ was significant for six traits, but not for yield. It accounted for less than 4% of the variation in $\overline{p}$.

The predominantly linear relation between $\overline{p}$ and $\mathscr{F}$ shows that epistasis was of minor importance.

In Section 3.4 it was shown that selfing in autotetraploid crops leads to a slow decrease in the frequency of heterozygous plants. Yet a single round of reproduction by means of selfing of a natural cross-fertilizing autotetraploid population yields strong inbreeding depression. Allard (1960, p. 217) reported for alfalfa that the S_1 yielded 32% less than the original variety. Busbice and Wilsie (1966) attributed the strong inbreeding depression to the strong reduction of the frequency of plants with a tri- or tetra-allelic heterozygous genotype, *i.e.* $BB\beta b$ or $B\mathbf{B}\beta b$. In artificially made autotetraploid plant material, *e.g.* rye, the inbreeding depression is less than in natural autotetraploid material. The difference is attributed to the lower frequency of plants with a tri- or tetra-allelic heterozygous genotype in artificial autotetraploid populations, but it might equally be due to the expression of deleterious recessive genes.

Both inbreeding depression and heterosis are due to unidirectional dominance of B-alleles, *i.e.* incomplete dominance, complete dominance, or even overdominance. Jinks (1981) concluded that the failure to find examples of 'true' overdominance is general. Thus, if epistatic effects are absent or of minor importance, inbreeding depression and heterosis will mainly occur in the case of dispersion of alleles with (in)complete dominance. This implies that it should be possible to develop pure lines performing as well as F_1 hybrids.

N.B. The phenomenon of pseudo-overdominance may give rise to erroneous conclusions about the genetic control of the considered trait. This is illustrated by Example 9.10.

Example 9.10 Consider loci B_1-b_1 and B_2-b_2, with $m = 2, a_1 = d_1 = a_2 = d_2 = 1$, *i.e.* complete dominance at both loci. The genotypic values of genotypes $b_1b_1b_2b_2, B_1B_1b_2b_2, b_1b_1B_2B_2$ and $B_1B_1B_2B_2$ are 0, 2, 2 and 4, respectively.

Both the cross $B_1B_1b_2b_2 \times b_1b_1B_2B_2$ and the cross $b_1b_1b_2b_2 \times B_1B_1B_2B_2$ yield an F_1 with genotype $B_1b_1B_2b_2$ with $\mathcal{G} = 4$.

If the two loci are strongly linked ($r_c \approx 0$) cross $B_1B_1b_2b_2 \times b_1b_1B_2B_2$ will segregate in the F_2 with a 1:1 segregation ratio with $E\underline{\mathcal{G}} = 3$, which could be explained as due to a single locus with overdominance. Cross $b_1b_1b_2b_2 \times B_1B_1B_2B_2$ will segregate in the F_2 with a 3:1 segregation ratio, which could be explained as due to a single locus with complete dominance.

Heterosis is exploited by developing varieties containing an excess of heterozygous plants in comparison to their frequency at the Hardy–Weinberg equilibrium. Such excess occurs after bulk crossing (Section 2.2.1). The heterosis of the plant material obtained by the bulk cross is:

$$\frac{1}{2}\sum_{i=1}^{K}(p_{1i}-p_{1i})^2 d_i \tag{9.14}$$

where

$$\frac{1}{2}(p_{1i}-p_{1i})^2$$

represents the excess of plants with genotype $B_i b_i$ if the difference in the frequency of allele B_i between the two parental populations amounts to $p_{1i} - p_{2i}$ (see Equation (2.9)).

Equation (9.14) implies that heterosis will be large:

1. If $(p_{1i}-p_{2i})^2$ is large. A bulk cross involving contrasting pure lines, *i.e.* lines with genotypes $b_i b_i$ and $B_i B_i$, yields the maximum value for $(p_{1i} - p_{2i})^2$, *viz.* 1. The resulting plant material is then heterozygous (and genetically uniform).
2. If K is large, *i.e.* if the parental populations, preferably pure lines, have a different homozygous single-locus genotype for a high number of loci.
3. If the parental populations, preferably pure lines with a different homozygous single-locus genotype for many loci, have homozygous genotypes for alleles differing in such a way that d_i is at its maximum. This should be pursued by trial and error.

According to Note 9.1 the above conditions describe, in quantitative genetic terms, the requirements for a high specific combining ability (see Section 11.5.2).

Note 9.1 It is to be expected that a superior hybrid will result from crossing pure lines differing in such a way that both K and d_i are large. It is then roughly correct to say that such lines have a high specific combining ability (Section 11.5.2). In fact, however, the concept of specific combining ability is defined in the framework of a statistical analysis. Its quantitative genetic interpretation is not straightforward.

Heterosis with regard to a complex trait, *i.e.* a trait of which the genetic variation is the result of the variation of a number of component traits, may tentatively be explained on the basis of additive inheritance (absence of dominance) of the components. The explanation is clarified by considering yield (Y) data of some crop, where yield is determined by number of fruits and (average) single fruit weight. When observing each candidate plant with regard to the following traits:

A: number of fruits

B: number of harvested grammes of product, *i.e.* yield (thus: $B = Y$)

One may, in the following way, calculate phenotypic values of the yield components X_1 and X_2:

$$X_1 = A: \text{ number of fruits per plant of the considered candidate}$$

$$X_2 = \frac{B}{A}: \text{ single fruit weight}$$

Thus

$$Y = A \times \frac{B}{A} = B \tag{9.15}$$

A specific case which pointed to the importance of components of complex characters, was the unexpected superiority of hybrids between African and Asian oil-palms. The latter were also of African origin but had undergone several generations of selection under totally different climatic conditions. Under African conditions, the local palms produced a high number of small bunches, whereas the imported Asian palms produced a few very large bunches. The hybrid was intermediate for both number and average weight of the bunches. This resulted in an overall yield far exceeding the mid-parent value.

It has often observed that parents having mutually complementing phenotypic values with regard to yield components, produce a single-cross hybrid with heterosis for yield or other complex characters. Example 9.11 illustrates this phenomenon for a self-fertilizing and a cross-fertilizing crop. It has become known as **recombinative heterosis** (Mac Key, 1976).

Example 9.11 Tables 9.3 and 9.4 illustrate the phenomenon of recombinative heterosis for a self-fertilizing and a cross-fertilizing crop, respectively.

For each of the two yield components the mean phenotypic value of the offspring lies within the range of the parental phenotypic values. Table 9.3 shows for both yield components incomplete dominance of the lower level of expression. In Table 9.4 this applies to one of the components. Yet in both tables the yield of the offspring exceeds those of the parents.

Table 9.3 The plant yield of single tomato plants, as the product of the number of fruits per plant and the mean single fruit weight of two pure lines and their single-cross hybrid (source: Powers, 1944)

Material	Number of fruits	Fruit weight (g)	Plant yield (g)
P_2	4.4	138	607
F_1	44.5	55	2,428
P_1	109.1	17	1,868

Table 9.4 The yearly bunch yield of single oil-palm trees as the product of the yearly number of bunches per palm and the mean single bunch weight of 2 *tenera* palms and their offspring (source: Van der Vossen, 1974, Table 12)

Material	Number of bunches	Bunch weight (kg)	Bunch yield
1.2229T	5.8	7.1	41.2
32.2612T × 1.2229T	8.5	6.3	53.6
32.2612T	16.3	2.8	45.6

One may speculate with regard to this phenomenon as follows. The yield of a plant may be assumed to be at its maximum if all organs and functions are mutually tuned. This may occur if the plant has an intermediate phenotypic value for each of a number of yield components, *e.g.* number of stems, number of flowers per stem, number of seeds per flower and seed size. If the intermediate phenotypic values for the components are due to heterozygous single-locus genotypes, it is understandable that plants with a heterozygous complex genotype have a superior value for the complex character.

The idea that a complex trait, *e.g.* grain yield, should be indirectly improved via improvement of its components may lead to an interest in the physiological processes underlying the complex trait. Thus, in addition to plant architectural features, *e.g.* ear size, crop physiological parameters may be used to describe the features of the ideal genotype, the so-called **ideotype**. The ideotype for rice is, for instance, characterized by erect leaves, compact and large panicles on a short and firm culm, a vigorous root system and absence of unproductive tillers.

An ideotype may be designed on the basis of estimates of the crop physiological parameters that are relevant to the crop growth model used. These estimates are usually obtained from evaluation of a limited set of genotypes. After having designed an ideotype, crop physiologists simply advise breeders to create it. In practice there are, however, complications: the majority of the traits that are to be assessed with this approach are hard to measure with the required accuracy. The assessment, for example, of the rate of reallocation of dry matter from stems and leaves to seeds is not feasible in a segregating population with many genotypes, each of which is represented by a single plant or by, at most, a small number of plants. Selection for such traits is thus mostly beyond the breeder's capability (Stam, 1998).

Furthermore it is assumed when designing an ideotype that parameter values can be combined at will in a single genotype. The possible existence of constraints, *e.g.* lack of genetic variation, and correlations among the parameters, especially correlations due to pleiotropic loci, is ignored.

Sparnaaij and Bos (1993) and Bos and Sparnaaij (1993) considered the analysis of complex characters as well as the phenomenon of recombinative heterosis and its prediction.

Equation (9.12) shows that inbreeding depression is due to a deficit of heterozygous plants in comparison with their Hardy-Weinberg frequency. Random variation of allele frequencies also leads to a decrease in the frequency of heterozygous plants. If $P_{nf,0}$ designates the probability that fixation with regard to locus B_i-b_i has not yet occurred in the initial population, $P_{nf,t}$ is expected to be $\psi P_{nf,0}$, where ψ represents the remaining part of $P_{nf,0}$ (Section 7.1).

The initial contribution of locus B_i-b_i to $\mathrm{E}\underline{\mathcal{G}}$ is $(p_i - q_i)a_i + 2p_iq_id_i$. At fixation of genotype B_iB_i, which occurs with probability p_i, the contribution is a_i; at fixation of genotype b_ib_i, which occurs with probability q_i, it is $-a_i$. Thus, at fixation, the expected contribution of this locus is $(p_i - q_i)a_i$. Consequently, at fixation due to random variation of allele frequencies its expected contribution to 'inbreeding' depression amounts to $-2p_iq_id_i$. The expected depression, due to fixation, is thus equal to the depression occurring in the case of continued inbreeding.

9.4.2 Hybrid Varieties

Comparison of a number of the annual Dutch lists of varieties shows both an increase in the total number of varieties for grain and silage maize, and a gradual shift in the most frequently included type of variety. The increase in the total number of varieties reflects the increase in acreage since 1970. Apparently breeders responded by offering more and more varieties. The main type of variety offered changed simultaneously: from open-pollinating varieties via double-cross hybrids (DC-hybrid) and threeway-cross hybrids (TC-hybrids) to single-cross hybrids (SC-hybrid) (Table 9.5).

Table 9.5 The number of varieties of grain and silage maize included in Dutch lists of recommended varieties and their distribution across open-pollinating varieties (OP), double-cross (DC), threeway-cross (TC) and single-cross (SC) hybrid varieties

	Type of variety				
Year	OP	DC	TC	SC	Total
1967	4	4	0	0	8
1977	0	3	6	0	9
1980	0	2	8	0	10
1984	0	1	12	0	13
1988	0	2	14	0	16
1990	0	2	19	0	21
1992	0	2	19	3	24
1994	0	1	26	16	43
1996	0	0	19	17	36
1998	0	0	19	19	38

The table shows that, in the past, DC-hybrids were more popular than SC-hybrids. Because DC-hybrid seed is produced by a vigorous SC-hybrid, it was much cheaper than SC-hybrid seed. (The latter is produced by an inbred line suffering from inbreeding depression). At present, however, relatively high yielding pure lines are available as maternal parent of a SC-hybrid. Already in 1980 about 80% of the acreage of maize grown in the Corn Belt of the USA consisted of SC-hybrids.

Two reasons for the present popularity of SC-hybrids are

1. Farmers prefer their greater uniformity
2. Breeders prefer to evaluate the lower number of all conceivable SC-hybrids instead of all conceivable TC- or DC-hybrids (see below)

Numbers of conceivable SC-, TC- and DC-hybrids
When having available N promising inbred lines, one might produce and test

- $\binom{N}{2}$ SC-hybrids
- $\binom{N}{2}(N-2)$ TC-hybrids
  As each of the $\binom{N}{2}$ SC-hybrids may be crossed with any of the $(N-2)$ remaining inbred lines, the number of TC-hybrids is $(N-2)$ times the number of SC-hybrids.
- $3\binom{N}{4}$ DC-hybrid
  This number is derived as follows. Each of the $\binom{N}{2}$ SC-hybrids may be crossed with any of the $\binom{N-2}{2}$ SC-hybrids among the $(N-2)$ remaining inbred lines. When reciprocal crosses are not distinguished, this yields $\frac{1}{2}\binom{N}{2}\binom{N-2}{2} = 3\binom{N}{4}$ DC-hybrids, *i.e.* $\frac{1}{4}(N-2)(N-3)$ times the number of SC-hybrids.

Example 9.12 shows that it is demanding or even impossible to produce and to test all conceivable TC- and DC-hybrids when N becomes larger than 15.

Example 9.12 The number of SC-hybrids, TC-hybrids and DC-hybrids that may be produced on the basis of N inbred lines amounts for $N = 5$, 15 and 50 to

N	Number of SC-hybrids	Number of TC-hybrids	Number of DC-hybrids
5	10	30	15
15	105	1365	4095
50	1225	58800	690900

Thus the five inbred lines V, W, X, Y and Z may give rise to 10 different SC-hybrids, *viz.* VW, VX, VY, VZ, WX, WY, WZ, XY, XZ and YZ. When making TC-hybrids each of these may be crossed with any of the three inbred lines not already used as its parent, *e.g.* VW may be crossed with X,

Y or Z. Alternatively, when making DC-hybrids one may cross each of the 10 SC-hybrids with any of the $\binom{3}{2} = 3$ SC-hybrids among the three remaining inbred lines. Pooling of reciprocal crosses yields $3\binom{5}{4} = 15$ DC-hybrids.

The costs of producing 1 tonne of SC-hybrid maize seeds are not necessarily higher than those required to produce 1 tonne of TC- or DC-hybrid seed, the reasons being:

1. Because of mutual isolation of maize fields, grown for maintenance of inbreds or their crossing, the production of TC- or DC-hybrid seed is more demanding than the production of SC-hybrid seed: to produce DC-hybrid seed at least seven isolated fields are required, instead of three when producing SC-hybrid seed (check this for yourself).
2. For a given successful SC-hybrid the alleles may be reshuffled to produce a new maternal and a new paternal inbred line, such that the new maternal line has a higher seed yield (Koutsika-Sotiriou, Bos and Fasoulas, 1990).

Of course, growers will be interested in the performance of G_1, *i.e.* the plant material obtained by open pollination in the hybrid variety. If the performance of G_1 would be satisfactory, they might decide to grow G_1-, G_2-, *etc.* material.

In the case of the absence of epistasis a single round of panmictic reproduction will yield plant material (G_1) with an expected genotypic value equal to that of any later generation obtained by panmixis, *i.e.* equal to $E\mathcal{G}_{RM}$ (Section 9.2). Then the reduction in the performance, occurring when growing G_1, G_2, *etc.* instead of the hybrid, is $E\underline{\mathcal{G}}_{hybrid} - E\underline{\mathcal{G}}_{RM}$, which is equal to the heterosis as defined by Equation (9.12). Example 9.13 illustrates the reduction occurring when growing plant material obtained by panmictic reproduction of a hybrid. In addition to the reduction in performance, the plant material will show a reduced uniformity.

Example 9.13 The four homozygous genotypes $b_3b_3b_4b_4, b_3b_3B_4B_4$, $B_3B_3b_4b_4$ and $B_3B_3B_4B_4$ of Example 8.12 may be coded W, X, Y and Z.

TC-hybrid YZ · W is produced by crossing SC-hybrid YZ, which has genotype $B_3B_3B_4b_4$, with inbred line W. The genotypic composition of hybrid YZ · W is described by

	Genotype	
	$B_3b_3B_4b_4$	$B_3b_3b_4b_4$
f	$\frac{1}{2}$	$\frac{1}{2}$
$\mathcal{G}$	14	13

Thus the expected genotypic value of the TC-hybrid is

$$E\underline{\mathcal{G}}_{YZ\cdot W} = \tfrac{1}{2}(14 + 13) = 13.5$$

Its allele frequencies are $p_3 = \frac{1}{2}$ and $p_4 = \frac{1}{4}$. As $m = 12.5, a_3 = d_3 = 1$ and $a_4 = d_4 = \frac{1}{2}$ (Example 8.12), Equation (9.6) yields

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} = 12.5 + (\tfrac{1}{2} - \tfrac{1}{2})1 + (\tfrac{1}{4} - \tfrac{3}{4})\tfrac{1}{2} + 2[\tfrac{1}{2} \cdot \tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot \tfrac{3}{4} \cdot \tfrac{1}{2}] = 12.94$$

Thus the heterosis amounts to $13.5 - 12.94 = 0.56$. This is the reduction of the performance when growing $\mathrm{G}_1, \mathrm{G}_2$, *etc.* obtained by continued panmictic reproduction starting with TC-hybrid $\mathrm{YZ} \cdot \mathrm{W}$.

If the number of SC-hybrid plants is insufficient to produce the desired amount of DC-hybrid seed, one may apply open pollination within both of the SC-hybrids underlying the DC-hybrid. Next the two G_1s are crossed. This procedure yields plant material with (approximately) the same genotypic composition as expected when crossing the two SC-hybrids. The explanation for this is as follows. The population resulting from open pollination of a SC-hybrid is identical to the population resulting from self-fertilization of the SC-hybrid. When applying selfing, the haplotype frequencies with regard to unlinked loci do not change. (In the case of linkage the change is insignificant, see Section 3.2.2). Thus a single round of panmictic reproduction of each of the two SC-hybrids hardly affects the genotypic composition of the DC-hybrid to be produced.

Prediction of the performances of TC-hybrids and DC-hybrids

Example 9.12 illustrated that it is, even for a rather low number of inbred lines (N), impossible to produce and to test all $\binom{N}{2}(N-2)$ TC- or all $3\binom{N}{4}$ DC-hybrids. The remainder of this section is dedicated to a way out: it has become a routine to predict, on the basis of data about the performances of the SC-hybrids, the performance of any conceivable TC- or DC-hybrid. This prediction can indeed be made for each TC- and DC-hybrid if data about all SC-hybrids are available. The TC- or DC-hybrids with the most favourable predicted performances are subsequently actually produced and tested.

The predictions are based on the following equations:

- For TC-hybrid $\mathrm{XY} \cdot \mathrm{Z}$:

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{XY} \cdot \mathrm{Z}} = \tfrac{1}{2}(\mathcal{G}_{\mathrm{XZ}} + \mathcal{G}_{\mathrm{YZ}}) \tag{9.16}$$

- For DC-hybrid $\mathrm{WX} \cdot \mathrm{YZ}$:

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{WX} \cdot \mathrm{YZ}} = \tfrac{1}{4}(\mathcal{G}_{\mathrm{WY}} + \mathcal{G}_{\mathrm{WZ}} + \mathcal{G}_{\mathrm{XY}} + \mathcal{G}_{\mathrm{XZ}}) \tag{9.17}$$

The performance of TC-hybrid $\mathrm{XY} \cdot \mathrm{Z}$, *i.e.* $\mathcal{G}_{\mathrm{XY} \cdot \mathrm{Z}}$, is therefore predicted as

$$\tfrac{1}{2}(\hat{\mathcal{G}}_{XZ} + \hat{\mathcal{G}}_{YZ}) \tag{9.18}$$

and the performance of DC-hybrid $\mathrm{WX} \cdot \mathrm{YZ}$, *i.e.* $\mathcal{G}_{\mathrm{WX} \cdot \mathrm{YZ}}$, as

$$\tfrac{1}{4}(\hat{\mathcal{G}}_{\mathrm{WY}} + \hat{\mathcal{G}}_{\mathrm{WZ}} + \hat{\mathcal{G}}_{\mathrm{XY}} + \hat{\mathcal{G}}_{\mathrm{YZ}}) \tag{9.19}$$

The performances predicted according to Equations (9.18) and (9.19) will be best if the performances of the SC-hybrids occurring in the equations are the best. The SC-hybrids to be used to produce the best possible TC- or DC-hybrid should thus not have the best possible performances.

The reliability of Equations (9.16) and (9.17) will now be illustrated for the case of absence of epistasis, implying that presence or absence of linkage is irrelevant. The illustration is only elaborated for loci B_1-b_1 and B_2-b_2.

The genotypes assumed for pure lines W, X, Y and Z are

Line code	Genotype	Genotypic value ($\mathcal{G}$)
W	$B_1B_1B_2B_2$	$m + a_1 + a_2$
X	$B_1B_1b_2b_2$	$m + a_1 - a_2$
Y	$b_1b_1B_2B_2$	$m - a_1 + a_2$
Z	$b_1b_1b_2b_2$	$m - a_1 - a_2$

This yields the following SC-hybrids:

Hybrid code	Genotype	Genotypic value ($\mathcal{G}$)
WX	$B_1B_1B_2b_2$	$m + a_1 + d_2$
WY	$B_1b_1B_2B_2$	$m + d_1 + a_2$
WZ	$B_1b_1B_2b_2$	$m + d_1 + d_2$
XY	$B_1b_1B_2b_2$	$m + d_1 + d_2$
XZ	$B_1b_1b_2b_2$	$m + d_1 - a_2$
YZ	$b_1b_1B_2b_2$	$m - a_1 + d_2$

TC-hybrid XY · Z is then described by

	Genotype			
	$b_1b_1b_2b_2$	$B_1b_1b_2b_2$	$b_1b_1B_2b_2$	$B_1b_1B_2b_2$
f	$\frac{1}{2}r_c$	$\frac{1}{2}(1 - r_c)$	$\frac{1}{2}(1 - r_c)$	$\frac{1}{2}r_c$
$\mathcal{G}$	$m - a_1 - a_2$	$m + d_1 - a_2$	$m - a_1 + d_2$	$m + d_1 + d_2$

Its expected genotypic value is

$$\begin{aligned} \mathrm{E}\underline{\mathcal{G}}_{\mathrm{XY\cdot Z}} &= m + a_1(-\tfrac{1}{2}r_c - \tfrac{1}{2} + \tfrac{1}{2}r_c) + d_1(\tfrac{1}{2} - \tfrac{1}{2}r_c + \tfrac{1}{2}r_c) \\ &\quad + a_2(-\tfrac{1}{2}r_c - \tfrac{1}{2} + \tfrac{1}{2}r_c) + d_2(\tfrac{1}{2} - \tfrac{1}{2}r_c + \tfrac{1}{2}r_c) \\ &= m - \tfrac{1}{2}a_1 + \tfrac{1}{2}d_1 - \tfrac{1}{2}a_2 + \tfrac{1}{2}d_2 \end{aligned}$$

It is easily verified that this is equal to

$$\tfrac{1}{2}(\mathcal{G}_{\mathrm{XZ}} + \mathcal{G}_{\mathrm{YZ}}) = \tfrac{1}{2}[(m + d_1 - a_2) + (m - a_1 + d_2)]$$

Similarly DC-hybrid WX · YZ is described by

	Genotype		
	$B_1b_1b_2b_2$	$B_1b_1B_2b_2$	$B_1b_1B_2B_2$
f	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$\mathcal{G}$	$m + d_1 - a_2$	$m + d_1 + d_2$	$m + d_1 + a_2$

Its expected genotypic value is

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{WX\cdot YZ}} = m + d_1 + \tfrac{1}{2}d_2$$

This is equal to

$$\begin{aligned}\tfrac{1}{4}(\mathcal{G}_{\mathrm{WY}} + \mathcal{G}_{\mathrm{WZ}} + \mathcal{G}_{\mathrm{XY}} + \mathcal{G}_{\mathrm{XZ}}) = \tfrac{1}{4}[&(m + d_1 + a_2) + (m + d_1 + d_2) \\ &+ (m + d_1 + d_2) + (m + d_1 - a_2)] \\ = m + d_1 + \tfrac{1}{2}d_2\end{aligned}$$

In this way it is illustrated that, for the case of absence of epistasis, the prediction is unbiased.

The expressions to predict TC- or DC-hybrid performances are due to Jenkins (1934). Applications were elaborated by Allard (1960, pp. 271–274) and Hallauer and Miranda (1981, pp. 352–357).

The predictions are based on estimates of the genotypic values of SC-hybrids. Inaccuracy of these estimates may lead to incorrect predictions. Other causes for differences between predicted and actual performances may be

- Genotype × environment interaction: the prediction may be based on observations made in 2007 whereas the verification occurred in 2008, possibly at a different location
- Maternal effects
- Presence of epistasis

Unexpected behaviour of plant material may determine the failure or the success of a breeder. Thus the predictions should be used as rough indications. Ample actual evaluation of promising hybrids, during several years and at several locations, is always required.

Example 9.14 shows (for $N = 4$) the prediction, on the basis of data about the performances of each of the six SC-hybrids, of the performances of all 12 conceivable TC-hybrids and all three conceivable DC-hybrids.

Example 9.14 The genotypic values of the $\binom{N}{2} = 6$ SC-hybrids conceivable for $N = 4$ inbred lines W, X, Y and Z were estimated to amount to

$$\begin{aligned}\mathcal{G}_{\mathrm{WX}} &= 14 \\ \mathcal{G}_{\mathrm{WY}} &= 13 \\ \mathcal{G}_{\mathrm{WZ}} &= 14 \\ \mathcal{G}_{\mathrm{XY}} &= 14 \\ \mathcal{G}_{\mathrm{XZ}} &= 7 \\ \mathcal{G}_{\mathrm{YZ}} &= 10\end{aligned}$$

According to Equation (9.18) the predictions of the expected genotypic values of the $\binom{N}{2}(N-2) = 12$ TC-hybrids amount to

$$\begin{aligned}
\hat{\mathcal{G}}_{\mathrm{WX\cdot Y}} &= \tfrac{1}{2}(13+14) = 13.5\\
\hat{\mathcal{G}}_{\mathrm{WX\cdot Z}} &= \tfrac{1}{2}(14+7) = 10.5\\
\hat{\mathcal{G}}_{\mathrm{WY\cdot X}} &= \tfrac{1}{2}(14+14) = 14\\
\hat{\mathcal{G}}_{\mathrm{WY\cdot Z}} &= \tfrac{1}{2}(13+10) = 11.5\\
\hat{\mathcal{G}}_{\mathrm{WZ\cdot X}} &= \tfrac{1}{2}(14+7) = 10.5\\
\hat{\mathcal{G}}_{\mathrm{WZ\cdot Y}} &= \tfrac{1}{2}(13+10) = 11.5\\
\hat{\mathcal{G}}_{\mathrm{XY\cdot W}} &= \tfrac{1}{2}(14+13) = 13.5\\
\hat{\mathcal{G}}_{\mathrm{XY\cdot Z}} &= \tfrac{1}{2}(7+10) = 8.5\\
\hat{\mathcal{G}}_{\mathrm{XZ\cdot W}} &= \tfrac{1}{2}(14+14) = 14\\
\hat{\mathcal{G}}_{\mathrm{XZ\cdot Y}} &= \tfrac{1}{2}(14+10) = 12\\
\hat{\mathcal{G}}_{\mathrm{YZ\cdot W}} &= \tfrac{1}{2}(13+14) = 13.5\\
\hat{\mathcal{G}}_{\mathrm{YZ\cdot X}} &= \tfrac{1}{2}(14+7) = 10.5
\end{aligned}$$

According to Equation (9.19) the predictions of the expected genotypic values of the $3\binom{N}{4} = 3$ DC-hybrids are

$$\begin{aligned}
\hat{\mathcal{G}}_{\mathrm{WX\cdot YZ}} &= \tfrac{1}{4}(13+14+14+7) = 12\\
\hat{\mathcal{G}}_{\mathrm{WY\cdot XZ}} &= \tfrac{1}{4}(14+14+14+10) = 13\\
\hat{\mathcal{G}}_{\mathrm{WZ\cdot XY}} &= \tfrac{1}{4}(14+13+7+10) = 11
\end{aligned}$$

Thus the most promising TC-hybrids are $\mathrm{WY}\cdot\mathrm{X}$ and $\mathrm{XZ}\cdot\mathrm{W}$. These are as good as the best three SC-hybrids WX, WZ and XY. The most promising DC-hybrid is $\mathrm{WY}\cdot\mathrm{XZ}$. This hybrid has a lower performance than the best SC- or TC-hybrid).

The inferior SC-hybrid XZ is identified as a parent of promising TC- or DC-hybrids.

Its parental pure lines X and Z give mostly rise to good-performing SC-hybrids, *e.g.* WX, WZ and XY, when crossed with pure lines W or Y.

9.4.3 Synthetic Varieties

Hermaphroditic cross-fertilizing crops exist in which neither a reliable system of cytoplasmic male sterility occurs, nor incompatibility, *e.g.* some herbage crops. The breeding and maintenance of hybrid varieties is then greatly hampered. In other crops hybrid varieties may be developed but are not actually produced because the additional costs for the grower, due to the more expensive hybrid seed, are not repaid by the additional yield or by the advantage of greater uniformity.

In these situations the breeding of a **synthetic variety** may be considered. Characteristic features of synthetic varieties are

1. Syn_1, *i.e.* generation 1 of the synthetic variety, is obtained by open pollination as occurring in a polycross.
2. The components are maintained by identical reproduction.
3. Syn_1 and later generations, *i.e.* Syn_2, Syn_3, etc., produce offspring by open pollination.

Production of Syn_1 by a polycross
The n parental components with a good combining ability may be identified on the basis of a **polycross** (see Section 6.3.6). Generally a good general combining ability requires unrelatedness. However, to develop a rather uniform synthetic variety the components should be phenotypically similar and, consequently, may have a similar genotype. This requirement may hamper the composition of a set of good combining components. For date of flowering the components should, by definition, be similar in any case.

Maintenance of the components by identical reproduction
The maintenance of the components by identical reproduction (see Section 8.1) may be done by vegetative reproduction (in grasses) or by continued sib mating (*e.g.* in rye). This implies that the components are mostly clones or inbred populations.

Production of Syn_2, Syn_3, etc. by open pollination
A synthetic variety is required to have a fairly constant performance when comparing successive generations. In the absence of epistasis a reduction of the expected genotypic value will only occur from Syn_1 to Syn_2 (see Example 9.15). Further reductions in later generations should be attributed to epistasis and/or (natural) selection.

Example 9.15 Inoue and Kaneko (1976, Table 27) observed the grain yield (in qu/ha) of successive generations of a synthetic variety of maize:

$$\begin{aligned} \overline{p}_{\text{Syn1}} &= 60.5 \\ \overline{p}_{\text{Syn2}} &= 50.2 \\ \overline{p}_{\text{Syn3}} &= 49.7 \\ \overline{p}_{\text{Syn4}} &= 50.4 \end{aligned}$$

Geiger, Diener and Singh (1981) present data concerning the performance of successive generations of synthetic varieties of rye.

When having N potential components available, the total number of conceivable synthetic varieties based on n components, where $n = 2$, or 3, or ..., N, amounts to:

$$\sum_{n=2}^{N} \binom{N}{n} = \sum_{n=0}^{N} \binom{N}{n} - N - 1 = 2^N - N - 1$$

This implies that already for $N = 15$, the development of as many as 32,752 different synthetic varieties may be considered. Prediction of the performances of synthetic varieties is thus very desirable. Such prediction is possible on the basis of the observed performances of material resulting from pairwise crosses between the components involved in the conceived synthetic variety. This is shown in Note 9.2.

Note 9.2 Assume panmictic reproduction of the set of n components. The expected genotypic value of the obtained plant material will then be

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{G}_{F_{ij}}}{n^2} = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} \mathcal{G}_{F_{ij}} + \sum_{i=1}^{n} \mathcal{G}_{F_{ii}}}{n^2}$$

where

- $\mathcal{G}_{\mathrm{F}ij}$ designates the genotypic value of F_{ij}, the plant material obtained from crossing maternal component i with paternal component j, and
- $\mathcal{G}_{\mathrm{F}ii}$ the genotypic value of F_{ii}, the plant material obtained from selfing component i.

In the case of inbred (thus homozygous) parents

$$\frac{\sum_{i=1}^{n} G_{F_{ii}}}{n}$$

is equal to the mean genotypic value of the parents, say $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}}$. The mean genotypic value of the plant material obtained from the crosses (these are hybrids in the case of homozygous parents) is equal to

$$\frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} G_{F_{ij}}}{n(n-1)}$$

say $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{F1}}$. It is, in fact the mean genotypic value of the synthetic variety obtained in the case of outbreeding. Thus $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{F1}} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn1}}$.

Altogether it is derived that

$$\begin{aligned}\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} &= \left(\frac{n-1}{n}\right)\left(\frac{1}{n(n-1)}\right)\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathcal{G}_{F_{ij}} + \frac{1}{n}\cdot\frac{\sum_{i=1}^{n}\mathcal{G}_{F_{ii}}}{n} \\ &= \left(\frac{n-1}{n}\right)\mathrm{E}\underline{\mathcal{G}}_{F_1} + \frac{1}{n}\mathrm{E}\underline{\mathcal{G}}_{P} = \mathrm{E}\underline{\mathcal{G}}_{F_1} - \frac{\mathrm{E}\underline{\mathcal{G}}_{F_1} - \mathrm{E}\underline{\mathcal{G}}_{P}}{n}\end{aligned}$$

Plant material obtained by panmixis has the Hardy–Weinberg genotypic composition. Thus the former expression presents $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn2}}$ and may be read as

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn_2}} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn_1}} - \frac{\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn_1}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}}}{n} \tag{9.20}$$

implying

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn_1}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} = \frac{\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn_1}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}}}{n} \tag{9.21}$$

The latter equation is illustrated in Example 9.16.

Example 9.16 Example 2.8, dealing with a polycross involving $n = 5$ components, is once more considered with regard to the complex genotypes with regard to the two loci B_1-b_1 and B_2-b_2. The genotypic values of the complex genotypes are

	b_2b_2	B_2b_2	B_2B_2
b_1b_1	5.5	13.5	13.5
B_1b_1	7.5	15.5	15.5
B_1B_1	9.5	17.5	17.5

The values of the components of the genotypic values are: $a_1 = 2$, $d_1 = 0$, $a_2 = d_2 = 4$, as in Example 8.10. From Table 2.3 the following derivations can be made: $p_1 = 0.8, q_1 = 0.2, p_2 = 0.4$ and $q_2 = 0.6$. Equation (9.6) yields then:

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}(=\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}) = 11.5 + (0.8-0.2)2 + (0.4-0.6)4 + 2\times 0.4\times 0.6\times 4 = 13.82$$

From Table 2.3 we may calculate

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}} = 0.2\times 5.5 + 0.4\times 9.5 + 0.4\times 17.5 = 11.9, \text{and}$$
$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn1}} = 0.2\times 7.5 + 0.2\times 15.5 + 0.1\times 9.5 + 0.4\times 17.5 + 0.1\times 17.5 = 14.3.$$

This implies that $\frac{E\underline{\mathcal{G}}_{Syn_1} - E\underline{\mathcal{G}}_{P}}{n}$ is equal to $\frac{14.3-11.9}{5} = 0.48$, which, according to Equation (9.21), indeed is equal to $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{Syn1}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} = 14.3 - 13.82$.

The n parental components need to be maintained in mutual isolation. Syn_1 is produced by mixed growing of the components followed by harvest, in bulk, of the seed produced after open pollination. The grower may purchase Syn_1 material, but will mostly buy Syn_2 and grow then several generations. If growers buy exclusively Syn_2 the reduction in performance from Syn_1 to Syn_2 is only the breeder's concern. Despite this reduction, Syn_2 should still perform attractively.

Syn_2 is obtained by random mating, implying $\text{E}\underline{\mathcal{G}}_{\text{Syn2}} = \text{E}\underline{\mathcal{G}}_{\text{RM}}$. The reduction in the performance occurring from Syn_1 to Syn_2 is thus equal to the heterosis of Syn_1 in comparison to Syn_2. Wright (1922) derived Equation (9.20), describing the heterosis of a synthetic variety developed from n parental components, with expected genotypic value $\text{E}\underline{\mathcal{G}}_{\text{P}}$. The equation implies that one may predict $\text{E}\underline{\mathcal{G}}_{\text{Syn2}}$ by

$$\overline{p}_{\text{Syn1}} - \left(\frac{\overline{p}_{\text{Syn1}} - \overline{p}_{\text{P}}}{n} \right) \tag{9.22}$$

and the heterosis of Syn_1 by

$$\frac{\overline{p}_{\text{Syn1}} - \overline{p}_{\text{P}}}{n} \tag{9.23}$$

The five assumptions underlying the derivation of Equation (9.20) (Note 9.2) are

1. Syn_1 originates from outbreeding, *i.e.* intercomponent crossing of the n parental components, in the absence of intracomponent crossing.

   This assumption can be justified if the components are self-incompatible, *e.g.* clones of grasses. The outbreeding causes an excess of heterozygous plants in Syn_1 compared to their Hardy–Weinberg equilibrium frequency occurring in Syn_2 or later generations. This excess gives rise to heterosis.
2. A diploid behaviour of the chromosomes.

   For many polyploid herbage crops, such as grasses or alfalfa, synthetic varieties have been developed. Thus this assumption cannot be justified for all crops for which synthetic varieties are developed.
3. The components are homozygous, at least for the loci controlling the traits considered by the breeder (the latter may be accomplished by assortative mating).

   In practice the components are often only partly inbred (possibly because of presence of self-incompatibility).
4. Absence of epistasis.
5. Syn_2 originates from panmixis.

   This assumption may even be justified in the presence of self-incompatibility. The gametophytic incompatibility occurring in grasses is due to two multiple allelic loci: the S- and the Z-locus. Syn_1 is expected

to produce, at gametogenesis, so many different haplotypes – each consisting of a unique combination of an S- and a Z-allele – that the frequency of incompatible pollinations can be neglected.

Predictions of the performance of Syn_2 or predictions of the heterosis of Syn_1, on the basis of Equations (9.22) and (9.23), respectively, may be inaccurate or biased. Reasons for this are

- Genotype × environment interaction, as mentioned in Example 9.6
- Inappropriateness of one or more of the assumptions used in the derivation of Equation (9.21).

Prediction on the basis of Equation (9.22) or (9.23) is indeed inappropriate in certain situations. Alternative expressions applying to specific situations have therefore been developed. Gallais (1967), for instance, developed an expression for self-compatible components, which are consequently partially inbred. His expression contains the inbreeding coefficient, making allowance for the appropriate degree of inbreeding. Gallais (1967, 2003) also developed expressions for autotetraploid crops. These take into consideration

- preferential fertilization, which has been shown to occur in alfalfa;
- epistasis and
- linkage.

Busbice (1969, 1970) proposed a general expression which can be applied at

- Several levels of ploidy
- Several degrees of relatedness of the parental components
- Several degrees of self-incompatibility

Example 9.16 derived the heterosis to be expected for a Syn_1 variety at specific allele frequencies and specific genotypic values. An expression for the heterosis of Syn_1 for the general case, but taking five assumptions into account, was shown to yield the same result. Indeed, Example 9.16 does not prove the usefulness for breeding practice of Equation (9.21). Such usefulness, however, appears from Example 9.17.

The components involved in a synthetic variety should preferentially be chosen on the basis of a test of the progenies resulting from pairwise crosses. A drawback of selecting among parental components on the basis of a polycross is elaborated in Section 11.3.

Example 9.17 Table 9.6 presents results of a study by Neal (1935) concerning grain yield data of maize lines and hybrids. The data allow calculation of the heterosis by comparing the grain yield of the hybrids with the grain yield of G_1 *i.e.* the material obtained from open pollination in the hybrid. For SC-hybrids the actual heterosis amounted to $62.8 - 44.2 = 18.6$ bu/acre.

Table 9.6 The grain yield of maize material: pure lines used to produce hybrids, the hybrids themselves and the offspring obtained by open pollination in the hybrids, say G_1 (source: Neal, 1935)

	Grain yield			
Type of	parental		G_1	
hybrid	lines	hybrids	observed	predicted*)
SC	23.7	62.8	44.2	43.2
TC	23.8	64.2	49.3	50.7
DC	25.0	64.1	54.0	54.3

*) predicted by using Equation (9.22)

The heterosis predicted on the basis of Equation (9.23) amounted for SC-hybrids: $(62.8 - 23.7)/2 = 19.6$. Then the predicted grain yield of the G_1 material is $62.8 - 19.6 = 43.2$ bu/acre.

Kiesselbach (1960) observed no further reduction in the case of continued reproduction by means of open pollination. This suggests absence of epistasis.

Mostly a synthetic variety is based on 6, 7 or 8 components. As n is smaller, $E\underline{\mathcal{G}}_{Syn1}$ could be higher, but this may be offset by an increase of $(E\underline{\mathcal{G}}_{Syn1} - E\underline{\mathcal{G}}_{P})/n$. There is, apparently, an optimum value for n. Becker (1982, 1988) reviewed the topic of synthetic varieties, including published optimal and actual values for n.

Chapter 10
Effects of the Mode of Reproduction on the Genetic Variance

This book focusses on the mean genotypic value as well as on the genetic variance. Breeders seek desired changes of the mean genotypic value. Presence of genetic variance is a prerequisite for success if the change is pursued by selection. The magnitude of the genetic variance, a measure for the diversity of the genotypic values of the candidates, depends on the genotypic composition of the population subjected to selection. At given allele frequencies, the coefficient of inbreeding is decisive for the genotypic composition. The effect of the mode of reproduction, the major factor determining the coefficient of inbreeding, on the genetic variance is therefore considered for both random mating and inbreeding.

10.1 Introduction

In the absence of epistasis the genotypic value of a complex genotype with regard to loci $B_1 - b_1, \ldots, B_K - b_K$ can be written as the sum of contributions due to the relevant single-locus genotypes (Section 8.3.2):

$$\mathcal{G}_{B_1-b_1,\ldots,B_K-b_K} = m + \sum_{i=1}^{K} \mathcal{G}'_{B_i-b_i}$$

or

$$\mathcal{G} = m + \sum_{i=1}^{K} \mathcal{G}'_i$$

Then

$$var(\underline{\mathcal{G}}) = var\left(\sum_{i=1}^{K} \underline{\mathcal{G}}'_i\right)$$

If $\text{cov}(\underline{\mathcal{G}}'_i, \underline{\mathcal{G}}'_j) = 0$ for all $i \neq j = 1, \ldots, K$ this simplifies to

$$var(\underline{\mathcal{G}}) = \sum_{i=1}^{K} var\left(\underline{\mathcal{G}}'_i\right) \tag{10.1}$$

implying that the variance of the genotypic values for a polygenically determined trait can be written as the sum of the contributions due to relevant single-locus genotypes.

The condition $\text{cov}\left(\underline{\mathcal{G}}'_i, \underline{\mathcal{G}}'_j\right) = 0$ applies if $\underline{\mathcal{G}}'_i$ and $\underline{\mathcal{G}}'_j$ are independent random variables, *i.e.* if the probability of a certain genotype for locus $B_i - b_i$

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 205–223.

does not depend on the genotype for locus $B_j - b_j$. Such independency is present:

- in cross-fertilizing crops if the considered population is in linkage equilibrium;
- in self-fertilizing crops in the populations designated as F_2, F_3, *etc.* in the case of unlinked loci (see, for example, Table 3.3).

In these situations the effect of the mode of reproduction on var($\underline{\mathcal{G}}$) depends exclusively on its effect on the contribution of separate loci to var($\underline{\mathcal{G}}$). Thus implications of random mating and (continued) self-fertilization for Equations (8.22) and (8.23) are considered in Sections 10.2 and 10.3, respectively.

10.2 Random Mating

We consider the genetic variance for a quantitatively varying trait, which is controlled by non-epistatic loci. For a population with the linkage equilibrium genotypic composition, var($\underline{\mathcal{G}}$) is easily obtained by summation across all relevant single loci (Equation (10.1)). Because $\mathscr{F} = 0$ we consider

	Genotype		
	bb	Bb	BB
f	q^2	$2pq$	p^2
$\mathcal{G}$	$m - a$	$m + d$	$m + a$

Substitution of $\mathscr{F} = 0$ in Equations (8.22) and (8.23) gives

$$\text{var}(\underline{\mathcal{G}}) = \text{var}(\underline{\gamma}) + \text{var}(\underline{\delta}) = 2pq[a - (p - q)d]^2 + 4p^2q^2d^2 \tag{10.2}$$

Extension to the case of K loci for a population in linkage equilibrium yields:

$$var(\underline{\mathcal{G}}) = 2\sum_{i=1}^{K} p_i q_i \left[a_i - (p_i - q_i)d_i\right]^2 + 4\sum_{i=1}^{K} p_i^2 q_i^2 d_i^2 \tag{10.3}$$

The part

$$2\sum_{i} p_i q_i \left[a_i - (p_i - q_i)d_i\right]^2 \tag{10.4}$$

is the additive genetic variance at $\mathscr{F} = 0$. It will be indicated by ${\sigma_a}^2$ (Section 8.3.3). The part

$$4\sum_{i} p_i^2 q_i^2 d_i^2 \tag{10.5}$$

is the dominance variance at $\mathscr{F} = 0$, which will be indicated by ${\sigma_d}^2$ (Section 8.3.3). Thus

$${\sigma_g}^2 := {\sigma_a}^2 + {\sigma_d}^2$$

In the absence of selection p and q are constant, implying constancy of var($\underline{\mathcal{G}}$). Note 10.1 presents an interesting application of Equation 10.3.
Example 10.1 illustrates the calculation of the genotypic variance and its components.

Note 10.1 For unlinked loci the plant material obtained by open pollination within a single cross hybrid variety is in linkage equilibrium for $p_i = \frac{1}{2}; i = 1, \ldots, K$. Substitution of these allele frequencies into Equation (10.3) yields

$$var(\underline{\mathcal{G}}) = \frac{1}{2}\sum_{i=1}^{K} {a_i}^2 + \frac{1}{4}\sum_{i=1}^{K} {d_i}^2 \tag{10.6}$$

The genotypic composition of the obtained population is identical to the genotypic composition of an F_2 population of a self-fertilizing crop. Table 10.3 presents, indeed, the above equation for var($\underline{\mathcal{G}}$) for an F_2 population.

Example 10.1 The genotypic variance is calculated for Example 9.4 by application of the definition for variance. Thus

$$\text{var}(\underline{\mathcal{G}}) = \text{E}\underline{\mathcal{G}}^2 - (\text{E}\underline{\mathcal{G}})^2$$

where:

$$\begin{aligned}\text{E}\underline{\mathcal{G}}^2 &= 0.36 \times 0.04 \times 11^2 + 0.48 \times 0.04 \times 13^2 + \cdots \\ &\quad + 0.16 \times 0.64 \times 14^2 = 176.2576 \\ (\text{E}\underline{\mathcal{G}})^2 &= (13.24)^2 = 175.2976\end{aligned}$$

This yields

$$\text{var}(\underline{\mathcal{G}}) = 0.96$$

Application of Equations (10.4) and (10.5) yields:

- for locus $B_3 - b_3$ with $p_3 = 0.4, q_3 = 0.6, a_3 = d_3 = 1$:

  $$2 \times 0.4 \times 0.6[1 - (0.4 - 0.6)]^2 + 4 \times 0.4^2 \times 0.6^2 = 0.6912 + 0.2304 = 0.9216$$

  and
- for locus $B_4 - b_4$ with $p_4 = 0.8, q_4 = 0.2, a_4 = d_4 = \frac{1}{2}$:

  $$\begin{aligned}&2 \times 0.8 \times 0.2[\tfrac{1}{2} - (0.8 - 0.2) \times \tfrac{1}{2}]^2 + 4 \times 0.8^2 \times 0.2^2 \times (\tfrac{1}{2})^2 \\ &\quad = 0.0128 + 0.0256 = 0.0384\end{aligned}$$

Altogether this yields

$$\sigma_a{}^2 = 0.6912 + 0.0128 = 0.704$$
$$\sigma_d{}^2 = 0.2304 + 0.0256 = 0.256$$
$$\sigma_g{}^2 = 0.704 + 0.256 = 0.960$$

N.B. At the end of Section 8.3.4 it was shown that, in the case of intrapopulation progeny testing, $\sigma_a{}^2$ is equal to the variance of the breeding values.

It is very desirable to know $\sigma_a{}^2$ because it is the numerator in the ratio $\frac{\sigma_a{}^2}{\sigma_p{}^2}$, which is called **heritability in the narrow sense**, designated by $h_n{}^2$. This ratio is a scale-independent quantity, which plays an important role in the theory of selection methods: it is possible to predict the response to selection when $h_n{}^2$ is known (Section 11.1).

Example 10.1 shows that even in the case of complete dominance σ_a^2 may be (considerably) larger than σ_d^2. For $d = a$ it can be shown that this applies if the frequency of allele B is less than $\frac{2}{3}$. Figure 10.1 illustrates $\sigma_g{}^2$, $\sigma_a{}^2$ and $\sigma_d{}^2$ for incomplete dominance, *i.e.* for $a = 2$ and $d = 1$, which corresponds to Fig. 9.1, graph (iv), and also for complete dominance, *viz.* for $a = d = 2$.

Figure 10.1 shows that in the case of incomplete dominance $\sigma_a{}^2$ is by far the larger component of $\sigma_g{}^2$.

The additive genetic variance is 0:

- if $p = 0$,
- if $p = 1$
- if $a - (p - q)d = a - (2p - 1)d = 0$,

i.e. if $\boldsymbol{p} = \frac{a+d}{2d} = \boldsymbol{p}_m$, the frequency of allele B for loci where $d > a$, such that the expected genotypic value attains its maximum if $d > 0$ or its minimum if $d < 0$ (see Section 9.2). One should realize that the above conditions for $\sigma_a{}^2 = 0$ imply absence of opportunities for further improvement of $\mathrm{E}\underline{\mathcal{G}}$ by selection.

By pollinating (and harvesting) the plants of some generation in a proper way, one can partition the genotypic variance (see Equation (10.2)) such that $\sigma_a{}^2$ (Equation (10.4)), the component deserving special interest, can be estimated. Two estimation procedures that require only a small effort are elaborated. They apply to the two modes of reproduction of cross-fertilizing crops most frequently employed:

1. Open pollination followed by separate harvesting of random plants, which yields HS-families (see Section 10.2.1).
2. Pairwise crossing of random plants followed by separate harvesting of the pairs of plants involved in a certain cross. This yields FS-families (see Section 10.2.2).

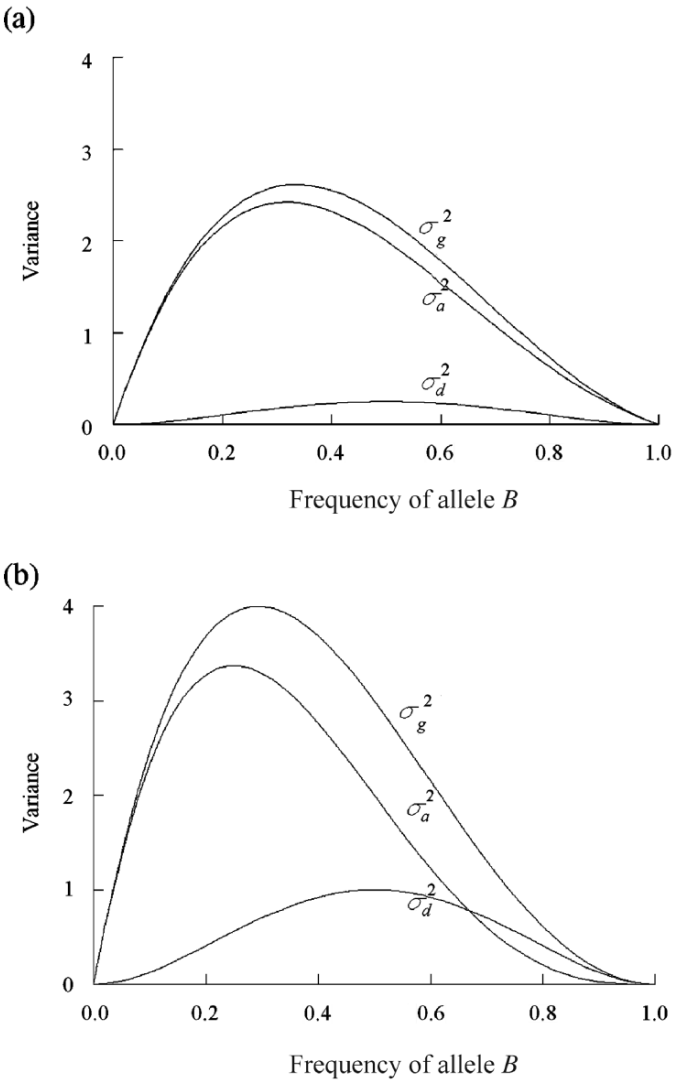


Fig. 10.1 The relation between the frequency of allele B and ${\sigma_g}^2, {\sigma_a}^2$ and ${\sigma_d}^2$ for (a) $a = 2$ and $d = 1$ (incomplete dominance) and (b) $a = d = 2$ (complete dominance)

The present chapter considers for both situations the partitioning of ${\sigma_g}^2$ into genetic variance between families and genetic variance within families. The partitioning is done in such a way that these components are written in terms of ${\sigma_a}^2$ and ${\sigma_d}^2$. Separate evaluation of either the HS- or the FS-families enables the estimation of ${\sigma_a}^2$. Actual experiments, required to estimate ${\sigma_a}^2$ are dealt with in Section 11.2.2

10.2.1 Partitioning of ${\sigma_g}^2$ in the case of open pollination

In the case of open pollination one may partition $\mathrm{var}(\underline{\mathcal{G}})$ as

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) + \mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{HS})}) \tag{10.7}$$

where

- $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}})$ designates the genetic variance between HS-families, *i.e.* the variance of the genotypic values of the HS-families, where $\underline{\mathcal{G}}_{\mathrm{HS}}$ is defined to be equal to the expected genotypic value of the plants representing some HS-family. Thus one may write

$$\underline{\mathcal{G}}_{\mathrm{HS}} = \mathrm{E}(\underline{\mathcal{G}}|\mathrm{HS})$$

- $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{HS})})$ designates the expected genetic variance within HS-families.

N.B. In the above the formulation 'expected genetic variance within HS-families' is incidentally used. Indeed the genetic variance within a HS-family depends on the genotype of its maternal parent.

In Section 8.3.4, Equation (8.29), it was derived that

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) = \tfrac{1}{4}{\sigma_a}^2 \tag{10.8}$$

This implies that

$$\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{HS})}) = \tfrac{3}{4}{\sigma_a}^2 + {\sigma_d}^2 \tag{10.9}$$

In addition to Equation (10.8), it is also possible to estimate ${\sigma_a}^2$ on the basis of the relationship between parents and offspring. Thus we consider the phenotypic value of random maternal plants, say $\underline{p}_{\mathrm{M}}$, as well as the phenotypic values of the HS-families they produce after open pollination, say $\underline{p}_{\mathrm{HS}}$, where $\underline{p}_{\mathrm{HS}}$ is the expected phenotypic value calculated across the plants constituting the considered HS-family. The relation between $\underline{p}_{\mathrm{M}}$ and $\underline{p}_{\mathrm{HS}}$ is of course of interest. In Note 10.2 it is shown that

$$\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) = \tfrac{1}{2}{\sigma_a}^2 \tag{10.10}$$

Thus, when evaluating HS-families derived from random plants, estimates for ${\sigma_a}^2$ are

$$4\hat{\mathrm{var}}(\underline{\mathcal{G}}_{\mathrm{HS}}) \tag{10.11}$$

and

$$2\hat{\mathrm{cov}}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) \tag{10.12}$$

Equations (10.8) and (10.10) imply a quantitative genetical interpretation of the statistical parameters $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}})$ and $\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}})$ in terms of ${\sigma_a}^2$. The conditions required to justify such an interpretation will now be considered. It will, all things being considered, be concluded that a possible bias in Equation (10.10) tends to be smaller than a possible bias in Equation (10.8). Then estimation of ${\sigma_a}^2$ according to Equation (10.12) is to be preferred over estimation according to Equation (10.11).

Note 10.2 When assigning individual plants at random to positions in the field, the covariance of a plant's genotypic value and the environmental deviation of the HS-family, obtained by open pollination of the plant, is zero: $\mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{M}}, \underline{e}_{\mathrm{HS}}) = 0$. Also the covariance of the plant's environmental deviation and the genotypic value of the HS-family, obtained by open pollination of the plant, is zero: $\mathrm{cov}(\underline{e}_{\mathrm{M}}, \underline{\mathcal{G}}_{\mathrm{HS}}) = 0$. Likewise $\mathrm{cov}(\underline{e}_{\mathrm{M}}, \underline{e}_{\mathrm{HS}}) = 0$. All this implies

$$\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) = \mathrm{cov}[(\underline{\mathcal{G}} + \underline{e})_{\mathrm{M}}, (\underline{\mathcal{G}} + \underline{e})_{\mathrm{HS}}] = \mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{M}}, \underline{\mathcal{G}}_{\mathrm{HS}})$$

Of course

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}} = \mathrm{E}[\mathrm{E}(\underline{\mathcal{G}}|\mathrm{HS})] = \mathrm{E}\underline{\mathcal{G}}$$

When considering some locus $B - b$, Equation (9.5) implies

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{M}} = \mathrm{E}\underline{\mathcal{G}} = m + (p - q)a + 2pqd$$

The parameter $\mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{M}}, \underline{\mathcal{G}}_{\mathrm{HS}}) = \mathrm{E}(\underline{\mathcal{G}}_{\mathrm{M}} \cdot \underline{\mathcal{G}}_{\mathrm{HS}}) - (\mathrm{E}\underline{\mathcal{G}}_{\mathrm{M}}) \cdot (\mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}})$ is derived from Table 10.1.

Table 10.1 The relationship between the genotypic value of a maternal plant ($\mathcal{G}_{\mathrm{M}}$) and the genotypic value of the corresponding HS-family ($\mathcal{G}_{\mathrm{HS}}$), *i.e.* the expected genotypic value of the plants constituting the considered HS-family

			HS-family			
Maternal plant			Genotypic composition			
genotype	f	$\mathcal{G}_{\mathrm{M}}$	bb	Bb	BB	$\mathcal{G}_{\mathrm{HS}}$
bb	q^2	$m - a$	q	p	0	$m - qa + pd$
Bb	$2pq$	$m + d$	$\frac{1}{2}q$	$\frac{1}{2}$	$\frac{1}{2}p$	$m + \frac{1}{2}(p - q)a + \frac{1}{2}d$
BB	p^2	$m + a$	0	q	p	$m + pa + qd$

As the constant m may be neglected, this yields

$$q^2(-a)(-qa + pd) + pq(d)[(p - q)a + d] + p^2(a)(pa + qd) - [(p - q)a + 2pqd]^2$$
$$= [q^3 + p^3 - (p - q)^2]a^2 - pq[q - (p - q) - p + 4(p - q)]ad + (pq - 4p^2q^2)d^2$$

When applying Equation (2.8) this is simplified into:

$$pqa^2 - 2pq(p-q)ad + pq(1-4pq)d^2 = pq[a-(p-q)d]^2$$

Thus

$$\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) = \tfrac{1}{2}\sigma_a^2$$

The interpretation of the statistical parameters in the left hand side of Equations (10.8) and (10.10) in terms of the quantitative genetic parameter σ_a^2 in the right-hand side can only be justified if the following conditions apply:

1. Absence of epistasis
2. The genotypic composition of the parental population is in linkage equilibrium
3. The parents produce offspring by means of panmixis
4. Absence of extra-chromosomal genetic variation affecting the genotypic values
5. Absence of genotype × environment interaction
6. Absence of covariance of genotypic value and environmental deviation

In the following, consequences of violations of these conditions are considered in detail. This results in the conclusion that Equation (10.12) gives rise to a smaller bias when estimating σ_a^2 than Equation (10.11).

Presence of epistasis

In the presence of epistasis Equations (10.8) and (10.10) are incorrect. This is illustrated by the effect of interaction of single-locus genotypes when considering only two loci. Falconer (1989, p. 157) presents for this case the following equations:

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) = \tfrac{1}{4}\sigma_a^2 + \tfrac{1}{16}\mathfrak{o}_{aa}^2$$

and

$$\mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{M}}, \underline{\mathcal{G}}_{\mathrm{HS}}) = \tfrac{1}{2}\sigma_a^2 + \tfrac{1}{4}\sigma_{aa}^2$$

where σ_{aa}^2 represents the genetic variance due to interaction between homozygous single-locus genotypes (see parameter *aa* in Table 8.5). When using Equation (10.11) to estimate σ_a^2, the bias amounts to $\frac{1}{4}\sigma_{aa}^2$; when using Equation (10.12) it amounts to $\frac{1}{2}\sigma_{aa}^2$, *i.e.* twice as high. Presence of epistasis implies overestimation of σ_a^2, especially when using Equation (10.12).

Parental population not in linkage equilibrium

Linkage equilibrium is required to justify the summation of single-locus genetic variances applied when determining the genetic variance for complex genotypes (Section 10.1). If the parental population is not in linkage equilibrium,

Equations (10.8) and (10.10) are incorrect. The bias occurring when estimating σ_a^2 by using Equation (10.11) or (10.12), will be relatively large in recently composed populations and in the case of selection.

Offspring not produced by panmixis

Panmixis implies, among other things, absence of selection. This means that the parental plants represent some specific population and that all parental genotypes produce the same number of offspring. In reality genotypes differ in fitness.

To be able to grow a progeny, the maternal plants should produce a certain minimum number of seeds. Plants not producing that minimum number are passed over. This may imply selection. What is the effect of this with regard to estimating σ_a^2? Falconer (1989, p. 183) said: 'The selection causes the variance between the parents to be reduced and consequently the covariance of sibs to be reduced'. In other words: the variance among the HS-families is reduced. Then the actual value of σ_a^2 will be underestimated, especially when estimating σ_a^2 on the basis of Equation (10.11). According to Kempthorne (1957, p. 329) the opinion that selection does not result in a biased estimate of σ_a^2 'will be true only if the regression of y on x is linear throughout the range of x'. In connection with this the statement that 'for non-normal frequency distributions, the regression generally deviates from linearity' (Spitters, 1979; p. 217), deserves attention.

The presence of so-called outcrossing devices may also disturb panmixis. Thus incompatibility, as in grass species, *Brassica oleracea* L. and rye, yields – compared to the Hardy – Weinberg genotypic composition – an excess of heterozygous plants. On the other hand, an excessive amount of selfing, implying a deficit of heterozygous plants, will occur in monoecious crops, such as maize, particularly if there is calm weather during the period of pollen release.

In summary, it is concluded that the bias due to (artificial) selection leads to an underestimation of σ_a^2 when using Equation (10.11).

Presence of extra-chromosomal genetic variation

The notion that extra-chromosomal factors affect plant development has evolved only slowly. Such factors may imply that the genotypic value of a plant is not only due to nuclear genes but to **plasmagenes** as well. One can make allowance for this by partitioning the genotypic value in the following way:

$$\underline{\mathcal{G}} = \underline{\mathcal{G}}_{\mathrm{n}} + \underline{\mathcal{G}}_{\mathrm{p}}$$

Then, in the case of absence of covariance of the contributions due to nuclear alleles and plasmagenes, one may derive

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) = \mathrm{var}[(\underline{\mathcal{G}}_{\mathrm{n}} + \underline{\mathcal{G}}_{\mathrm{p}})_{\mathrm{HS}}] = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{nHS}}) + \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{pHS}}) = \tfrac{1}{4}\sigma_a^2 + \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{p}})$$

and

$$\begin{aligned}\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) &= \mathrm{cov}[(\underline{\mathcal{G}}_{\mathrm{n}} + \underline{\mathcal{G}}_{\mathrm{p}})_{\mathrm{M}}, (\underline{\mathcal{G}}_{\mathrm{n}} + \underline{\mathcal{G}}_{\mathrm{p}})_{\mathrm{HS}}] \\ &= \mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{nM}}, \underline{\mathcal{G}}_{\mathrm{nHS}}) + \mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{pM}}, \underline{\mathcal{G}}_{\mathrm{pHS}}) = \tfrac{1}{2}\sigma_a^2 + \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{p}})\end{aligned}$$

Equations (10.11) and (10.12) will, consequently, yield a biased estimate of σ_a^2 if condition 4 does not apply. Because of the coefficients 4 and 2 in Equations (10.11) and (10.12), respectively, the bias due to using Equation (10.11) is larger than the bias due to using Equation (10.12).

Of course, $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}})$ may be estimated correctly if plasmagenes play a role, and successful selection may be partly due to selection for effects of plasmagenes, but interpretation of $\hat{\mathrm{cov}}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}})$ or $\hat{\mathrm{var}}(\underline{\mathcal{G}}_{\mathrm{HS}})$ in terms of σ_a^2 is then incorrect.

Variation among families may partly be due to variation in the physiological conditions of the maternal plants at harvest time (*e.g.* the degree of seed maturity). Effects of common environments are then to be expected. These include not only maternal effects, but also developmental time trends, as different families experience different environmental conditions at the same stage of development.

Presence of genotype × environment interaction

Interaction of genotype and macro-environmental conditions affects $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}})$. In Chapter 13 it is shown that effects of such interactions are included in the genotypic values of the HS-families when evaluating these only in a single growing season. Such interaction biases the estimate of σ_a^2 when based on Equation (10.11). However, it does not bias the estimate based on Equation (10.12) because $\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}})$ is not affected by genotype × growing season interaction if the maternal plants and the corresponding HS-families are evaluated in different growing seasons. Equation (10.11) tends thus to yield estimates of σ_a^2 more biased by $g \times e$ interaction than Equation (10.12). Estimates of σ_a^2 due to Equation (10.11) tend, consequently, to be larger than estimates due to Equation (10.12). This is supported by data presented in Example 11.11. Casler (1982) stressed that overestimation of the heritability in the narrow sense (h_n^2) is to be expected, when estimating h_n^2 on the basis of regression of offspring on parent where offspring and parents are grown in the same season. (The latter is possible in the case of vegetative maintenance.)

Presence of covariance of genotypic value and environmental deviation

Presence of covariance of genotypic value and environmental deviation implies presence across the families of a negative or a positive correlation of genotypic value and the quality of growing conditions. Proper randomization, ensuring that the entries to be evaluated are assigned positions in the field in a random

way, warrants absence of such a correlation and contributes to avoidance of a biassed estimate of σ_a^2.

10.2.2 Partitioning of $\sigma_g{}^2$ in the case of pairwise crossing

Pairwise crossing yield FS-families. When evaluating these families $\mathrm{var}(\underline{\mathcal{G}})$ is partitioned as

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}}) + \mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{FS})}) \tag{10.13}$$

where

- $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}})$ designates the genetic variance between FS-families, *i.e.* the variance of the genotypic values of the FS-families, where $\underline{\mathcal{G}}_{\mathrm{FS}}$ is defined to be equal to the expected genotypic value of the plants representing some FS-family. One may write

$$\underline{\mathcal{G}}_{\mathrm{FS}} = \mathrm{E}(\underline{\mathcal{G}}|\mathrm{FS})$$

- $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{FS})})$ designates the expected genetic variance within FS-families.

N.B. The formulation 'expected genetic variance within FS-families' is incidentally used.

Indeed, the genetic variance within a FS-family depends on the genotypes of its parents. In Note 10.3 it is derived that

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}}) = \tfrac{1}{2}\sigma_a^2 + \tfrac{1}{4}\sigma_d^2 \tag{10.14}$$

implying:

$$\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{FS})}) = \tfrac{1}{2}\sigma_a^2 + \tfrac{3}{4}\sigma_d^2 \tag{10.15}$$

Note 10.3 For reasons similar to those applying to HS-families (see Note 10.2) one may write with regard to randomly crossed pairs of plants and the resulting FS-families

$$\mathrm{cov}(\underline{p}_{\mathrm{P}}, \underline{p}_{\mathrm{FS}}) = \mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{P}}, \underline{\mathcal{G}}_{\mathrm{FS}})$$

Likewise, it applies that

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}} = \mathrm{E}[\mathrm{E}(\underline{\mathcal{G}}|\mathrm{FS})] = \mathrm{E}\underline{\mathcal{G}}$$

Thus, when considering some locus $B - b$, Equation (9.5) implies

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}} = \mathrm{E}\underline{\mathcal{G}} = m + (p - q)a + 2pqd$$

where
$\underline{\mathcal{G}}_{\mathrm{P}}$ designates the expected genotypic value of a pair of randomly crossed parents.

The genetic variance between FS-families, *i.e.* $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}})$, is derived from Table 10.2.

Table 10.2 The relationship between the average genotypic value of two parental plants ($\mathcal{G}_{\mathrm{P}}$) and the genotypic value of the corresponding FS-family ($\mathcal{G}_{\mathrm{FS}}$), *i.e.* the expected genotypic value of the plants constituting the considered FS-family

Parental plants			FS-family			
			Genotypic composition			
cross	f	$\mathcal{G}_{\mathrm{P}}$	bb	Bb	BB	$\mathcal{G}_{\mathrm{FS}}$
$bb \times bb$	q^4	$m-a$	1	0	0	$m-a$
$bb \times Bb$	$4pq^3$	$m-\frac{1}{2}a+\frac{1}{2}d$	$\frac{1}{2}$	$\frac{1}{2}$	0	$m-\frac{1}{2}a+\frac{1}{2}d$
$bb \times BB$	$2p^2q^2$	m	0	1	0	$m+d$
$Bb \times Bb$	$4p^2q^2$	$m+d$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$m+\frac{1}{2}d$
$Bb \times BB$	$4p^3q$	$m+\frac{1}{2}a+\frac{1}{2}d$	0	$\frac{1}{2}$	$\frac{1}{2}$	$m+\frac{1}{2}a+\frac{1}{2}d$
$BB \times BB$	p^4	$m+a$	0	0	1	$m+a$

Thus $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}}) = \mathrm{E}\underline{\mathcal{G}}^2_{\mathrm{FS}} - (\mathrm{E}\underline{\mathcal{G}})^2$

$$
\begin{aligned}
&= \mathrm{q}^4(-a)^2 + 4pq^3(-\tfrac{1}{2}a+\tfrac{1}{2}d)^2 + 2p^2q^2d^2 + 4p^2q^2(\tfrac{1}{2}d)^2 \\
&\quad +4p^3q(\tfrac{1}{2}a+\tfrac{1}{2}d)^2 + p^4(a)^2 - [(p-q)a+2pqd]^2 \\
&= [q^4+pq^3+p^3q+p^4-(p-q)^2]a^2 + [-2pq^3+2p^3q-4pq(p-q)]ad \\
&\quad +[pq^3+2p^2q^2+p^2q^2+p^3q-4p^2q^2]d^2
\end{aligned}
$$

Application of Equation (2.8) and some simplifications yield:

$$
\begin{aligned}
\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}}) &= pqa^2 - 2pq[q^2-p^2+2(p-q)]ad + pq(q^2+2pq+pq+p^2-4pq)d^2 \\
&= pqa^2 - 2pq(p-q)ad + pq(1-4pq)d^2 + p^2q^2d^2
\end{aligned}
$$

According to Note 10.2 this is equal to:

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}}) = \tfrac{1}{2}\sigma^2_{\mathrm{a}} + \tfrac{1}{4}\sigma^2_{\mathrm{d}}$$

Besides on the basis of Equations (10.14) and (10.15), one may also estimate σ^2_a on the basis of the relationship between pairs of parents and their offspring. Thus we consider the average phenotypic values of random pairs of parental plants, say $\underline{p}_{\mathrm{P}}$, as well as the phenotypic values of the FS-families they produce after pairwise crossing, say $\underline{p}_{\mathrm{FS}}$, where $\underline{p}_{\mathrm{FS}}$ is the mean phenotypic value calculated across the plants constituting the considered FS-family. The relationship between $\underline{p}_{\mathrm{P}}$ and $\underline{p}_{\mathrm{FS}}$ is thus considered. In Note 10.4 it is derived that

$$\mathrm{cov}(\underline{p}_{\mathrm{P}}, \underline{p}_{\mathrm{FS}}) = \tfrac{1}{2}\sigma^2_a \tag{10.16}$$

Note 10.4 Table 10.2 is used to derive $\mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{P}}, \underline{\mathcal{G}}_{\mathrm{FS}})$.

$$\begin{aligned}\mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{P}}, \underline{\mathcal{G}}_{\mathrm{FS}}) &= \mathrm{E}(\underline{\mathcal{G}}_{\mathrm{P}} \cdot \underline{\mathcal{G}}_{\mathrm{FS}}) - (\mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}}) \cdot (\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}}) \\ &= q^4(-a)^2 + 4pq^3(-\tfrac{1}{2}a + \tfrac{1}{2}d)^2 + 4p^2q^2(\tfrac{1}{2}d^2) \\ &\quad +4p^3q(\tfrac{1}{2}a + \tfrac{1}{2}d)^2 + p^4a^2 - [(p-q)\mathrm{a} + 2pqd]^2 \\ &= [p^4 + p^3q + pq^3 + q^4 - (p-q)^2]a^2 + [2p^3q - 2pq^3 - 4pq(p-q)]ad \\ &\quad +[p^3q + 2p^2q^2 + pq^3 - 4p^2q^2]d^2\end{aligned}$$

According to Equation (2.8) and some derivations in Note 10.3 this is equal to:

$$pqa^2 - 2pq(p-q)ad + pq(p^2 + 2pq + q^2 - 4pq)d^2 = \tfrac{1}{2}\sigma_a^2.$$

Thus

$$\mathrm{cov}(\underline{p}_{\mathrm{P}}, \underline{p}_{\mathrm{FS}}) = \tfrac{1}{2}\sigma_a^2$$

Thus, when evaluating FS-families derived from random pairs of plants, estimates for σ_a^2 are:

$$3\hat{\mathrm{v}}\mathrm{ar}(\underline{\mathcal{G}}_{\mathrm{FS}}) - \hat{\mathrm{v}}\mathrm{ar}(\underline{\mathcal{G}}_{(\mathrm{FS})}) \tag{10.17}$$

and

$$2\hat{\mathrm{c}}\mathrm{ov}(\underline{p}_{\mathrm{P}}, \underline{p}_{\mathrm{FS}}) \tag{10.18}$$

10.3 Self-Fertilization

When dealing with the breeding of a self-fertilizing crop, the decision concerning the initial crosses to be made should be made with great care. This was already emphasized in Section 9.3 and is further considered in Section 11.4. Of course the parents should be chosen such that the goal of the breeding programme might be attained. This in turn requires the development of a well-defined goal. One should thus be able to specify in what degree certain characters are desired to change. Often the breeder will distinguish between short-term and long-term objectives. With regard to short-term objectives it might be best to choose parents that will produce, in the segregating populations obtained after the initial crossing, lines approaching the specified goals as close as possible. This simply means that the parents should be similar to the target genotype. For long-term-objective breeding it is most important to cross divergent lines, such that sufficient genetic variation is generated in the segregating generations.

Mostly the choice of parents to be crossed is made on subjective grounds. Efforts to find reliable, objective grounds for parental selection employing mathematical tools (encompassing the calculation of genetic distances between

parents, component analysis (see Bos and Sparnaaij (1993)), index selection or even artificial intelligence) have not been entirely successful. Certainly the important traits of the potential parents need to be evaluated.

It is assumed that the successive generations of a certain population trace back to an initial cross between two pure lines. As long as selection does not occur, the allele frequencies of segregating loci will be $p = q = \frac{1}{2}$. The genotypic composition of generation t, where $t = 1$ for population F_2 (see Tables 3.1 and 9.1), is then completely determined by the inbreeding coefficient $\mathscr{F}_t$. In as far as the K relevant segregating loci are unlinked and non-epistatic, the variance of the genotypic values of the complex genotypes is equal to the sum of contributions due to single loci. The size of these single-locus contributions follows from substituting $p = q = \frac{1}{2}$ in Equations (8.22) and (8.23). The genotypic variance of any generation is consequently:

$$\begin{aligned} var(\underline{\mathcal{G}}) &= \tfrac{1}{2}(1+\mathscr{F}_t)\sum_{i=1}^{K} {a_i}^2 + \left(\tfrac{1-\mathscr{F}_t}{1+\mathscr{F}_t}\right)\sum_{i=1}^{K} {d_i}^2 \left[\mathscr{F}_t + \tfrac{1}{4}(1-\mathscr{F}_t)^2\right] \\ &= \tfrac{1}{2}\,(1+\mathscr{F}_t)\sum_{i=1}^{K} {a_i}^2 + \left(\tfrac{1-\mathscr{F}_t}{1+\mathscr{F}_t}\right)\left(\tfrac{1}{2}(1+\mathscr{F}_t)\right)^2 \sum_{i=1}^{K} {d_i}^2 \\ &= \tfrac{1}{2}\,(1+\mathscr{F}_t)\sum_{i=1}^{K} {a_i}^2 + \tfrac{1}{4}\left(1-\mathscr{F}_t^2\right)\sum_{i=1}^{K} {d_i}^2 \end{aligned} \tag{10.19}$$

It appears that var($\underline{\mathcal{G}}$) consists of two components, $\sum_i {a_i}^2$ and $\sum_i {d_i}^2$, with coefficients depending on the inbreeding coefficient $\mathscr{F}_t$, *i.e.* on the considered generation. (The expected genotypic value was also shown to be a simple function of $\mathscr{F}_t$, see Equation (9.11).)

With continued selfing the value of $\mathscr{F}_t$ in successive generations follows from Equation (3.4), *i.e.* $\mathscr{F}_t = \frac{1}{2}(1+\mathscr{F}_{t-1})$, where the inbreeding coefficient of generation 1, *i.e.* F_2, is 0. Substitution of the appropriate value for $\mathscr{F}_t$ in Equation (10.19) yields the genotypic variance in a certain generation of a self-fertilizing crop (Table 10.3)

If

$$\sum_i {a_i}^2 \geq \sum_i {d_i}^2$$

var($\underline{\mathcal{G}}$) will gradually increase in course of the generations.

Component $\sum_i {a_i}^2$ of var($\underline{\mathcal{G}}$) is equal to var($\underline{\mathcal{G}}_{F_\infty}$). It represents the genetic variance of the completely homozygous plant material eventually obtained if, indeed, selection is not applied. Knowledge of var($\underline{\mathcal{G}}_{F_\infty}$) *i.e.* of $\sum_i {a_i}^2$, in an early stage of the breeding process, before selection has even started, is of great interest to the breeder because it allows calculation of the probability of occurrence, in the F_∞-population yet to be obtained, of plant material with a superior genotypic value (Section 11.4.2). For this reason estimation of $\sum_i {a_i}^2$ in an early generation, on the basis of partitioning of var($\underline{\mathcal{G}}$), is considered.

Table 10.3 The genotypic variance ($\mathrm{var}(\underline{\mathcal{G}})$) of successive generations of a self-fertilizing crop. The inbreeding coefficients ($\mathscr{F}_t$) are derived from Table 3.1b

Generation	*Population*	$\mathscr{F}_t$	$\mathrm{var}(\underline{\mathcal{G}})$
0	F_1	-1	0
1	F_2	0	$\frac{1}{2}\sum_i {a_i}^2 + \frac{1}{4}\sum_i {d_i}^2$
2	F_3	$\frac{1}{2}$	$\frac{3}{4}\sum_i {a_i}^2 + \frac{3}{16}\sum_i {d_i}^2$
3	F_4	$\frac{3}{4}$	$\frac{7}{8}\sum_i {a_i}^2 + \frac{7}{64}\sum_i {d_i}^2$
4	F_5	$\frac{7}{8}$	$\frac{15}{16}\sum_i {a_i}^2 + \frac{15}{256}\sum_i {d_i}^2$
.			
∞	F_∞	1	$\sum_i {a_i}^2$

The partitioning is elaborated in Section (10.3.1); the actual estimation of $\sum_i {a_i}^2$ is dealt with in Section 11.2.3.

N.B. The quantity $\sum_i {d_i}^2$ is not of much practical interest because this component of $\mathrm{var}(\underline{\mathcal{G}})$ is due to heterozygous plants, which are bound to disappear with continued self-fertilization. It plays however a role in efforts to estimate the range of genotypic values (see Section 11.4.2).

10.3.1 Partitioning of $\sigma_g{}^2$ in the case of self-fertilization

In the partitioning of $\mathrm{var}(\underline{\mathcal{G}})$ allowing estimation of $\sum_i {a_i}^2$, separate plants, representing generation t, *i.e.* representing population F_{t+1}, produce the lines constituting generation $t+1$ (population F_{t+2}). Then the genotypic variance in population F_{t+2} may be partitioned as

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}}) + \mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$$

where

- $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$ designates the genetic variance between lines, *i.e.* the variance of the genotypic values of the lines, where $\underline{\mathcal{G}}_{\mathrm{L}}$ is defined to be equal to the expected genotypic value of the plants representing some line.
- $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$ designates the expected genetic variance within lines. (The formulation 'expected genetic variance within lines' is used, as the genetic variance within a line depends on the number of heterozygous loci in the parental plant. This number varies across the plants (see Section 3.2.3). The genetic variance within a line will, consequently, vary across the lines.)

In Note 10.5 it is derived that the genetic variance between the lines constituting population F_{t+2} can be written as

$$var(\underline{\mathcal{G}}_L) = \tfrac{1}{2}(1+\mathscr{F}_t)\sum_i {a_i}^2 + \tfrac{1}{16}(1-{\mathscr{F}_t}^2)\sum_i {d_i}^2 \qquad (10.20)$$

Note 10.5 The components $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}})$ and $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$ of $\mathrm{var}(\underline{\mathcal{G}})$ are derived for the lines obtained by self-fertilization of plants representing generation t (population F_{t+1}). The derivation proceeds with the help of Table 10.4.

Table 10.4 The relationship between the genotypic value of a parental plant occurring in generation t, *i.e.* $\underline{\mathcal{G}}_{\mathrm{P}}$, and the genotypic value of the corresponding line ($\underline{\mathcal{G}}_{\mathrm{L}}$), *i.e.* the expected genotypic value of the plants constituting the considered line; as well as the expected genetic variance within the line, *i.e.* $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$

			Line				
Parental plant			Genotypic composition				
genotype	f	$\mathcal{G}_{\mathrm{P}}$	bb	Bb	BB	$\mathcal{G}_{\mathrm{L}}$	$\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$
bb	$\frac{1}{4}(1+\mathscr{F}_t)$	$m-a$	1	0	0	$m-a$	0
Bb	$\frac{1}{2}(1-\mathscr{F}_t)$	$m+d$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$m+\frac{1}{2}d$	$\frac{1}{2}a^2+\frac{1}{4}d^2$
BB	$\frac{1}{4}(1+\mathscr{F}_t)$	$m+a$	0	0	1	$m+a$	0

The quantity to be derived is

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}}) = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}} - m) = \mathrm{E}(\underline{\mathcal{G}}_{\mathrm{L}} - m)^2 - [\mathrm{E}(\underline{\mathcal{G}}_{\mathrm{L}} - m)]^2$$

where

$$\begin{aligned}\mathrm{E}(\underline{\mathcal{G}}_{\mathrm{L}} - m)^2 = {} & \tfrac{1}{4}(1+\mathscr{F}_t)(-a)^2 + \tfrac{1}{2}(1-\mathscr{F}_t)(\tfrac{1}{2}d)^2 \\ & + \tfrac{1}{4}(1+\mathscr{F}_t)a^2 = \tfrac{1}{2}(1+\mathscr{F}_t)a^2 + \tfrac{1}{8}(1-\mathscr{F}_t)d^2\end{aligned}$$

and

$$[\mathrm{E}(\underline{\mathcal{G}}_{\mathrm{L}} - m)]^2 = [\tfrac{1}{2}(1-\mathscr{F}_t)(\tfrac{1}{2}d)]^2 = \tfrac{1}{16}(1-\mathscr{F}_t)^2 d^2$$

This yields

$$\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}}) = \tfrac{1}{2}(1+\mathscr{F}_t)a^2 + \tfrac{1}{16}(1-\mathscr{F}_t{}^2)d^2$$

It is easy to see that the expected genetic variance within lines amounts to

$$\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})} = \tfrac{1}{4}(1-\mathscr{F}_t)a^2 + \tfrac{1}{8}(1-\mathscr{F}_t)d^2$$

and the expected genetic variance within these lines as

$$var(\underline{\mathcal{G}}_{(L)}) = \tfrac{1}{4}(1-\mathscr{F}_t)\sum_i a_i{}^2 + \tfrac{1}{8}(1-\mathscr{F}_t)\sum_i d_i{}^2 \qquad (10.21)$$

The appropriate value of the coefficient of inbreeding is the value applying to the parental generation, *i.e.* generation t. The derivation in Note 10.5 is in terms of a single locus. In Section 10.1 it was explained that the resulting equations can be extended to any number of unlinked, non-epistatic loci.
Verification of the equation

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}}) + \mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$$

proceeds for Equations (10.20) and (10.21), which are in terms of the inbreeding coefficient of the parental population (generation t), as follows:

$$\begin{aligned} var(\underline{\mathcal{G}}_L) + var(\underline{\mathcal{G}}_{(L)} &= \tfrac{1}{2}(1+\mathscr{F}_t)\sum_i {a_i}^2 + \tfrac{1}{16}(1-\mathscr{F}_t^2)\sum_i {d_i}^2 \\ &\quad + \tfrac{1}{4}(1-\mathscr{F}_t)\sum_i {a_i}^2 + \tfrac{1}{8}(1-\mathscr{F}_t)\sum_i {d_i}^2 \\ &= \left(\tfrac{3}{4}+\tfrac{1}{4}\mathscr{F}_t\right)\sum_i {a_i}^2 + \left(\tfrac{3}{16}-\tfrac{1}{8}\mathscr{F}_t-\tfrac{1}{16}{\mathscr{F}_i}^2\right)\sum_i {d_i}^2 \end{aligned} \tag{10.22}$$

As Equation (3.4), *i.e.*

$$\mathscr{F}_t = \tfrac{1}{2}(1+\mathscr{F}_{t-1})$$

implies

$$\mathscr{F}_{t+1} = \tfrac{1}{2}(1+\mathscr{F}_t)$$

we get

$$\mathscr{F}_t = 2\mathscr{F}_{t+1} - 1$$

Substitution in Equation (10.22) of

$$\mathscr{F}_t$$

by

$$2\mathscr{F}_{t+1} - 1$$

yields the following equation for var($\underline{\mathcal{G}}$) in terms of generation $t+1$:

$$\begin{aligned} var(\underline{\mathcal{G}}) &= \left[\tfrac{3}{4}+\tfrac{1}{4}(2\mathscr{F}_{t+1}-1)\right]\sum_i {a_i}^2 + \left[\tfrac{3}{16}-\tfrac{1}{8}(2\mathscr{F}_{t+1}-1)\right. \\ &\quad \left.-\tfrac{1}{16}(2\mathscr{F}_{t+1}-1)^2\right]\sum_i {d_i}^2 \\ &= \tfrac{1}{2}(1+\mathscr{F}_{t+1})\sum_i {a_i}^2 + \tfrac{1}{4}(1-{\mathscr{F}_{t+1}}^2)\sum_i {d_i}^2 \end{aligned}$$

This equation is in accordance with Equation (10.19).

For reasons similar to those applying to HS-families (see Note 10.2) one may write with regard to random parental plants and their lines, *i.e.* their offspring obtained by selfing,

$$\mathrm{cov}(\underline{p}_\mathrm{P}, \underline{p}_\mathrm{L}) = \mathrm{cov}(\underline{\mathcal{G}}_\mathrm{P}, \underline{\mathcal{G}}_\mathrm{L})$$

The covariance between the genotypic value of a random parental plant occurring in generation t, and the expected genotypic value of the line obtained from the plant is derived in Note 10.6.

Note 10.6 In the absence of correlation of genotypic value and environmental deviation the following applies to the covariance of $\underline{p}_{\mathrm{P}}$ and $\underline{p}_{\mathrm{L}}$:

$$\mathrm{cov}(\underline{p}_{\mathrm{P}}, \underline{p}_{\mathrm{L}}) = \mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{P}}, \underline{\mathcal{G}}_{\mathrm{L}})$$

Using Table 10.4 one can derive

$$\begin{aligned}\mathrm{cov}(\underline{\mathcal{G}}_{\mathrm{P}}, \underline{\mathcal{G}}_{\mathrm{L}}) &= \mathrm{E}(\underline{\mathcal{G}}_{\mathrm{P}} \cdot \underline{\mathcal{G}}_{\mathrm{L}}) - (\mathrm{E}\underline{\mathcal{G}}_{\mathrm{P}}) \cdot (\mathrm{E}\underline{\mathcal{G}}_{\mathrm{L}}) \\ &= \tfrac{1}{2}(1+\mathscr{F}_t)a^2 + \tfrac{1}{4}(1-\mathscr{F}_t)d^2 - [\tfrac{1}{2}(1-\mathscr{F}_t)d][\tfrac{1}{4}(1-\mathscr{F}_t)d] \\ &= \tfrac{1}{2}(1+\mathscr{F}_t)a^2 + (1-\mathscr{F}_t{}^2)d^2\end{aligned}$$

It appears that

$$cov(\underline{p}_P, \underline{p}_L) = \tfrac{1}{2}(1+\mathscr{F}_t)\sum_i a_i{}^2 + \tfrac{1}{8}(1-\mathscr{F}_t^2)\sum_i d_i{}^2 \tag{10.23}$$

The gradual increase in over the course of the generations of $\mathrm{var}(\underline{\mathcal{G}})$, at

$$\sum_i a_i{}^2 \geq \sum_i d_i{}^2$$

is the result of a progressing increase of $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{L}})$ and decrease of $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{L})})$.

The earliest opportunity for generating lines is offered by the F_2 population, generation 1. The appropriate value of the inbreeding coefficient, to be substituted in Equations (10.20), (10.21) and (10.23), is then $\mathscr{F}_1$, *i.e.* 0. This yields

$$var(\underline{\mathcal{G}}_{\mathrm{LF3}}) = \tfrac{1}{2}\sum_i a_i{}^2 + \tfrac{1}{16}\sum d_i{}^2 \tag{10.24}$$

$$var(\underline{\mathcal{G}}_{(\mathrm{LF3})}) = \tfrac{1}{4}\sum_i a_i{}^2 + \tfrac{1}{8}\sum_i d_i{}^2 \tag{10.25}$$

Indeed

$$var(\underline{\mathcal{G}}_{\mathrm{F3}}) = var(\underline{\mathcal{G}}_{\mathrm{LF3}}) + var(\underline{\mathcal{G}}_{(\mathrm{LF3})}) = \tfrac{3}{4}\sum_i a_i{}^2 + \tfrac{3}{16}\sum_i d_i{}^2$$

(as indicated by Table 10.3)
An unbiased estimate for $\sum_i a_i{}^2$, based on the equation

$$2var(\underline{\mathcal{G}}_{\mathrm{LF3}}) - var(\underline{\mathcal{G}}_{(\mathrm{LF3})}) = \tfrac{3}{4}\sum_i a_i{}^2 \tag{10.26}$$

requires estimates of $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{LF3}})$ and $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{LF3})})$. It is rather demanding to get accurate and unbiased estimates of these genetic variance components.

An alternative procedure for estimating $\sum_i {a_i}^2$ is therefore proposed in Section 11.2.3.

The covariance between $\underline{p}_{\mathrm{PF2}}$, *i.e.* the phenotypic value of a random F_2 plant, and $\underline{p}_{\mathrm{LF3}}$, *i.e.* the phenotypic value of the derived F_3-line, is

$$cov(\underline{p}_{\mathrm{PF}_2}, \underline{p}_{\mathrm{LF}_3}) = \tfrac{1}{2} \sum_i {a_i}^2 + \tfrac{1}{8} \sum_i {d_i}^2 \tag{10.27}$$

The quantity

$$\sum_i {d_i}^2$$

can be estimated from the equation

$$2var(\underline{\mathcal{G}}_{(\mathrm{LF3})}) - var(\underline{\mathcal{G}}_{\mathrm{LF3}}) = \tfrac{3}{16} \sum_i {d_i}^2 \tag{10.28}$$

The latter equation might be used to estimate, from an estimate for $\sum_i {d_i}^2$, the quantity $\sum_i a_i$ (see Section 11.4.2).

In studies dedicated to the estimation of $\sum_i {a_i}^2$ or $\sum_i {d_i}^2$, the estimator is often based on different equations in terms of $\sum_i {a_i}^2$ or $\sum_i {d_i}^2$. Estimation of $\sum_i {a_i}^2 = \mathrm{var}(\underline{\mathcal{G}}_{\mathrm{F}\infty})$ from data obtained from plants belonging to an earlier generation than F_∞ is possible in various ways, but an estimate on the basis of F_3 plant material, due to an unbiased estimator, is considered to be most attractive because that estimate can be obtained far ahead of the actual presence of the F_∞ population. In this case $\sum_i {a_i}^2$ is estimated from Equation (10.26):

$$2var(\underline{\mathcal{G}}_{\mathrm{LF3}}) - var(\underline{\mathcal{G}}_{(\mathrm{LF3})}) = \tfrac{3}{4} \sum_i {a_i}^2$$

It requires estimation of $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{LF3}})$ and of $\mathrm{var}(\underline{\mathcal{G}}_{(\mathrm{LF3})})$. It is rather demanding to get accurate and unbiased estimates of these variance components. A possible approach could be to estimate each of these genetic variance components by subtracting from the corresponding estimates of phenotypic variance an appropriate estimate of the environmental variance.

For plant breeders this approach is unattractive because it requires too large an effort. In Section 11.2.3 a procedure for estimating $\sum_i {a_i}^2$ from F_3 plant material is described that

- fits into a regular breeding programme,
- avoids separate estimation of components of environmental variance and
- yields an accurate estimate.

Chapter 11
Applications of Quantitative Genetic Theory in Plant Breeding

In the preceding chapters dealing with traits with quantitative variation, a number of important concepts were introduced, such as phenotypic value and genotypic value (Chapter 8), expected genotypic value (Chapter 9) and genotypic variance (Chapter 10). The present chapter focusses on applications of these concepts that are important in the context of this book. Thus the response to selection, both its predicted and its actual value, is considered. The prediction of the response is based on estimates of the heritability. Procedures for the estimation of this quantity are elaborated for plant material that can identically be reproduced (clones of crops with vegetative reproduction, pure lines of self-fertilizing crops and single-cross hybrids). It is shown how the heritability value depends on the number of replications.

In addition to the partitioning of the genotypic value in terms of parameters defined in the framework of the F_∞-metric (Section 8.3.2), or in terms of additive genotypic value and dominance deviation (Section 8.3.3), here the rather straightforward partitioning in terms of general combining ability and specific combining ability is elaborated.

11.1 Prediction of the Response to Selection

When dealing with selection with regard to quantitative variation the concepts of **selection differential**, designated by S, and **response to selection**, designated by R, play a central role. These concepts, see also Fig. 11.1, are defined as follows:

$$S := \mathrm{E}\underline{p}_{s,t} - \mathrm{E}\underline{p}_t \tag{11.1}$$

$$R := \mathrm{E}\underline{p}_{t+1} - \mathrm{E}\underline{p}_t \tag{11.2}$$

where

- $\mathrm{E}\underline{p}_{s,t}$ designates the expected phenotypic value of the candidates (plants, clones, families or lines) in generation t of the considered population with a phenotypic value greater than the phenotypic value minimally required for selection (p_{min}). $\mathrm{E}\underline{p}_{s,t}$ designates thus the expected phenotypic value of the selected candidates.
- $\mathrm{E}\underline{p}_t$ designates the expected phenotypic value calculated across all candidates belonging to generation t of the population subjected to selection.
- $\mathrm{E}\underline{p}_{t+1}$ designates the expected phenotypic value calculated across the offspring of the selected candidates.

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 225–287.

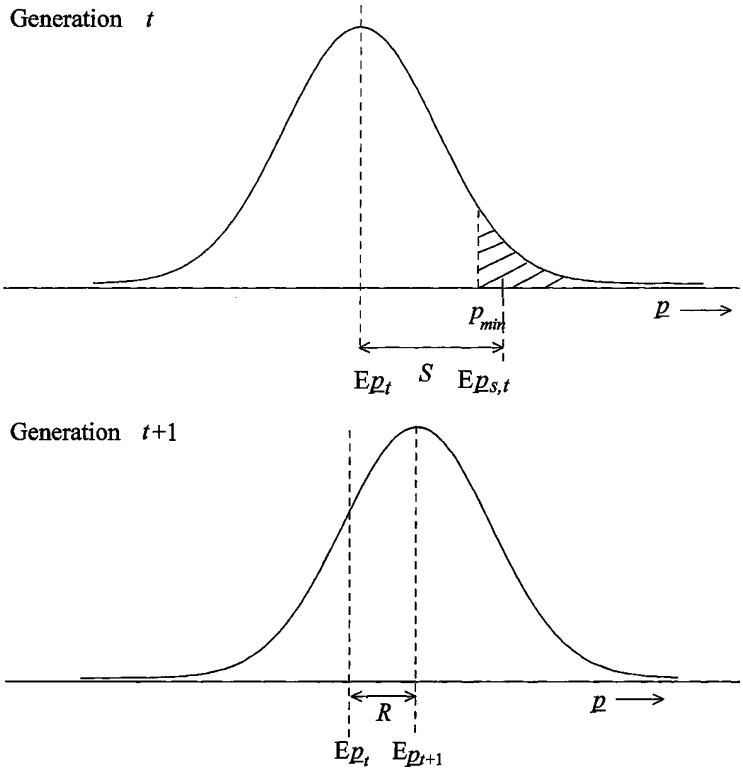


Fig. 11.1 The density function for the phenotypic value $\underline{p}$ in generation t and in generation $t+1$, obtained by selecting in generation t all candidates with a phenotypic value greater than p_{min}. The selection differential (S) in generation t and the response to the selection (R) are indicated. The shaded area represents the probability that a candidate has a phenotypic value larger than the minimally required phenotypic value (p_{min})

In Section 8.2 it was derived that

$$\mathrm{E}\underline{p} = \mathrm{E}\underline{\mathcal{G}}$$

This implies that one may write $\mathrm{E}\underline{\mathcal{G}}_t$ instead of $\mathrm{E}\underline{p}_t$ and $\mathrm{E}\underline{\mathcal{G}}_{t+1}$ instead of $\mathrm{E}\underline{p}_{t+1}$.

The quantities $\mathrm{E}\underline{p}_{s,t}, \mathrm{E}\underline{p}_t$ and $\mathrm{E}\underline{p}_{t+1}$, *i.e.* the quantities S and R, can be estimated from the phenotypic values of a random sample of the (selected) candidates and their offspring, *i.e.* from $\overline{p}_t, \overline{p}_{s,t}$ and $\overline{p}_{t+1}$, As the symbol $\hat{R}$ will be used to indicate the predicted response to selection, the values estimated for S and R will be written in terms of $\overline{p}_t, \overline{p}_{s,t}$ and $\overline{p}_{t+1}$.

The response to selection is now considered for three situations:

1. The hypothetical case of absence of environmental deviations, as well as absence of dominance and epistasis
2. Absence of environmental deviations, presence of dominance and/or epistasis
3. Presence of environmental deviations, dominance and/or epistasis

Absence of environmental deviations, dominance and epistasis

In the absence of environmental deviations, dominance and epistasis, both the genotypic value and the phenotypic value of a candidate can be described by a linear combination of the parameters $a_1, \ldots, a_K$ defined in Section 8.3.2. Selection of candidates with the highest possible phenotypic value implies selection of candidates with genotype $B_1B_1 \ldots B_KB_K$ and with genotypic value $m + \sum_{i=1}^{K} a_i$. The offspring of these candidates will have the same phenotypic and genotypic value as their parents. This applies to self-fertilizing crops as well as cross-fertilizing crops, when the selection occurs before pollen distribution. Under the described conditions R will be equal to S.

Absence of environmental deviations, presence of dominance and/or epistasis

In the case of absence of environmental deviations but presence of dominance and/or epistasis, selected candidates, with the same highest possible phenotypic value, may have a homozygous or a heterozygous genotype. Then the offspring of the selected candidates are expected to comprise plants with genotype bb for one or more loci, giving rise to an inferior phenotypic value compared to that of the selected candidates. In the case of complete dominance, for instance, candidates with the highest possible phenotypic value for a trait controlled by loci $B_1 - b_1$ and $B_2 - b_2$ will have genotype $B_1 \cdot B_2 \cdot$. Selection of such candidates will yield offspring including plants with genotype $b_1b_1b_2b_2$, $b_1b_1B_2 \cdot$ or $B_1 \cdot b_2b_2$, having an inferior genotypic and phenotypic value. Under these conditions R will be less than S.

Presence of environmental deviations, dominance and/or epistasis

In actual situations environmental deviations, dominance and epistasis should be expected to be present. Among the selected candidates their phenotypic values will tend to be (much) higher than their genotypic values. Furthermore, except in the case of identical reproduction, the genotypic composition of the selected candidates will deviate from that of their offspring. Under these conditions R will be (much) smaller than S.

Selected maternal plants coincide with the selected paternal plants in the case of self-fertilizing crops, as well as in case of hermaphroditic cross-fertilizing

crops if the selection is applied before pollen distribution. In other situations, the set of selected maternal parents providing the eggs differs from the set of selected paternal parents providing the pollen. Then one should determine S_f for the candidates selected as maternal parents and S_m for the candidates selected as paternal parents. Because both sexes contribute equal numbers of gametes to generate the next generation we may write

$$S = \tfrac{1}{2}(S_f + S_m) \tag{11.3}$$

Equation (11.3) does not only apply at selection in dioecious crops, but also when selecting in hermaphroditic cross-fertilizing crops when the selection is done after pollen distribution. In the latter case there is no selection with regard to paternal parents. This implies $S_m = 0$ and consequently $S = \frac{1}{2}S_f$.

Actual situations tend to be more complicated. Consider selection before pollen distribution with regard to some trait X. In the case of an association between the expression for trait X and the expression for trait Y, the selection differential for X implies a **correlated selection differential** with regard to Y, say *CS*. Thus

$$CS_{\mathrm{Y}} := \mathrm{E}\underline{p}_{\mathrm{Y}_{s,t}} - \mathrm{E}\underline{p}_{\mathrm{Y},t} \tag{11.4}$$

where

- $\mathrm{E}\underline{p}_{Y_{s,t}}$ designates the expected phenotypic value with regard to trait Y of the candidates selected in generation t because their phenotypic value with regard to trait X being greater than minimally phenotypic value (p_{Xmin}) and
- $\mathrm{E}\underline{p}_t$ designates the expected phenotypic value with regard to trait Y calculated across all candidates belonging to generation t of the population subjected to selection with regard to trait X.

When considering a linear relationship between the phenotypic values for traits X and Y, the coefficient of regression of $\underline{p}_{\mathrm{Y}}$ on $\underline{p}_{\mathrm{X}}$, *i.e.*

$$\beta_{p_Y,p_X} = \frac{cov(\underline{p}_Y, \underline{p}_X)}{var(\underline{p}_X)}$$

may be used to write

$$CS_Y = \beta_{p_Y,p_X} S_X$$

The **indirect selection** (see Section 12.3) for trait Y, via trait X, may be followed, after pollen distribution, by direct selection for Y. The effective selection differential for Y comprises then a correlated selection differential. Example 11.1 presents an illustration.

Example 11.1 Van Hintum and Van Adrichem (1986) applied selection in two populations of maize with the goal of improving biomass.

Population A consisted of 1184 plants. Mass selection for biomass (say trait Y) was applied at the end of the growing season, *i.e.* after pollen

distribution. The mean biomass (in g/plant), calculated across all plants, was $\overline{p}_{\mathrm{Y}} = 245\,\mathrm{g}$. For the 60 selected plants it amounted to $\overline{p}_{\mathrm{Ys}} = 446\,\mathrm{g}$. Thus

$$S_f = 446 - 245 = 201\,\mathrm{g}$$

and

$$S_m = 0\,\mathrm{g}$$

This implies

$$S_{\mathrm{Y}} = \tfrac{1}{2}(201 + 0) = 100.5\,\mathrm{g}.$$

Population B consisted of 1163 plants. Immediately prior to pollen distribution the following was done. The volumes of the plants (say trait X) were roughly calculated from their stalk diameter and their height. The 181 plants with the highest phenotypic values for X were identified. These plants were selected as paternal parents. The 982 other plants were emasculated by removing the tassels. At the end of the growing season among all 1163 plants, the 60 plants with the highest biomass were selected. For the 1163 plants of population B it was found that:

$$\overline{p}_{\mathrm{Y}} = 246\,\mathrm{g},$$

and

$$\overline{p}_{\mathrm{X}} = 599\,\mathrm{cm}^3.$$

For the 181 plants selected as paternal parents (because of superiority for X) it was established that:

$$\overline{p}_{\mathrm{Y}s} = 320\,\mathrm{g},$$
$$\overline{p}_{\mathrm{X}s} = 983\,\mathrm{cm}^3,$$

and

$$CS_{\mathrm{Y}_m} = 320 - 246 = 74\,\mathrm{g}.$$

For the 60 plants selected for Y the following was established:

$$\overline{p}_{\mathrm{Ys}} = 418\,\mathrm{g}$$
$$\overline{p}_{\mathrm{X}s} = 931\,\mathrm{cm}^3$$

and

$$S_{\mathrm{Y}f} = 418 - 246 = 172\,\mathrm{g}$$

The selection differential in population B amounted thus to

$$S_{\mathrm{Y}} = \tfrac{1}{2}(74 + 172) = 123\,\mathrm{g}$$

Due to the correlated selection differential because of selection among the paternal parents with regard to trait X, this is clearly higher than the selection differential in population A.

If the considered trait has a normal distribution, $E\underline{p}_{s,t}$, *i.e.* the expected phenotypic value of those candidates with a phenotypic value larger than the value minimally required for selection, may be calculated prior to the actual selection. This will now be elaborated.

A normal distribution of the phenotypic values for some trait is often designated by

$$\underline{p} = N(\mu, \sigma^2)$$

where

- $\mu = E\underline{p}$, and
- $\sigma^2 = \text{var}(\underline{p})$.

Standardization, *i.e.* the transformation of $\underline{p}$ into $\underline{z}$ according to

$$\frac{\underline{p} - \mu}{\sigma} = \underline{z}$$

implies that $\underline{z}$ has a standard normal distribution characterized by

$$\mu_z = 0 \text{ and}$$
$$\sigma_z = 1.$$

Thus

$$\underline{z} = N(0, 1).$$

Selection of candidates with a phenotypic value exceeding the phenotypic value minimally required for selection (p_{min}) is called **truncation selection**. Selection of superior performing candidates up to a proportion v implies applying a value for p_{min} such, that

$$v = P(\underline{p} > p_{min})$$

Standardization of p_{min} yields the standardized minimum phenotypic value z_{min}:

$$z_{\min} = \frac{p_{\min} - \mu}{\sigma} \tag{11.5}$$

Thus

$$v = P\left(\underline{p} > p_{\min}\right) = P\left(\underline{z} > z_{\min}\right) = \int_{z_{\min}}^{\infty} f(z).dz$$

where

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

is the density function of the standard normal random variate $\underline{z}$.

In Fig. 11.1 the shaded area corresponds with v. Most statistical handbooks (*e.g.* Kuehl, 2000, Table I) contain for the standard normal random variate $\underline{z}$

a table presenting z_{min} such $\mathrm{P}(\underline{z} > z_{min})$ is equal to some specified value v. Then one can calculate p_{min} according to

$$p_{min} = \mu + \sigma z_{min} \tag{11.6}$$

Example 11.2 gives an illustration of this.

Example 11.2 It was desired to select the 168 best yielding plants from the 5016 winter rye plants occurring at the central plant positions of the population which is mentioned in Example 11.7. The proportion to be selected amounted thus to:

$$v = \frac{168}{5016} = 0.0335$$

The standardized minimum phenotypic value z_{min} should thus obey:

$$0.0335 = \mathrm{P}(\underline{z} > z_{min})$$

According to the appropriate statistical table, his implies

$$z_{min} = 1.83.$$

The mean and the standard deviation of the phenotypic values for grain yield were calculated to be 50 dg and 28.9 dg, respectively. When assuming a normal distribution for grain yield, substitution of these values in Equation (11.5) yielded:

$$p_{min} = 50 + (28.9 \times 1.83) = 102.9\,\mathrm{dg}.$$

To measure the selection differential in a scale-independent yardstick, a parameter, called **selection intensity** and designated by the symbol i, has been defined:

$$i = \frac{S}{\sigma} \tag{11.7}$$

There is a simple relationship between the proportion of selected candidates (v) and i if the phenotypic values of the considered trait follow a normal distribution, namely

$$i = \frac{f(z_{min})}{v} \tag{11.8}$$

where $f(z_{min})$ represents the value at $z = z_{min}$ of the density function of the standard normal random variate $\underline{z}$. Equation (11.8) is derived in Note 11.1.

Note 11.1 Equation (11.6) implies that, in the case of a normal distribution of the phenotypic values, the expected phenotypic value of candidates with a phenotypic value larger than p_{min} amounts to

$$\mathrm{E}\underline{p}_{s,t} = \mathrm{E}(\underline{p}|\underline{p} > p_{min}) = \mu + \sigma \mathrm{E}\underline{z}_{s,t}$$

where

- p_{min} may be obtained from Equation (11.5)
- $\mathrm{E}\underline{z}_{s,t} = \mathrm{E}(\underline{z}|\underline{z} > z_{min})$, where z_{min} follows from Equation (11.5)

The quantity $\mathrm{E}\underline{z}_{s,t}$ is now derived.

The density function of the conditional random variable $(\underline{z}|\underline{z} > z_{min})$ is

$$f(z|\underline{z} > z_{min}) = \frac{f(z)}{P(\underline{z} > z_{min})} = \frac{f(z)}{v}$$

Thus

$$\begin{aligned} E\underline{z}_s &= E(\underline{z}|\underline{z} > z_{\min}) = \int_{z=z_{\min}}^{\infty} z f(z|\underline{z} > z_{\min}) dz = \int_{z_{\min}}^{\infty} z \frac{f(z)}{v} dz \\ &= \frac{1}{v\sqrt{2\pi}} \cdot \int_{z_{\min}}^{\infty} z e^{-\frac{1}{2}z^2} dz = \frac{1}{v\sqrt{2\pi}} \cdot \int_{z_{\min}}^{\infty} e^{-\frac{1}{2}z^2} d\left(\frac{1}{2}z^2\right) \\ &= \frac{-1}{v\sqrt{2\pi}} \left[e^{-\frac{1}{2}z^2}\right]_{z=z_{\min}}^{\infty} = \frac{-1}{v\sqrt{2\pi}} \left[0 - e^{-\frac{1}{2}z_{\min}^2}\right] = \frac{f(z_{\min})}{v} \end{aligned}$$

This means that

$$E\underline{p}_{s,t} = \mu + \sigma \left[\frac{f(z_{\min})}{v}\right]$$

Because $\mu = \mathrm{E}\underline{p}$, Equation (11.1) can be written as

$$S = \sigma \left[\frac{f(z_{\min})}{v}\right]$$

Thus when applying truncation selection with regard to a trait with a normal distribution and selecting the proportion v the selection intensity is:

$$i = \frac{f(z_{\min})}{v} = E\underline{z}_{s,t}$$

One can easily calculate i for any value for v and next $\mathrm{E}\underline{p}_{s,t} = \mu + \sigma i$, see Example 11.3. Falconer (1989, Appendix Table A) presents a table for the relation between i and v.

Example 11.3 In Example 11.2 it was derived that the standardized minimum phenotypic value z_{min} is 1.83 when selecting the proportion $v = 0.0335$. In the case of a normal distribution of the phenotypic values the selection intensity amounts then to

$$\frac{f(1.83)}{0.0335} = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1.83)^2}}{0.0335} = \frac{0.3989 \times 0.1874}{0.0335} = 2.232$$

Thus

$$\mathrm{E}\underline{p}_s = 50 + 28.9 \times 2.232 = 114.5\,\mathrm{dg}.$$

Among the 168 plants with the highest grain yield, the grain yield of the plant with the lowest phenotypic value amounted to 102 dg. The actual minimum phenotypic value was thus 102 dg. Their mean grain yield amounted to 117.5 dg, implying

$$S = 117.5 - 50 = 67.5\,\mathrm{dg}$$

and

$$i = \frac{67.5}{28.9} = 2.34$$

Also the measurement of the response to selection (R) deserves closer consideration. It requires determination of $\mathrm{E}\underline{p}$ in the two successive generations t and $t+1$. To exclude an effect of different growing conditions these two generations should preferably be grown in the same growing season. This is possible by

1. Testing simultaneously plant material representing generation $t+1$ (say population $\mathrm{P'}_{t+1}$), obtained by harvesting candidates selected in generation t, and – from remnant seed – plant material representing generation t (say population P_t)
2. Testing simultaneously plant material representing generation $t+1$, obtained by harvesting candidates selected in generation t (population $\mathrm{P'}_{t+1}$), and plant material, also representing generation $t+1$, obtained by harvesting in generation t random candidates (population P_{t+1})

Simultaneous testing of populations P'_{t+1} and P_t

Measurement of R by simultaneous testing of populations $\mathrm{P'}_{t+1}$ and P_t will be biased if these populations differ due to other causes than the selection. Such differences may be due to

- the fact that the remnant seed is older and has, consequently, lost viability;
- the remnant seed representing P_t was produced under conditions deviating from the conditions prevailing when producing the seed representing $\mathrm{P'}_{t+1}$ or
- a difference in the genotypic compositions of $\mathrm{P'}_{t+1}$ and P_t which is not due to the selection. This is to be expected when dealing with self-fertilizing crops: $\mathrm{P'}_{t+1}$ tends to contain a reduced frequency of heterozygous plants in comparison to P_t.

When testing populations P'_{t+1} and P_t simultaneously, no allowance is made for the possible quantitative genetic effect of the reduction of heterozygosity occurring in self-fertilizing crops.

Simultaneous testing of populations P'_{t+1} and P_{t+1}

The causes for the bias mentioned above do not apply to simultaneous testing of populations P'_{t+1} and P_t. Furthermore, this method allows – for cross-fertilizing crops – estimation of the coefficient of regression of the phenotypic value of offspring on parental phenotypic value. Such an estimate may be interpreted in terms of the narrow sense heritability (Section 11.2.2).

One should realize that R as defined by Equation (11.2) does not represent a lasting response to selection if $\sum_{i=1}^{K} d_i \neq 0$. For self-fertilizing crops populations after generation $t+1$, obtained in the absence of selection, will – due to the ongoing reduction of the frequency of heterozygous plants – tend to have an expected genotypic value deviating from $E\underline{p}_{t+1} = E\underline{p}_t + R$. The same applies to selection after pollen distribution in cross-fertilizing crops: population P'_{t+1} results then from a bulk cross and will, consequently, contain an excess of heterozygous plants compared to population P_{t+2} obtained – in the absence of selection – from population P'_{t+1}. In the case of selection before pollen distribution, population P'_{t+1} is in Hardy–Weinberg equilibrium and P'_{t+1} and P_{t+2} will then, in the absence of epistasis, have the same expected genotypic value.

A procedure to predict R is, of course, of great interest to breeders, because such prediction may be used as a basis for a decision with regard to further breeding efforts dedicated to the plant material in question.

As the prediction is based on linear regression theory, a few important aspects of that theory are reminded. In the case of linear regression of y on x the y-value for some x-value is predicted by

$$\hat{y} = \alpha + \beta x,$$

where

$$\beta = \frac{cov(\underline{x}, \underline{y})}{var(\underline{x})} = \frac{E(\underline{x} \cdot \underline{y}) - (E\underline{x}) \cdot (E\underline{y})}{E\underline{x}^2 - (E\underline{x})^2} \tag{11.9}$$

and, because of

$$E\underline{y} = \alpha + \beta \cdot E\underline{x}$$

the intercept α is equal to

$$\alpha = E\underline{y} - \beta.E\underline{x} \tag{11.10}$$

Thus

$$\hat{y} = (E\underline{y} - \beta \cdot E\underline{x}) + \beta x = E\underline{y} + \beta(x - E\underline{x}) \tag{11.11}$$

implying

$$\hat{y} - E\underline{y} = \beta(x - E\underline{x}) \tag{11.12}$$

This means in the present context

$$\mathrm{E}\underline{p}_{t+1} - \mathrm{E}\underline{p}_t = \beta(\mathrm{E}\underline{p}_{s,t} - \mathrm{E}\underline{p}_t)$$

or

$$R = \beta S \tag{11.13}$$

It is common practice to substitute parameter β in Equation (11.13) either by the wide or by the narrow sense heritability:

1. In the case of identical reproduction, this applies when dealing with clones, pure lines and single-cross hybrids, β is substituted by the ratio $\frac{{\sigma_g}^2}{{\sigma_p}^2}$, *i.e.* the **heritability in the wide sense**, commonly designated by ${h_w}^2$. Thus

$$R = {h_w}^2 S \tag{11.14}$$

   In this situation the genotypes of the selected entries are preserved. Note 11.2 presents the derivation of Equation (11.14).
2. In the case of non-identical reproduction of the selected candidate plants of a cross-fertilizing crop β is substituted by $\frac{{\sigma_a}^2}{{\sigma_p}^2}$, *i.e.* the **heritability in narrow sense**, commonly designated by ${h_n}^2$. Thus

$$R = {h_n}^2 S \tag{11.15}$$

   The possible bias introduced with this substitution is taken for granted.

In Note 11.2 a few interesting results of quantitative genetic theory are derived, namely that amongst the candidates

- the coefficient of correlation of $\underline{\mathcal{G}}$ and $\underline{p}$, *i.e.* $\rho_{g,p}$, is equal to the square root of the heritability in the wide sense:

$$\rho_{g,p} = h_w \tag{11.16}$$

- the coefficient of regression of $\underline{\mathcal{G}}$ on $\underline{p}$, *i.e.* β, is equal to the heritability in the wide sense:

$$\beta = {h_w}^2 \tag{11.17}$$

Note 11.2 The degree of linear association of the genotypic value ($\underline{\mathcal{G}}$) and the phenotypic value ($\underline{p}$) is of course of interest with regard to the success of selection. Indeed, selection intends to improve the expected genotypic value by selecting plants with superior phenotypic values. The coefficient of correlation measures the degree of linear association. In the absence of covariance of genotypic value and environmental deviation, thus at

$$\mathrm{cov}(\underline{\mathcal{G}}, \underline{e}) = 0,$$

the coefficient of correlation of $\underline{\mathcal{G}}$ and $\underline{p}$, *i.e.* $\rho_{g,p}$, amounts to

$$\rho_{g,p} = \frac{cov(\underline{\mathcal{G}}, \underline{p})}{\sigma_g \sigma_p} = \frac{cov(\underline{\mathcal{G}}, \underline{\mathcal{G}} + \underline{e})}{\sigma_g \sigma_p} = \frac{{\sigma_g}^2}{\sigma_g \sigma_p} = \frac{\sigma_g}{\sigma_p} = h_w$$

The coefficient of regression of $\underline{\mathcal{G}}$ on $\underline{p}$, *i.e.* β, amounts to

$$\beta = \frac{cov(\underline{\mathcal{G}}, \underline{p})}{{\sigma_p}^2} = \frac{cov(\underline{\mathcal{G}}, \underline{\mathcal{G}} + \underline{e})}{{\sigma_p}^2} = \frac{{\sigma_g}^2}{{\sigma_p}^2} = {h_w}^2$$

At identical reproduction, the regression of $\underline{p}_{\mathrm{O}}$, *i.e.* the phenotypic value of the offspring, on $\underline{p}_{\mathrm{P}}$, *i.e.* the phenotypic value of the parent, amounts to

$$\frac{cov(\underline{p}_O, \underline{p}_P)}{var(\underline{p}_P)} = \frac{cov(\underline{\mathcal{G}}_O, \underline{\mathcal{G}}_P)}{var(\underline{p}_P)} = \frac{{\sigma_g}^2}{{\sigma_p}^2} = {h_w}^2$$

Equation (11.12) can be rewritten as

$$\hat{y} - E\underline{y} = \frac{cov(\underline{x}, \underline{y})}{{\sigma_x}^2} \cdot (x - E\underline{x})$$

Thus, if one substitutes in

$$\frac{cov(\underline{x}, \underline{y})}{{\sigma_x}^2}$$

$\underline{x}$ by $\underline{p}_{\mathrm{P}}$, $\underline{y}$ by $\underline{p}_{\mathrm{O}}$, $x - \mathrm{E}\underline{x}$ by S, and $\hat{y} - \mathrm{E}\underline{y}$ by R, one gets

$$R = {h_w}^2 S \tag{11.18}$$

In addition to this it is interesting to know that within candidates

- the coefficient of correlation of the additive genotypic value ($\underline{\gamma}$, see Section 8.3.3) and $\underline{p}$, *i.e.* $\rho_{\gamma,p}$, is equal to the square root of the heritability in the narrow sense:

$$\rho_{\gamma,p} = h_n \tag{11.19}$$

(see Note 11.3)

Note 11.3 The coefficient of correlation of the additive genotypic value ($\underline{\gamma}$) and $\underline{p}$, *i.e.* $\rho_{\gamma,p}$, is considered. Application of Equation (8.9), *i.e.*

$$\underline{\mathcal{G}} = \underline{\gamma} + \underline{\delta}$$

implies

$$\rho_{\gamma,p} = \frac{cov(\underline{\gamma}, \underline{p})}{\sigma_a \sigma_p} = \frac{cov(\underline{\gamma}, \underline{\gamma} + \underline{\delta} + \underline{e})}{\sigma_a \sigma_p} = \frac{{\sigma_a}^2}{\sigma_a \sigma_p} = \frac{\sigma_a}{\sigma_p} = h_n$$

Because $S = i\sigma$ (see Equation (11.7), Equation (11.13) can also be written as

$$R = \beta.i\sigma$$

Equation (11.14) can thus be written as

$$R = {h_w}^2 i\sigma_p = ih_w \left(\frac{\sigma_g}{\sigma_p}\right) \sigma_p = ih_w\sigma_g \tag{11.20}$$

When selecting, after pollen distribution, in a cross-fertilizing crop one can similarly write

$$R = \tfrac{1}{2}i{h_n}^2\sigma_p = \tfrac{1}{2}ih_n \left(\frac{\sigma_a}{\sigma_p}\right) \sigma_p = \tfrac{1}{2}ih_n\sigma_a \tag{11.21}$$

Higher selection intensities occur at lower proportions of selected plants. One should thus be careful when using the terms 'selection intensity' and 'proportion selected candidates.'

In the situation of non-identical reproduction of plants belonging to an early segregating population of a self-fertilizing crop substitution of β by the heritability cannot be justified. If, in this case, $\sum_{i=1}^{K} d_i \neq 0$, then $\mathrm{E}\underline{p}_{t+1}$ will deviate from $\mathrm{E}\underline{p}_t$, even in the absence of selection. This is due to the autonomous process of progressing inbreeding. According to Equation (11.13), however, absence of selection, *i.e.* $S = 0$, would imply $R = 0$, *i.e.* $\mathrm{E}\underline{p}_{t+1} = \mathrm{E}\underline{p}_t$. Prediction of R at $S \neq 0$ on the basis of the heritability is not possible in this situation.

If β is estimated to be b, then the response to selection with selection differential S is predicted to be

$$\hat{R} = bS \tag{11.22}$$

In practice, estimation of β involves estimation of either ${h_w}^2$ or ${h_n}^2$. This is possible

1. On the basis of estimates of the components of variance involved in the heritability. (examples are given in Section 11.2.1)
2. By means of estimation of the coefficient of regression of the phenotypic value of offspring on the phenotypic value of their parent(s) (Section 11.2.2)

It is emphasized that a high heritability does not necessarily imply a large genetic variance, nor that a large genetic variance necessarily implies a high heritability. At $h^2 = 1$ the ratio R/S amounts to 1, whereas at $h^2 = 0$ it is 0. The quantity h^2, a scale independent parameter, indicates thus the **efficiency of the selection**. The difference between S and R amounts to

$$S - R = S - h^2S = (1 - h^2)S \tag{11.23}$$

The part $(1 - h^2)$ of the selection differential does thus not give rise to a selection response. As ${h_w}^2 \geq {h_n}^2$ (this follows from the previous definitions of

${h_w}^2$ and ${h_n}^2$), the non-responding part of S will be smaller at identical reproduction of the selected candidates than at cross-fertilization of the selected candidates.

As

$$\mathrm{E}\underline{p}_s = \mathrm{E}(\underline{p}|\underline{p} > p_{min})$$

one may write

$$\mathrm{E}\underline{p}_s = \mathrm{E}(\underline{\mathcal{G}}|\underline{p} > p_{min}) + \mathrm{E}(\underline{e}|\underline{p} > p_{min}) = \mathrm{E}\underline{\mathcal{G}}_s + \mathrm{E}\underline{e}_s$$

Thus

$$S = \mathrm{E}\underline{p}_s - \mathrm{E}\underline{p} = \mathrm{E}\underline{\mathcal{G}}_s + \mathrm{E}\underline{e}_s - \mathrm{E}\underline{p} = (\mathrm{E}\underline{\mathcal{G}}_s - \mathrm{E}\underline{\mathcal{G}}) + (\mathrm{E}\underline{e}_s - \mathrm{E}\underline{e})$$

The quantity

$$\mathrm{E}\underline{\mathcal{G}}_s - \mathrm{E}\underline{\mathcal{G}}$$

represents the genetic superiority of the selected candidates. At identical reproduction it is equal to R, the response to selection, *i.e.* to ${h_w}^2 S$. The remainder, $\mathrm{E}\underline{e}_s - \mathrm{E}\underline{e} = \mathrm{E}\underline{e}_s$ (as $\mathrm{E}\underline{e} = 0$), is due to fortuitous favourable growing conditions of the selected candidates.

Then

$$\mathrm{E}\underline{e}_s = S - R = (1 - {h_w}^2)S = {e_w}^2 S$$

when defining

$${e_w}^2 = \frac{var(\underline{e})}{var(\underline{p})} = 1 - {h_w}^2 \tag{11.24}$$

This implies that selected candidates tend to have a positive environmental deviation. Their phenotypic superiority S is partly due to superior growing conditions, *i.e.* ${e_w}^2 S$, and partly due to genetic superiority, *i.e.* ${h_w}^2 S$.

The heritability value depends on the way the evaluation of the candidates is carried out. When each candidate genotype is represented by just a single plant the heritability of the candidates will be (considerably) smaller than when each candidate genotype is represented by a (large) number of plants (either or not evaluated on replicated plots). According to Equations (11.14) and (11.15), the response to **directional selection** depends on the heritability as well as on the selection differential. With regard to the former parameter, as applying to the situation where each candidate is represented by a single plant, the following rule of thumb guideline for selection in a cross-fertilizing crop may be given:

- At a single-plant value for ${h_n}^2$ amounting at least 0.40, mass selection will be successful
- At a single-plant value for ${h_n}^2$ in the interval $0.15 < {h_n}^2 < 0.40$, family selection may offer good prospects (depending on the extensiveness of the evaluation of the candidates)

- At a single-plant value for ${h_n}^2$ amounting less than 0.15, successful selection requires such great evaluation efforts that it is advised

  (a) to introduce new genetic variation
  (b) to stop dedicating efforts to the considered plant material
  (c) to assess the trait in a new way

It is admitted that these decision rules are only based on the heritability. The decision actually made by a breeder may also be based on additional considerations.

Phenotypic values and, consequently, genotypic values depend highly on the macro-environmental growing conditions. Thus not only the phenotypic and genotypic variance depend on the macro-environmental conditions (Example 8.8), but also the heritability (Example 11.4).

Example 11.4 When growing tomatoes outdoors, a quick and uniform emergence after sowing is desired. This may be pursued by selection. El Sayed and John (1973) studied, therefore, the heritability of speed of emergence under different temperature regimes. The following estimates were obtained:

Temperature regime	$\hat{h}^2$
Simulation of 10 years' average daily ambient maximum and minimum temperature	0.35
55° F constant temperature	0.55
daily 16^h 80° F and 8^h 63° F	0.64
50° F constant temperature	0.68

It is concluded that the temperature regime affects the heritability.

This leads to the following general question: At what macro-environmental conditions, *i.e.* the conditions prevailing during a certain growing season (year) at a certain site, is the efficiency of selection maximal? This topic is of course very important in the context of this book. It is also considered in Sections 12.3.3 and 15.2.1. Here three suggested answers are only briefly considered:

1. Macro-environmental conditions maximizing ${\sigma_g}^2$ or h^2
2. Macro-environmental conditions identical to those of the target environment, *i.e.* the conditions applied by a major group of growers
3. Macro-environmental conditions characterized by absence of interplant competition, *i.e.* use of a very low plant density

Macro-environmental conditions maximizing ${\sigma_g}^2$ or h^2

It can be said that a breeder should look for macro-environmental conditions such, that the heritability is high. This requires the macro-environment to be uniform, *i.e.* ${\sigma_e}^2$ is small, and the genetic contrasts to be large, *i.e.* ${\sigma_g}^2$ is large.

However, for different traits different sets of macro-environmental conditions may then be required (see Example 11.6). For example: selection for a high yield per plant may require a low plant density, but selection for a high yield per m^2 may require a high plant density.

For traits with a negligible genotype × environment interaction the selection may be done on the basis of testing in a single environment. Thus in order to select in oats for resistance against the crown rust disease, a number of oat genotypes may be inoculated in the laboratory with crown rust fungal spores. This maximizes the heritability of the degree of susceptibility (differences in the susceptibility do not show up in the absence of the disease). Then (on the assumption that laboratory tests are reflected in field performance) all resistant oat genotypes are expected to be resistant under commercial growing conditions. For traits with important g × e interaction, however, selection in the single macro-environment yielding maximum heritability may imply selection of genotypes that do not perform in a superior way in the target environment.

In Example 11.5 it is reported that differences among entries were larger under favourable growing conditions than under unfavourable conditions.

Example 11.5 In 1980 and 1981 Castleberry, Crum and Krull (1984) compared maize varieties bred in six different decades, *viz.*:

- ten open pollinating varieties bred 1930–40,
- three DC-hybrid varieties bred 1940–50,
- one DC- and two SC-hybrids bred 1950–60,
- three DC-, one TC- and one SC-hybrid bred 1960–70,
- two TC- and two SC-hybrids bred 1970–80 and
- two SC-hybrids bred 1980–90.

The comparison occurred at

- different locations
- high as well as at low soil fertility
- in the presence and in the absence of irrigation

For each decade-group the mean grain yield (in kg/ha) across the involved varieties was determined and plotted against the pertaining year (decade). The coefficient of regression was estimated to be $b = 82$ kg/ha. This figure represents the increase of the grain yield per year. Modern varieties yielded better than old varieties, both under intensive and extensive growing conditions (also reported in Example 13.10).

In the present context it is of special interest that the differences among the six groups of varieties were larger under favourable growing conditions, where the yield ranged from 6 to 12 t/ha, than under unfavourable conditions, where the yield ranged from 4.5 to 8.5 t/ha. The authors advised consequently to evaluate yield potentials under favourable growing conditions and to test for stress-tolerance in separate tests.

Macro-environmental conditions identical to those of the target environment

The suggestion to select under macro-environmental conditions identical to those of the target environment is generally accepted as a good guideline. However, with regard to plant density this suggestion implies a problem: due to the intergenotypic competition occurring when selecting under the high plant density applied at commercial cultivation, candidates may be selected that perform disappointingly when grown *per se, i.e.* in the absence of intergenotypic competition. Intergenotypic competition is a phenomenon which does not show up in the target environment provided by farmers growing genetically uniform varieties. With regard to competition it is, in fact, impossible to apply selection under conditions identical to those of the target environment. This topic is further considered in Section 12.3.3.

Fasoulas and Tsaftaris (1975) suggested that breeders should provide favourable growing conditions when selecting. The latter seems to be supported by the results of the experiment mentioned in Example 11.5, but the example also supports the idea that selection should be done under macro-environmental conditions similar to those of the target environment. Example 12.11 illustrates that selection aiming to increase grain yield under less-favourable conditions was the most effective when applied under the poor conditions of the target environment.

Macro-environmental conditions characterized by absence of interplant competition

The idea of avoiding interplant competition by applying a very low plant density is supported by the problem indicated in the former paragraph. Gotoh and Osanai (1959) and Fasoulas and Tsaftaris (1975) advocated application of selection at such a low plant density that interplant competition does not occur.

An objection against selecting at a very low plant density is its inefficiency if genotype × plant density interaction occurs. Thus some (*e.g.* Spitters, 1979, p. 117) have defended the opinion that selection should be applied at the plant density of commercial cultivation. This, however, would generate the problem of intergenotypic competition, a problem not occurring at a very low plant density (see the previous paragraph). Example 11.6 reports some experimental results.

Example 11.6 Vela-Cardenas and Frey (1972) established that a high plant density was optimal when selecting for reduced plant height of oats and that a low density was optimal when selecting for a high number of spikelets per panicle. When selecting for a larger kernel size all studied macro-environmental conditions were equally suited. Thus a general guideline cannot be derived from this study. The same applies to an empirical

study by Pasini and Bos (1990a,b) dedicated to the plant density to be preferred when selecting for a high grain yield in spring rye. They could not unambiguously substantiate a preference for either a high or a very low plant density. However, weak indications in favour of a low plant density were obtained.

The predicted response to selection as calculated from Equation (11.14) or (11.15) should only be considered as a rough indication. Example 11.7 shows that the discrepancy between the predicted response and the actual response may be considerable.

Example 11.7 In a population of winter rye consisting of 5263 plants, the 168 plants with the highest grain yield were selected (see Bos, 1981, Chapter 3). Because:

$$\overline{p} = 50\,\text{decigrams(dg)}$$

and

$$\overline{p}_{\text{s}} = 117.5\,\text{dg},$$

the selection differential, Equation (11.3), amounted to

$$S = \tfrac{1}{2}(67.5 + 0.0) = 33.75\,\text{dg}.$$

The narrow sense heritability was estimated to be 0.048 (see Example 11.10). The predicted response to the selection amounted thus to:

$$\hat{R} = 0.048 \times 33.75 = 1.6\,\text{dg}, \textit{i.e.}\ 3.2\%.$$

The average grain yield of the offspring of 84 random plants was 56.95 dg, whereas the average yield of the offspring of the 168 selected plants was 59.8 dg. The actual response to the selection was thus 2.85 dg, *i.e.* 5.0%.

Four reasons for such a discrepancy are mentioned here:

1. If linkage and/or epistasis occur, estimators for the heritability based on the assumption of their absence are biased.
2. The estimators of the heritability have some inaccuracy.
3. The macro-environmental conditions experienced by population P_t, the population subjected to selection, may differ from those experienced by population P'_{t+1}, the population obtained from the selected candidates. This relates both to imposed conditions, such as plant density, and uncontrollable conditions, such as climatic conditions. The actual response, appearing from a comparison of populations P'_{t+1} and P_t, is then to be regarded as a **correlated response** due to **indirect selection** P_t (Section 12.3). In this situation the result of deliberate selection is sometimes hardly better than the result of 'selection at random'.

4. Because the phenotypic values for different quantitatively varying traits tend to be correlated (Section 8.1), selection with regard to a certain trait implies indirect selection with regard to other, related traits. The correlated response to such indirect selection may turn out to be negative with regard to pursuing a certain ideotype.

   The indirect selection for biomass of maize, via selection for plant volumes (see Example 11.1), for instance, gave rise to a population susceptible to lodging. In the long-lasting selection programme of maize described in Example 8.4, selection for oil content implied indirect selection with regard to many other traits. A correlated response to selection was observed for: grain yield, earliness, plant height, tillering, *etc.*

Notwithstanding the often observed discrepancy between the predicted and the actual response to selection, the relation $R = \beta S$ is for plant breeders one of the most useful results of quantitative genetic theory. Based on this relationship the concept of **realized heritability**, designated as ${h_r}^2$, has been defined. It is calculated after having established the actual response to selection at some selection differential. When selecting among identical reproducing candidates, or when selecting before pollen distribution in a population of a cross-fertilizing crop the definition is

$${h_r}^2 = \frac{R}{S}$$

When selecting after pollen distribution in a population of a cross-fertilizing crop this definition turns out to be equivalent to

$${h_r}^2 = \frac{2R}{S_f}$$

Because R has already been established, the quantity ${h_r}^2$ can not be used to predict R. It indicates afterwards the efficiency of the applied selection procedure.

11.2 The Estimation of Quantitative Genetic Parameters

The main activity of a plant breeder does not consist of making quantitative genetic studies of a number of traits, but the development of new varieties. This means that breeders are unwilling to dedicate great efforts to the estimation of quantitative genetic parameters. Thus only estimation procedures demanding hardly any additional effort, fitting in a regular breeding programme, are presented in this section.

First attention is given to some problems involved in obtaining appropriate estimates of $\text{var}(\underline{e})$, the environmental variance. Because of these problems, in the present section procedures for estimating $\text{var}(\underline{\mathcal{G}})$ or h^2 not requiring estimation of $\text{var}(\underline{e})$ are emphasized.

Breeders may measure the phenotypic variation for a trait of some genetically heterogeneous population. They may do so by estimating $\mathrm{var}(\underline{p})$. However, their main interest lies in exploiting the genetic variation. As

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{p}) - \mathrm{var}(\underline{e}) \tag{11.25}$$

an appropriate way to estimate $\mathrm{var}(\underline{\mathcal{G}})$ consists of subtracting $\hat{\mathrm{var}}(\underline{e})$ from $\hat{\mathrm{var}}(\underline{p})$.

The estimate for $\mathrm{var}(\underline{e})$ should be derived from similar but genetically homogeneous plant material, grown in the same macro-environmental conditions as the population of interest. A complication arises if the genotypes differ in their capacity to buffer variation in the growing conditions. Then the candidates representing one genotype are more (or less) affected by the prevailing variation in the quality of the micro-environmental growing conditions than the candidates plants representing another genotype. This was already dealt with in Example 8.9 and its preceding text.

To account for this, the environmental variance assigned to the F_2 population of a self-fertilizing crop is sometimes estimated to be:

$$\tfrac{1}{4}\hat{\mathrm{var}}(\underline{p}_{\mathrm{P1}}) + \tfrac{1}{2}\hat{\mathrm{var}}(\underline{p}_{\mathrm{F1}}) + \tfrac{1}{4}\hat{\mathrm{var}}(\underline{p}_{\mathrm{P2}}) \tag{11.26}$$

Plants of the F_2 generation are more heterozygous than those of P_1 or P_2, but less than those of the F_1. Heterogeneity among plants of the F_1 may be partly due to the manipulations applied to produce the F_1 seed, *i.e.* emasculation and pollination of the parent (instead of spontaneous selfing). Manipulation certainly contributes to heterogeneity in the case of cloning. Thus the usual way of cloning (*e.g.* of grass or rye plants) gives clones such that the within-clone phenotypic variance overestimates the environmental variance appropriate to the segregating plant material not subjected to the manipulation required for the cloning. Example 11.8 illustrates the present concern of using a non-representative estimate of $\mathrm{var}(\underline{e})$.

Example 11.8 A straightforward estimate of $\mathrm{var}(\underline{e})$ for the maize material described in Example 8.9 is

$$\hat{\mathrm{var}}(\underline{e}) = \tfrac{1}{6}(185 + 256 + 90.3 + 285.6 + 424.4 + 240.3) = 246.9\,(\mathrm{cm})^2$$

This yields for the DC-hybrid WXYZ:

$$\hat{\mathrm{var}}(\underline{\mathcal{G}}) = 475.3 - 246.9 = 228.4\,(\mathrm{cm})^2$$

and

$$\hat{h}_w{}^2 = \tfrac{228.4}{475.3} = 0.48$$

This approach is risky because of the positive relationship between $\overline{p}$ and $\hat{\text{var}}(\underline{p})$. Thus a higher estimate for the environmental variance of the DC-hybrid than 246.9 cm^2 is likely to be more appropriate. That would imply a lower value for ${h_w}^2$.

11.2.1 Plant Material with Identical Reproduction

Clones, pure lines and single-cross hybrids can be reproduced with the same genotype. For such plant material, estimation of the heritability in the wide sense may proceed as elaborated in this section.

A random sample consisting of I genotypes is taken from a population of entries with identical reproduction; $I > 1$. Each sampled genotype is evaluated by growing it in J plots, each containing K plants; $J > 1, K \geq 1$. These plots may be assigned to

1. A completely randomized experiment
2. Randomized (complete) blocks.

Table 11.1 presents the analysis of variance for either design.

The test of the null hypothesis H_0: "${\sigma_g}^2 = 0$" requires calculation of the F value, MS_g/MS_r. This value is compared with critical values tabulated for different levels of significance.

Unbiased estimates of σ^2 and ${\sigma_g}^2$ are

$$\hat{\sigma}^2 = MS_r \tag{11.27}$$

$$\hat{\sigma}_g^2 = \frac{MS_g - MS_r}{J} \tag{11.28}$$

Table 11.1 The structure of the analysis of variance of data obtained from I genotypes evaluated at J plots

(a) Completely randomized experiment

Source of variation	df	SS	MS	E(MS)
Genotypes	$I-1$	SS_g	MS_g	$\sigma^2 + J{\sigma_g}^2$
Residual	$I(J-1)$	SS_r	MS_r	σ^2

(b) Randomized complete block design

Source of variation	df	SS	MS	E(MS)
Blocks	$J-1$	SS_b	MS_b	$\sigma^2 + I{\sigma_b}^2$
Genotypes	$I-1$	SS_g	MS_g	$\sigma^2 + J{\sigma_g}^2$
Residual	$(J-1)(I-1)$	SS_r	MS_r	σ^2

For each entry the mean phenotypic value calculated across the J plots constitutes the basis for the decision to select it or not. Thus the appropriate environmental variance when testing each genotype at each of J plots is

$$\sigma_e^2 = \frac{\sigma^2}{J}$$

The wide sense heritability is thus

$$h_w{}^2 = \frac{\sigma_g{}^2}{\sigma_g{}^2 + \sigma_e{}^2} = \frac{\sigma_g{}^2}{\sigma_g{}^2 + \frac{\sigma^2}{J}} \tag{11.29}$$

It should be noted that substitution of the unbiased estimates for $\sigma_e{}^2$ and for $\sigma_g{}^2$ in Equation (11.29) does not yield an unbiased estimate for $h_w{}^2$. Example 11.8 illustrates the estimation of a few statistical parameters with an interesting quantitative genetic interpretation.

Example 11.8 A random sample of $I = 3$ genotypes were evaluated in each of $J = 4$ blocks. The observations were

		Block				
		1	2	3	4	Total
Genotype	1	6	8	7	6	27
	2	6	6	5	5	22
	3	7	9	8	7	31
	Total	19	23	20	18	80

An analysis of variance of these data as if resulting from a completely randomized experiment (Table 11.1(a)), yields

Source of variation	df	SS	MS	E(MS)
Genotypes	2	10.17	5.09	$\sigma^2 + 4\sigma_g{}^2$
Residual	9	6.50	0.722	σ^2

The F value, *i.e.* $5.09/0.722 = 7.05$, indicates that the null hypothesis H_0: $\sigma_g{}^2 = 0$ is rejected ($P < 0.025$). The estimates of the variance components are

$$\hat{\sigma}^2 = 0.722,$$

and

$$\hat{\sigma_g}{}^2 = 1.09.$$

According to these estimates the (biased!) estimate of $h_w{}^2$ amounts to 0.86.

Analysis of variance of these data according to a randomized complete block design yields

Source of variation	df	SS	MS	E(MS)
Blocks	3	4.67	1.56	$\sigma^2 + 3\sigma_b{}^2$
Genotypes	2	10.17	5.09	$\sigma^2 + 4\sigma_g{}^2$
Residual	6	1.83	0.305	σ^2

The F value, *i.e.* 16.7, indicates that the null hypothesis H_0: ${\sigma_g}^2 = 0$ is rejected ($P < 0.005$). The F value for the blocks, *i.e.* 5.1, indicates that the null hypothesis H_0: ${\sigma_b}^2 = 0$ is rejected ($P < 0.05$). The estimates of the variance components are

$$\hat{\sigma}^2 = 0.305,$$

and

$$\hat{\sigma}_g^2 = 1.196.$$

According to these estimates the biased estimate of ${h_w}^2$ amounts to 0.94. Partitioning of the trial field in blocks yielded a somewhat higher heritability, implying a somewhat higher efficiency of selection.

According to the F value for genotypes and its significance level, the power of the randomized block design was higher than that of the completely randomized experiment.

The intention of replicated testing of entries in several plots is a reduction of the environmental variance. This induces the heritability to be higher at higher values for J. The ratio

$$\frac{{h_J}^2}{{h_1}^2},$$

i.e. the heritability when testing each entry in several plots to the heritability when testing each entry at a single plot, is now considered.

In doing so, in the remainder of this section symbols with the subscript 1 refer to non-replicated testing ($J = 1$), and symbols with the subscript J to replicated testing ($J \geq 2$). The heritability appropriate when testing each entry at each of J plots is thus designated by

$${h_J}^2 = \frac{{\sigma_g}^2}{{\sigma_J}^2} \tag{11.30}$$

where ${\sigma_J}^2$ represents the phenotypic variance of the means of the entries across J plots, *i.e.*

$${\sigma_J}^2 = {\sigma_g}^2 + \left(\frac{\sigma^2}{J}\right) \tag{11.31}$$

Then

$${h_1}^2 = \frac{{\sigma_g}^2}{{\sigma_g}^2 + \sigma^2} = \frac{{\sigma_g}^2}{{\sigma_1}^2} \tag{11.32}$$

which implies

$${\sigma_g}^2 = {h_1}^2 {\sigma_1}^2,$$

and

$$\sigma^2 = {\sigma_1}^2 - {\sigma_g}^2 = {\sigma_1}^2 - {h_1}^2 {\sigma_1}^2.$$

Thus

$${\sigma_J}^2 = {h_1}^2 {\sigma_1}^2 + \left(\frac{{\sigma_1}^2 - {h_1}^2 {\sigma_1}^2}{J}\right)$$

Table 11.2 The ratio of the heritability when testing each entry at J plots to the heritability when testing each entry at a single plot ($h_1{}^2$), for several values for $h_1{}^2$ and J

	$h_1{}^2$				
J	0.1	0.2	0.3	0.4	0.5
2	1.82	1.67	1.54	1.43	1.33
3	2.50	2.14	1.88	1.67	1.50
4	3.08	2.50	2.11	1.82	1.60

or

$$\frac{\sigma_J{}^2}{\sigma_1{}^2} = h_1{}^2 + \left(\frac{1 - h_1{}^2}{J}\right) = \frac{1 + h_1{}^2(J-1)}{J} \tag{11.33}$$

From Equations (11.30) and (11.32) it follows that

$$\frac{h_J{}^2}{h_1{}^2} = \frac{\sigma_1{}^2}{\sigma_J{}^2} = \frac{J}{1 + h_1{}^2(J-1)} \tag{11.34}$$

Table 11.2 presents the ratio $\frac{h_J{}^2}{h_1{}^2}$ for several values for $h_1{}^2$ and J.

Especially for a (very) low value for $h_1{}^2$ application of additional replications may be rewarding because of the large (relative) increase of the heritability. The largest relative improvement occurs when applying $J = 2$ instead of $J = 1$. Thus potato breeders should consider a system where each first-year-clone is represented by 2 seed potatoes instead of only 1, which is customary; see Pfeffer *et al.* (1982).

As a general conclusion it is stated that replicated testing promotes the efficiency of selection. If the replicated testing involves different macro-environments it gives an indication of the stability as well.

In Section 16.1 attention is given to the optimum number of replications, say J_{opt}. It is the number of replications giving rise to the maximum response to selection at a fixed number of plots. The ratio $h_J{}^2/h_1{}^2$ is shown to play a crucial role in the derivation of J_{opt}.

In connection with the foregoing, we consider the ratio

$$\frac{\sigma_b{}^2}{\sigma_b{}^2 + \sigma_w{}^2} \tag{11.35}$$

where

$\sigma_{\mathrm{b}}{}^2$ represents the between-entry component of variance and

$\sigma_{\mathrm{w}}{}^2$ the within-entry component of variance.

The ratio may be considered if from each entry $J > 1$ observations are available. This occurs in perennial crops, such as apple and oil palm, when

observing in successive years the yield per year of individual plants. The quantitative genetic interpretations of these components of variance are

$\sigma_w{}^2$: environmental variance in course of time and
$\sigma_b{}^2$: genetic variance + variance due to variation in permanent environmental conditions (because of the permanent position in the field).

In statistics the ratio is called **intraclass correlation coefficient** or **repeatability** (Snedecor and Cochran, 1980, p. 243). The numerator of the ratio tends to be larger than $\sigma_g{}^2$, which causes the ratio to be larger than $h_w{}^2$.

In certain situations estimation of h^2 is not as easy as estimation of the repeatability. Then one may simply estimate the repeatability as this quantity indicates the upper limit of $h_w{}^2$.

Observations repeated in the course of time do not only allow estimation of the repeatability or the heritability, they also indicate the stability, for instance the presence or absence of certain genotype × year interaction effects.

11.2.2 Cross-fertilizing Crops

In the introduction to Section 11.2 it was indicated that procedures for estimating var($\underline{\mathcal{G}}$) or h^2 not requiring separate estimation of var($\underline{e}$) will be considered. In Section 10.2 it was concluded that estimation of the additive genetic variance ($\sigma_a{}^2$) on the basis of regression, *i.e.* according to Equation (10.12), is to be preferred over estimation on the basis of an analysis of variance, *i.e.* according to Equation (10.11). However, for the sake of completeness first the estimation of $\sigma_a{}^2$ and h^2 on the basis of an analysis of variance is briefly considered.

Estimation on the basis of an analysis of variance

Estimation of $\sigma_a{}^2$ on the basis of an analysis of variance, *i.e.* according to Equation (10.8), is now considered. The number of HS-families in the random sample taken from the whole set of HS-families is designated by the symbol I. These I families are evaluated by means of a randomized complete block design involving J blocks, each consisting of I plots of K plants; $I > 1, J > 1, K \geq 1$. Table 11.3 presents the structure of the analysis of variance.

Variance component σ_f^2, *i.e.* var($\underline{\mathcal{G}}_{\mathrm{HS}}$), is estimated as

$$\hat{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) = \frac{MS_f - MS_r}{J} \tag{11.36}$$

Table 11.3 The analysis of variance of data obtained from I HS-families each evaluated at J plots, distributed across J blocks

Source of variation	df	SS	MS	E(MS)
Blocks	$J-1$	SS_b	MS_b	$\sigma^2 + I{\sigma_b}^2$
HS-families	$I-1$	SS_f	MS_f	$\sigma^2 + J{\sigma_f}^2$
Residual	$(J-1)(I-1)$	SS_r	MS_r	σ^2

and next ${\sigma_a}^2$, according to Equation (10.11), as

$$\hat{\sigma}_a^2 = 4\hat{\text{var}}(\underline{\mathcal{G}}_{\text{HS}}) \tag{11.37}$$

When selecting among the families on the basis of their mean phenotypic value calculated across the J plots, the heritability may be estimated according to Equation (11.29). Example 11.9 gives an illustration.

Example 11.9 $I = 3$ HS-families were evaluated in each of $J = 2$ blocks. The observations were

		Block		
		1	2	Total
Family	1	15.8	16.4	32.2
	2	18.2	17.4	35.6
	3	17.4	16.6	34.0
	Total	51.4	50.4	101.8

Analysis of variance of these data according to a randomized complete block design yields

Source of variation	df	SS	MS	E(MS)
Blocks	1	0.167	0.167	$\sigma^2 + 3{\sigma_b}^2$
Families	2	2.893	1.447	$\sigma^2 + 2{\sigma_f}^2$
Residual	2	0.654	0.327	σ^2

According to the estimates $\hat{\sigma}^2 = 0.327$ and $\hat{\sigma}_f^2 = 0.560$, the biased estimate of h^2 – as applying to way in which the HS-families were evaluated – amounts to 0.77. The additive genetic variance is estimated to be $4 \times 0.560 = 2.24$.

Estimation on the basis of regression analysis

In the present section, emphasis is on estimation of ${\sigma_a}^2$ and ${h_n}^2$ on the basis of regression of the phenotypic value of offspring on the phenotypic value of parents.

The statistical meaning of the regression coefficient β is that it indicates how the performance of offspring are expected to change with a one-unit change in the performance of parents. In this respect the response to selection is directly

at issue. Note 11.4 gives attention to the problem of the shape of the function to be fitted when considering the relationship between offspring and parents.

Note 11.4 The graph relating the genotypic value of the offspring and the phenotypic value of the parents may be expected to be a sigmoid curve instead of a straight line. This is explained as follows.

Indeed, across the whole population $\mathrm{E}\underline{e} = 0$ due to $\mathrm{E}\underline{p} = \mathrm{E}\underline{\mathcal{G}}$. However, in Section 11.1, it was shown that

$$\mathrm{E}\underline{e}_{\mathrm{s}} = \mathrm{E}(\underline{e}|\underline{p} > \mathrm{p}_{min}) = {e_w}^2 S > 0.$$

When selecting candidates with a low phenotypic value one may, likewise, derive

$$\mathrm{E}\underline{e}_{\mathrm{s}} = \mathrm{E}(\underline{e}|\underline{p} < \mathrm{p}_{max}) = {e_w}^2 S < 0.$$

Thus the regression coefficient estimated on the basis of a random sample of parental candidates and their offspring may overestimate the performance of the offspring of selected candidates having a phenotypic values located in the tail of the distribution.

1. Regression of HS-family performance on maternal plant performance.
In the case of open pollination, the paternal plants cannot be identified. Then only the coefficient of regression of HS-family performance on maternal plant performance can be estimated. According to Equation (10.10) ${\sigma_a}^2$ and ${h_n}^2$ may then be estimated on the basis of the following expressions:

$${\sigma_a}^2 = 2\mathrm{cov}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) \tag{11.38}$$

$${h_n}^2 = \frac{{\sigma_a}^2}{{\sigma_p}^2} = \frac{2cov(\underline{p}_M, \underline{p}_{HS})}{var(\underline{p}_M)} = 2\beta_{HS,M} \tag{11.39}$$

Example 11.10 gives an illustration.

Example 11.10 In the growing season 1975–76 a population of winter rye plants comprising 5263 plants was grown (Bos, 1981). The mean phenotypic value for grain yield was $\overline{p} = 50\,\mathrm{dg}$. After harvest a random sample of 84 plants was taken under the condition that each random plant produced enough seeds to grow the required number of offspring. The average grain yield of these 84 plants amounted to 56.95 dg.

In 1976–77 the offspring of each random plant was grown as a single-row plot of 20 plants, in each of two blocks. The coefficient of regression of offspring on maternal parent was estimated to be $b = 0.024$. The heritability in the narrow sense of grain yield of individual plants was thus estimated to be 0.048. The estimated coefficient of correlation amounted only to $r = 0.04$. It did not differ significantly from 0.

N.B. Absence of selection was one the conditions, considered in Section 10.2.1, to justify interpretation of estimates of statistical parameters

in terms of quantitative genetical parameters. The reason for this is that the relationship between offspring and selected parents may differ from that between offspring and parents in the absence of selection. It may thus, even when the relationship would have been significant, be questioned whether the obtained estimate for ${h_n}^2$ yields an unbiased prediction of the response to selection.

2. Regression of FS-family performance on parental performance.
In the case of pairwise crosses one may estimate the coefficient of regression of FS-family performance on the mean performance across both parents. According to Equation (10.16) ${\sigma_a}^2$ and ${h_n}^2$ can then be estimated on the basis of the following expressions:

$$ {\sigma_a}^2 = 2\mathrm{cov}(\underline{p}_{\mathrm{P}}, \underline{p}_{\mathrm{FS}}) \tag{11.40} $$

$$ {h_n}^2 = \frac{{\sigma_a}^2}{{\sigma_p}^2} = \frac{2cov(\underline{p}_P, \underline{p}_{FS})}{2var(\underline{p}_P)} = \beta_{FS,P} \tag{11.41} $$

A discussion in Section 10.2.1 suggests that estimates of ${\sigma_a}^2$ according to Equation (11.37) will tend to be higher than estimates according to Equation (11.38) or (11.40). Example 11.11 presents results of a comparison of the two ways of estimating ${\sigma_a}^2$.

Example 11.11 Bos (1981, p. 138) estimated ${\sigma_a}^2$ both on the basis of regression, *i.e.* Equation (11.38), and on the basis of an analysis of variance, *i.e.* Equation (11.37). The estimates were calculated from data from random samples of plants taken from a population of winter rye subjected to continued selection aiming at higher grain yield and reduced plant height. The estimates concerned grain yield (in dg) and plant height (in cm). The following estimates were obtained:

Growing season of the parental plants	Grain yield		Plant height	
	Regression	Anova	Regression	Anova
1974–75	215.5	268.0	63.3	87.6
1975–76	24.9	193.2	41.7	71.6
1976–77	476.6	0.0	99.6	131.9
1977–78	95.7	54.2	64.0	56.6

For five of the eight pairs of estimates the 'anova-estimate' appeared to be higher than the corresponding 'regression-estimate'.

With open pollination each plant will predominantly be pollinated by a few of its neighbours. If each plant was pollinated by only one neighbour, $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}})$ would in fact be equal to $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}})$. Equations (10.8), *i.e.* $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{HS}}) = \frac{1}{4}{\sigma_a}^2$, and (10.14), *i.e.* $\mathrm{var}(\underline{\mathcal{G}}_{\mathrm{FS}}) = \frac{1}{2}{\sigma_a}^2 + \frac{1}{4}{\sigma_d}^2$, show that pollination by a few neighbours tends to cause an upward bias when estimating ${\sigma_a}^2$ by $4\hat{\mathrm{var}}(\underline{\mathcal{G}}_{\mathrm{HS}})$.

Polycrosses aim to produce real panmixis. This is promoted by planting the plants representing the involved clones at positions according to the patterns proposed by Oleson and Oleson (1973) and Oleson (1976). In these patterns each clone has each other clone equally often as a neighbour; if desired, even equally often as a neighbour in each of the four directions of the wind. Morgan (1988) presents schemes for N clones, each represented by N^2 plants. These schemes consist of N squares of $N \times N$ plants. Each clone has each other clone N times as a direct neighbour in each of the four directions of the wind, and $N-2$ times as a direct neighbour in each of the four intermediate directions. Each clone is $N-1$ times its own direct neighbour in each of the four intermediate directions.

Comstock and Robinson (1948, 1952) proposed mating designs yielding progenies in such a way that the estimates for ${\sigma_a}^2$ or ${\sigma_d}^2$ are unbiased. These mating designs are known as North Carolina mating design I, II and III. They require effort, especially the making of additional crosses, not coinciding with normal breeding procedures. For this reason these designs are not considered further here.

The degree of linear association of two random variables, $\underline{x}$ and $\underline{y}$, is measured by the coefficient of correlation, say $\rho_{x,y}$. The linear relation itself is described by the function

$$\hat{y} = \alpha + \beta x, \tag{11.42}$$

where

β is the coefficient of regression of y on x and

$\hat{y}$ is the value predicted for $\underline{y}$ if $\underline{x}$ assumes the value x.

In the preceding text the regression of offspring performance ($\underline{y}$) on parental plant performance ($\underline{x}$) was considered. The parental plants and their offspring are usually evaluated in different growing seasons, *i.e.* under different macro-environmental conditions. Thus E$\underline{x}$ may differ from E$\underline{y}$ and var($\underline{x}$) may differ from var($\underline{y}$). For this reason one may consider standardization of the observations obtained from parents and offspring prior to the calculation of the regression coefficients α and β. In Note 11.5 it is shown that the coefficient of regression of standardized values for $\underline{y}$, *i.e.* $\underline{z}_y$, on standardized values for $\underline{x}$, *i.e.* $\underline{z}_x$, is equal to the coefficient of correlation of $\underline{x}$ and $\underline{y}$. Thus calculation of the coefficient of regression of $\underline{z}_y$ on $\underline{z}_x$ yields the same figure as calculation of the coefficient of correlation of $\underline{x}$ and $\underline{y}$. For this reason Frey and Horner (1957) introduced for ρ the term **heritability in standard units**.

N.B. Frey and Horner (1957) calculated the coefficient of regression of offspring on parent for oats, a self-fertilizing crop. However, for self-fertilizing crops a simple quantitative genetic interpretation of β in terms of 'the' heritability is not possible (see Section 11.1). Nevertheless Smith and Kinman (1965) presented a relationship allowing the derivation of the

Note 11.5 Standardization of the variable $\underline{x}$ yields the variable $\underline{z}_x$:

$$\underline{z}_x = \frac{\underline{x} - \mu_x}{\sigma_x}$$

Likewise one may determine

$$\underline{z}_y = \frac{\underline{y} - \mu_y}{\sigma}$$

We now calculate β', *i.e.* the coefficient of regression of $\underline{z}_y$ on $\underline{z}_x$. Equation (11.42) implies that

$$var(\underline{\hat{y}}) = var(\alpha + \beta\underline{x}) = \beta^2 var(\underline{x}) = \frac{cov^2(\underline{x}, \underline{y})}{var(\underline{x}) \times var(\underline{y})} \times var(\underline{y}) = \rho^2 var(\underline{y}) \tag{11.43}$$

When regressing $\underline{z}_y$ on $\underline{z}_x$, Equation (11.43) implies

$$(\beta')^2 \mathrm{var}(\underline{z}_x) = \rho^2(\underline{z}_x, \underline{z}_y)\mathrm{var}(\underline{z}_y)$$

Since

$$\mathrm{var}(\underline{z}_x) = \mathrm{var}(\underline{z}_y) = 1$$

and

$$\rho(\underline{z}_x, \underline{z}_y) = \rho_{x,y}$$

Equation (11.43) can be simplified to

$$\beta' = \rho_{x,y} \tag{11.44}$$

heritability from β. It is questionable whether that relationship is correct. In this book it is taken for granted that the bias due to inbreeding depression does not justify prediction of the response to selection in segregating generations of a self-fertilizing crop.

11.2.3 Self-fertilizing Crops

First attention will be given to the estimation of m, the origin in the F_∞-metric. It is the contribution to the genotypic value due to the common genotype for all non-segregating loci. It is equal to the unweighted mean genotypic value across the 2^K complex homozygous genotypes with regard to the K segregating loci (Section 8.3.2).

If epistasis does not occur, one may estimate m in a very direct way. This can be justified for any value for K, but here the justification is elaborated

for only two loci B_1-b_1 and B_2-b_2 (which may be linked). According to its definition we have

$$m = \tfrac{1}{4}(\mathcal{G}_{b1b1b2b2} + \mathcal{G}_{B1B1b2b2} + \mathcal{G}_{b1b1B2B2} + \mathcal{G}_{B1B1B2B2})$$

Absence of epistasis means

$$\mathcal{G}_{B1\text{-}b1,B2\text{-}b2} = m + \mathcal{G}'_{B1\text{-}b1} + \mathcal{G}'_{B2\text{-}b2}$$

(Equations (1.1) and (8.3)). This implies

$$\begin{aligned} m &= \tfrac{1}{4}(m + \mathcal{G}'_{b1b1} + \mathcal{G}'_{b2b2} + m + \mathcal{G}'_{B1B1} + \mathcal{G}'_{b2b2} + m + \mathcal{G}'_{b1b1} + \mathcal{G}'_{B2B2} \\ &\quad + m + \mathcal{G}'_{B1B1} + \mathcal{G}'_{B2B2}) \\ &= \tfrac{1}{2}(2m + \mathcal{G}'_{b1b1} + \mathcal{G}'_{b2b2} + \mathcal{G}'_{B1B1} + \mathcal{G}'_{B2B2}) \\ &= \tfrac{1}{2}(\mathcal{G}_{b1b1b2b2} + \mathcal{G}_{B1B1B2B2}) = \tfrac{1}{2}(\mathcal{G}_{b1b1B2B2} + \mathcal{G}_{B1B1b2b2}) \\ &= \tfrac{1}{2}(\mathcal{G}_{\mathrm{P}_1} + \mathcal{G}_{\mathrm{P}_2}) \end{aligned}$$

if P_1 and P_2 are the homozygous genotypes which were crossed to give rise to the considered segregating plant material. Example 11.12 illustrates this.

Example 11.12 If the genotype of P_1 is $b_1b_1B_2B_2b_3b_3$ and that of P_2 $B_1B_1b_2b_2B_3B_3$, then the genotypic values of P_1 and P_2 are, in the absence of epistasis, partitioned as

$$\mathcal{G}_{\mathrm{P}_1} = m - a_1 + a_2 - a_3$$

and

$$\mathcal{G}_{\mathrm{P}_2} = m + a_1 - a_2 + a_3$$

yielding

$$\tfrac{1}{2}(\mathcal{G}_{\mathrm{P}_1} + \mathcal{G}_{\mathrm{P}_2}) = m$$

whatever the degree of linkage of these three loci.

Generally absence of epistasis implies

$$m = \tfrac{1}{2}\left(\mathcal{G}_{P_1} + \mathcal{G}_{P_2}\right) \tag{11.45}$$

This allows estimation of m by

$$\hat{m} = \tfrac{1}{2}\left(\overline{p}_{P_1} + \overline{p}_{P_2}\right) \tag{11.46}$$

whatever the strength of linkage of the involved loci. An interesting application of the present result is illustrated in Section 11.4.2.

In Section 10.3 interest in $\sum_i {a_i}^2$ was explained. It was shown that from F_3 plant material an unbiased estimate of $\sum_i {a_t}^2$ can be derived based on Equation (10.26), *i.e.*

$$2var(\underline{\mathcal{G}}_{\mathrm{LF3}}) - var(\underline{\mathcal{G}}_{(\mathrm{LF3})}) = \tfrac{3}{4}\sum_i {a_i}^2$$

This would require estimation of $\text{var}(\underline{\mathcal{G}}_{\text{LF3}})$ and of $\text{var}(\underline{\mathcal{G}}_{(\text{LF3})})$. It is rather demanding to get accurate and unbiased estimates of these variance components. A possible approach could be estimation of each of these genetic variance components by subtracting from the corresponding estimates of phenotypic variance an appropriate estimate of the environmental variance. For plant breeders this approach is unattractive because it requires too large an effort. The present section presents a procedure for estimating $\sum_i {a_i}^2$ from F_3 plant material that

- fits into a regular breeding programme,
- avoids separate estimation of components of environmental variance and
- yields an accurate estimate.

This is all attained by estimating $\text{var}(\underline{\mathcal{G}}_{\text{LF3}})$ for a random sample of F_3 lines and estimating $\sum_i {a_i}^2$ by $2\hat{\text{var}}(\underline{\mathcal{G}}_{\text{LF3}})$.

Variance component $\text{var}(\underline{\mathcal{G}}_{\text{LF3}})$ can be estimated on the basis of a very simple experimental design. This proceeds as follows. Each of I F_3 lines, which are obtained in the absence of selection from I F_2 plants, is evaluated at J plots, each comprising K plants; $I > 1, J > 1, K \geq 1$. The J plots per F_3 line are distributed across J complete blocks. The structure of the appropriate analysis of variance is presented in Table 11.4.

An unbiased estimate for ${\sigma_l}^2$ is

$$\hat{var}(\underline{\mathcal{G}}_{\text{LF3}}) = \frac{MS_l - MS_r}{J}$$

According to Equation (10.24) the quantitative genetic interpretation of ${\sigma_l}^2$ is

$$var(\underline{\mathcal{G}}_{\text{LF3}}) = \tfrac{1}{2}\sum_i {a_i}^2 + \tfrac{1}{16}\sum_i {d_i}^2$$

Thus estimation of $\sum_i {a_i}^2$ by

$$\sum_i \hat{a}_i^2 = 2\hat{var}(\underline{\mathcal{G}}_{\text{LF3}}) \tag{11.47}$$

Table 11.4 The analysis of variance of data obtained from I F_3 lines evaluated at J plots, distributed across J blocks

Source of variation	df	SS	MS	E(MS)
Blocks	$J-1$	SS_b	MS_b	$\sigma^2 + I{\sigma_\text{b}}^2$
F_3 lines	$I-1$	SS_l	MS_l	$\sigma^2 + J{\sigma_l}^2$
Residual	$(J-1)(I-1)$	SS_r	MS_r	σ^2

implies the use of a biased estimator. However, in many cases – depending on the heritability in F_∞, the experimental design and the size of $\sum_i {d_i}^2$ – this estimator is much more accurate than an unbiased estimator (Van Ooijen, 1989). Then the probability of correct ranking of F_3, F_4, etc. populations with regard to $\sum_i a_i^2$ is larger.

This estimation procedure requires replicated testing ($J \geq 2$). Replicated testing can be attractive because non-replicated testing implies confounding of line effects and plot effects, including effects of intergenotypic competition (see Note 11.6). Replicated testing claims, however, a part of the testing capacity and requires for some crops that the plants of the F_2 population are grown at a low plant density in order to guarantee that these produce a sufficient amount of seed for replicated testing of the F_3 lines. The response to selection when evaluating F_3 lines at $J \geq 2$ plots instead of only a single plot is considered in Chapter 16.

Note 11.6 Intergenotypic competition tends to enlarge $\text{var}(\underline{\mathcal{G}})$, Example 8.8. Intergenotypic competition between F_3 lines may thus be responsible for a part of $\text{var}(\underline{\mathcal{G}}_{\text{LF3}})$. However, the F_∞ lines to be developed are to be used in large fields were intergenotypic competition does not cause inflation of the genetic variance. The variance of the genotypic values of the pure lines, *i.e.* $\sum_i a_i^2$, is therefore overestimated by $\hat{\text{var}}(\underline{\mathcal{G}}_{\text{LF3}})$ if intergenotypic competition occurs.

11.3 Population Genetic and Quantitative Genetic Effects of Selection Based on Progeny Testing

Section 8.3.3 introduced the concept of breeding value as a rather abstract quantity applying in the case of random mating (see Equation (8.12)). In Section 8.3.4 it was emphasized that the concept is of great importance when selecting among candidates on the basis progeny testing. The present section aims to clarify population genetic and quantitative genetic effects of such selection.

The progenies to be evaluated are obtained by crossing of candidates with a so-called tester population. In Section 3.2.2 it was shown that, in the case of selfing, haplotype frequencies hardly change in course of the generations. Thus it does not matter so much whether one evaluates the breeding value of individual plants or the breeding value of lines derived from these plants. The obtained progenies are HS-families.

The tester population may be

1. The population to which the candidates belong (intrapopulation testing)
2. Another population (interpopulation testing)

Intrapopulation testing

In the case of intrapopulation testing the allele frequencies of the tester population are equal to the allele frequencies of the population of candidates: p and q. Open pollination, as in the case of a polycross, is of course the simplest way of obtaining the progenies.

Interpopulation testing

When applying interpopulation testing, the tester population is another population than the population of candidates. Its allele frequencies are designated p' and q'. The aggregate of all families resulting from the test-crosses is then equal to the population resulting from bulk crossing (Section 2.2.1). Interpopulation testing occurs at **top-crossing** and at **reciprocal recurrent selection** (Section 11.3). In top-crossing a set of (pure) lines, which have been emasculated, are pollinated by haplotypically diverse pollen, possibly produced by an SC-hybrid or by a genetically heterogeneous population. At so-called **early testing**, young lines are involved in the top-cross (Section 11.5.2).

With regard to the candidates being tested, we now consider

1. The effect of the allele frequencies in the tester population on the ranking of the candidates with regard to their breeding value
2. The effect of selection of candidates with a high breeding value on the allele frequencies and, as a consequence, the expected genotypic value

The effect of the allele frequencies in the tester population on the ranking of the candidate genotypes with regard to their breeding value

When selecting (parental) plants with regard to their breeding values, plants with the most attractive (possibly: the highest) breeding values are selected. However, the ranking of the breeding values of plants with genotype *bb, Bb* or *BB* is not straightforward. It depends on the frequency of allele B in the tester population. This complicating factor is now considered.

The selection among the candidates is based on the quality of their offspring, *i.e.* on their breeding value. Table 8.6 shows that, for a given allele frequency (p), the ranking of the candidates with regard to their breeding value depends on whether α' (Equation (8.26a)) is positive, zero or negative. The ranking depends thus on whether

$$a' = a - (p' - q')d = a - (2p' - 1)d = (a + d) - 2p'd \tag{11.48}$$

is positive, zero or negative. This depends for a given locus, *i.e.* for given values for a and d, on p', the gene frequency in the tester population. The values for p' making α' either positive, or zero or negative will now be derived. Because of the tendency that $d \geq 0$ for most of the loci (Section 9.4.1), these values

will only be derived for loci with $d \geq 0$. When considering Equation (11.48) it is easily derived that

- $\alpha' > 0$: for loci with $0 \leq d \leq a$, if $0 \leq p' < 1$; and for loci with $d > a$ if $p' < p_m$, where $p_m = \frac{a+d}{2d}$ (Equation (9.9))
- $\alpha' = 0$: for loci with $d = a$ if $p' = 1$; and for loci with $d > a$ if $p' = p_m$, *i.e.* if the expected genotypic value of the tester population is at its maximum for such loci
- $\alpha' < 0$: for loci with $d > a$ if $p' > p_m$.

The reader is reminded that p_m is the allele frequency giving rise to the maximum of $\mathrm{E}\underline{\mathcal{G}}$ in the case of the Hardy–Weinberg genotypic composition (Section 9.2). At $d = a$ it amounts to 1, whereas $d > a$ implies $0 < p_m < 1$. Example 11.13 illustrates how α' depends on p'.

Example 11.13 Equation (11.48) describes how α' depends, for given values for a and d, on the allele frequency p' in the tester population. We consider the equation for loci B_3-b_3, B_4-b_4 and B_5-b_5, with $a_3 = a_4 = a_5 = 2$ and $d_3 = 0, d_4 = 1$ and $d_5 = 3$ of Example 9.5. According to Equation (9.9) $\mathrm{E}\underline{\mathcal{G}} - m$ attains for the locus with overdominance, *i.e.* locus B_5-b_5, a maximum value if $p_m = 0.833$. Figure 11.2 depicts α' as a function of p' for the three loci.

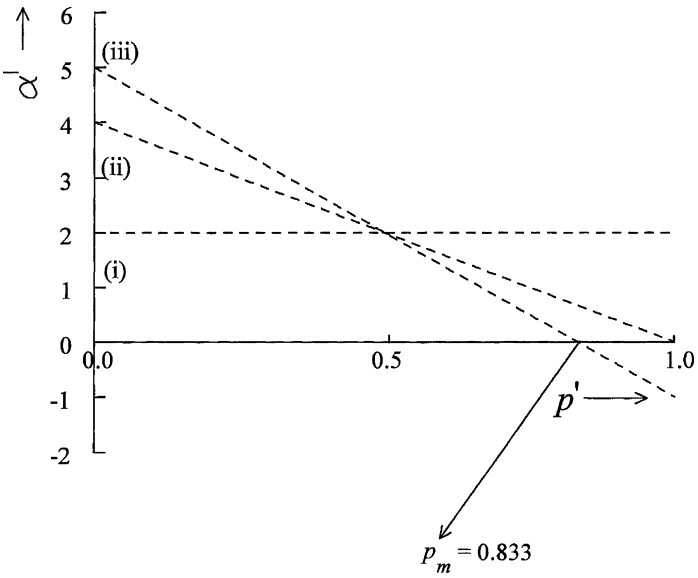


Fig. 11.2 The average effect of an allele substitution, *i.e.* α', as a function of p', the frequency of allele B in the tester population, for loci B_3-b_3, B_4-b_4 and B_5-b_5, with $a_3 = a_4 = a_5 = 2$ and $d_3 = 0(i), d_4 = 1(ii)$ and $d_5 = 3(iii)$

Ranking of the candidate genotypes for increasing breeding value, *i.e.* increasing value for

$$bv_j = (j - 2p)\alpha',$$

yields thus

- if $\alpha' > 0$
  $bv_{bb} < bv_{Bb} < bv_{BB}$, or: $bv_0 < bv_1 < bv_2$
- if $\alpha' = 0$
  $bv_0 = bv_1 = bv_2$
  Ranking is impossible for loci with $d \geq a$, if $p' = p_m$,
- if $\alpha' < 0$
  $bv_2 < bv_1 < bv_0$

Example 11.14 provides a numerical illustration of the foregoing.

Example 11.14 Locus B_5-b_5 of Example 11.13, with $a = 2$ and $d = 3$ is further considered (similar to Example 8.20). For this locus we have $p_m = 0.833$. We may calculate, according to Equation (8.26a), the average effect of an allele substitution for a population with $p = 0.875$ and $q = 0.125$:

$$\alpha' = 2 - (0.875 - 0.125)3 = -0.25$$

The allele effects (Equations (8.15) and (8.16) are thus

$$\begin{aligned} \alpha'_0 &= -0.875(-0.25) = 0.21875 \\ \alpha'_1 &= 0.125(-0.25) = -0.03125 \end{aligned}$$

and the breeding values (Equation (8.6) or (8.27b):

$$\begin{aligned} bv_0 &= 2(0.21875) = 0.4375 = (0 - 1.75)(-0.25) \\ bv_1 &= 0.21875 + (-0.03125) = 0.1875 = (1 - 1.75)(-0.25) \end{aligned}$$

and

$$bv_2 = 2(-0.03125) = -0.0625 = (2 - 1.75)(-0.25)$$

Because $d > a$ and $p' > p_m$ genotype bb is indeed the genotype with the highest breeding value.

In Section 11.2.2 it was shown how one might estimate $\text{var}(\underline{bv}) = \sigma_a^2$. In the case of a high value for $\text{var}(\underline{bv})$ prospects for successful selection are good. One may help achieve that by using an appropriate tester population as well as uniform environmental conditions in the progeny test. The choice of the tester is especially relevant for loci with overdominance or pseudo-overdominance. One should avoid using, with respect to such loci, a tester with $p' \approx p_m$, as such a tester would yield equivalent progenies. Figure 11.2 shows that α', and

consequently var($\underline{b\nu}$), is smaller as p' approaches either 1 or p_m. The former concerns loci with (in)complete dominance, the latter loci with overdominance. In both these cases the tester population will have a high expected genotypic value.

In practice it has often been observed that ${\sigma_a}^2$ does not decrease when applying selection (Hallauer and Miranda, 1981, p. 137; Bos, 1981, p. 91).

The effect of selection of candidates with a high breeding value on the expected genotypic value

In the context of progeny testing, the goal of the selection of candidates with a high breeding value is improvement of the genotypic value expected for the population subjected to the selection. It will be shown that this goal can not always be attained.

When combining the preceding text and the implications of Fig. 9.1, it can be deduced that selection of candidate plants with a high breeding value implies

- if $\alpha' > 0$
  An increase of p. This is associated with an increase of E$\underline{\mathcal{G}}$ if $0 \leq d \leq a$, or if $d > a$ as long as $p < p_m$. It is associated with a decrease of E$\underline{\mathcal{G}}$ if $d > a$ and $p > p_m$.
- if $\alpha' = 0$
  No change in p, *i.e.* no change in E$\underline{\mathcal{G}}$.
- if $\alpha' < 0$
  A decrease of p. This is associated with an increase of E$\underline{\mathcal{G}}$ as long as $p > p_m$. It is associated with a decrease of E$\underline{\mathcal{G}}$ if $p < p_m$.

It is assumed that absence of overdominance is the rule. The usual situation of presence of partial dominance or additivity, *i.e.* $0 \leq d \leq a$, implies then preferential selection of plants with genotype BB, *i.e.* an increase of p until $p = 1$. This is associated with an increase of E$\underline{\mathcal{G}}$.
For the relatively rare loci with overdominance ($d > a$) three situations concerning the tester population, namely $p' = p_m, p' < p_m$ and $p' > p_m$, have to be distinguished:

1. $p' = p_m$
   A tester population with $p' = p_m$ prohibits meaningful progeny testing for the involved loci: the progeny test does not allow successful selection among the candidates with regard to their breeding values.
2. $p' < p_m$
   In this case the tester produces pollen with haplotype b in such a frequency that candidates with genotype BB tend to yield superior offspring, if indeed $d > a$. Such candidates will be selected on the basis of the progeny test. The frequency of gene B will consequently increase.

3. $p' > p_m$
   When using a tester population with $p' > p_m$, candidates with genotype bb tend to produce superior offspring. Selection on the basis of the progeny test implies then a decrease of the frequency of allele B.

The above three situations for loci with overdominance require a more detailed treatment, both for

1. intrapopulation progeny testing and for
2. interpopulation progeny testing.

Intrapopulation progeny testing

Figure 11.3 illustrates how the allele frequency p will change, starting from the initial value p_0, in the case of continued selection of candidates with a high breeding values. This is done for a locus with $p_0 > p_m$ as well as for a locus with $p_0 < p_m$. The actual value of p_m depends, of course, on the values for a and d of the considered locus. In both cases p approaches p_m asymptotically. The closer p_m is approached, the smaller the differences in breeding and the smaller the heritability, *i.e.* the less efficient the selection. The changes in p become then smaller. At $p = p_m$ all genotypes have the same breeding value. In that situation the expected genotypic value ($E\underline{G}$) is maximal. Further improvement is then impossible.

Figure 11.4 depicts the same initial situation. Now, however, it is assumed that the selection results immediately in gene fixation, *i.e.* in $p_1 = 0$ (if $p_0 > p_m$) or in $p_1 = 1$ (if $p_0 < p_m$). This may occur when selecting only a few candidate genotypes on the basis of testing progenies obtained from a polycross.

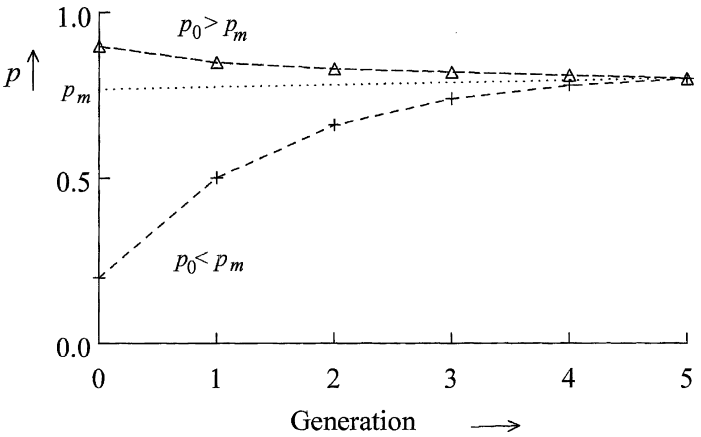


Fig. 11.3 The presumed frequency of allele B in successive generations with selection, based on intrapopulation testing, of candidates with a high breeding value; for a locus with $p_0 > p_m$ as well as a locus with $p_0 < p_m$ in the case of continuous change of p

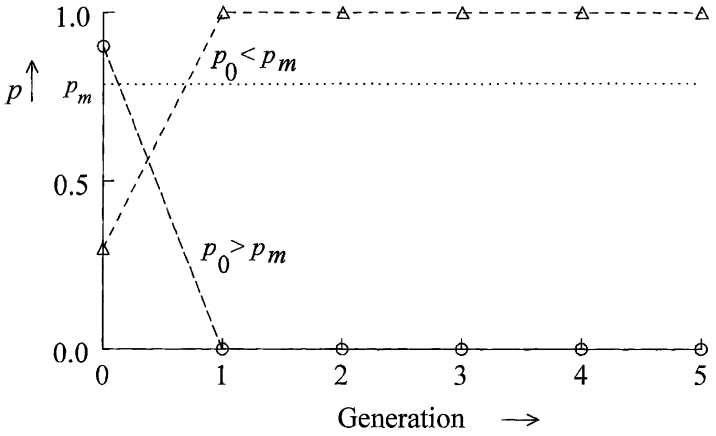


Fig. 11.4 The presumed frequency of allele B in successive generations when selecting, based on intrapopulation testing, candidates with a high breeding value; for a locus with $p_0 > p_m$ as well as a locus with $p_0 < p_m$ in the case of fixation after selection in generation 0

If the aim is to develop a synthetic variety the result may be disappointing: the maximum value for $\mathrm{E}\underline{\mathcal{G}}$ will never be attained.

Still another possibility is that selection starting with $p_0 < p_m$ gives successively rise to $p_1 > p_m, p_2 < p_m, p_3 > p_m$, *etc.* (or that selection starting with $p_0 > p_m$ gives successively rise to $p_1 < p_m, p_2 > p_m, p_3 < p_m$, *etc.*). Then p oscillates around p_m. Notwithstanding the presence of genetic variation the selection results in at most a small progress of $\mathrm{E}\mathcal{G}$, associated with dampening of the oscillation.

Interpopulation progeny testing

Interpopulation progeny testing occurs when applying recurrent selection (for general combining ability or specific combining ability, Section 11.5) or reciprocal recurrent selection. In this paragraph attention is focussed on **reciprocal recurrent selection** (RRS). In RRS two populations, say A and B, are involved. Plants in population A are selected because of their breeding values when using population B as tester. Likewise, and simultaneously, plants in population B are selected because of their breeding values when using population A as tester. (In an annual crop such as maize the S_1 lines obtained from the plants appearing to have a superior breeding value are used to continue the programme.)

It is likely that the allele frequencies of populations A and B differ more as these populations are less related. If indeed the allele frequencies are very different, it is probable that

$p_A > p_m > p_B$, or – at a different labelling of the populations – that $p_A < p_m < p_B$,

where p_A designates the allele frequency in population A and p_B the allele frequency in population B. The first situation implies testing of candidates representing population A with a population with p_B such that

$$\alpha' = (a + d) - 2p_B d > 0$$

(see Equation (11.48)). Selection in population A will then tend to yield an increase of p_A. It also implies testing of candidates representing population B with a tester with p_A such that $\alpha' < 0$. Selection in population B tends then to yield an decrease of p_B. These tendencies are illustrated in Fig. 11.5.

Continued selection will then, eventually, yield the desired goal, viz. two populations mutually adapted such that a bulk cross between them yields, with regard to loci affecting the considered trait and with $d > a$, exclusively heterotic, heterozygous plants.

Figure 11.6 depicts the development of the allele frequencies if the initial value of p_A is equal to p_m. This implies for the candidates genotypes in population B that $\alpha' = 0$. Effective selection of candidates with a high breeding value is then impossible in population B. The results eventually obtained is, however, the same as in Fig. 11.5. This may even occur if $p_A < p_m$ and $p_B \ll p_A$. Then, due to the first cycle of reciprocal recurrent selection, p may be increased in both populations such, that $p_A > p_m$ and $p_B < p_m$ (Fig. 11.7).

To help ensure that populations A and B have very different allele frequencies with regard to a large number of loci with $d > a$, these populations may be chosen on the basis of an evaluation of the performance of plant material produced by bulk crossing of a number of populations. Eligible populations are: open pollinating varieties, synthetic varieties, DC-, TC- and SC-hybrid varieties. If for a certain locus p_A and p_B are very similar, interpopulation

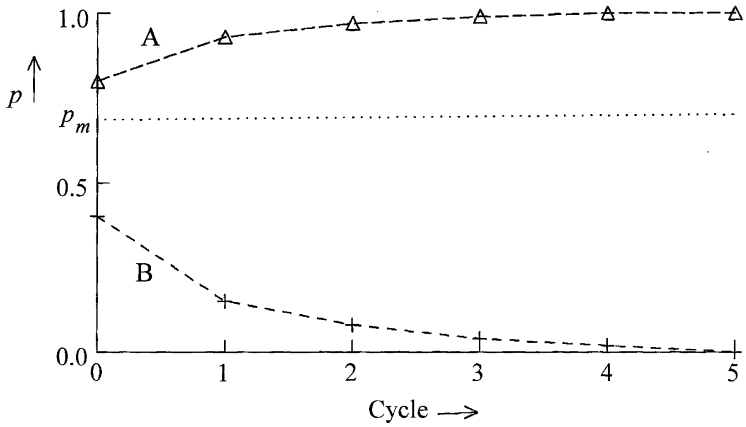


Fig. 11.5 The presumed frequency of allele B in successive cycles of reciprocal recurrent selection in populations A and B, for a locus with an initial allele frequency (p_0) such that $p_0 > p_m$ in population A and $p_0 < p_m$ in population B

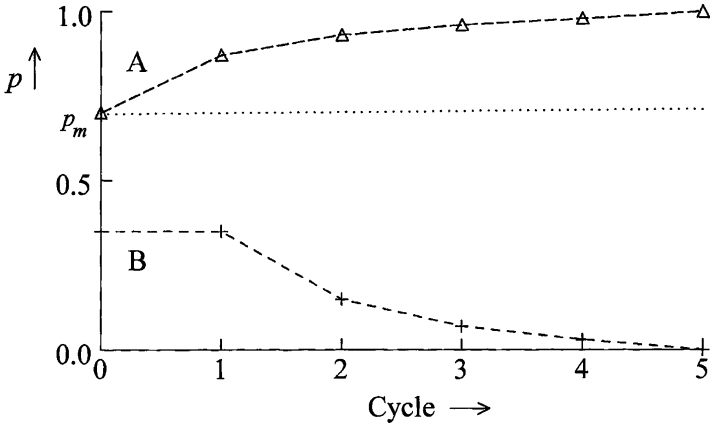


Fig. 11.6 The presumed frequency of allele B in successive cycles of reciprocal recurrent selection in populations A and B, for a locus with an initial allele frequency (p_0) such that $p_0 = p_m$ in population A and $p_0 < p_m$ in population B

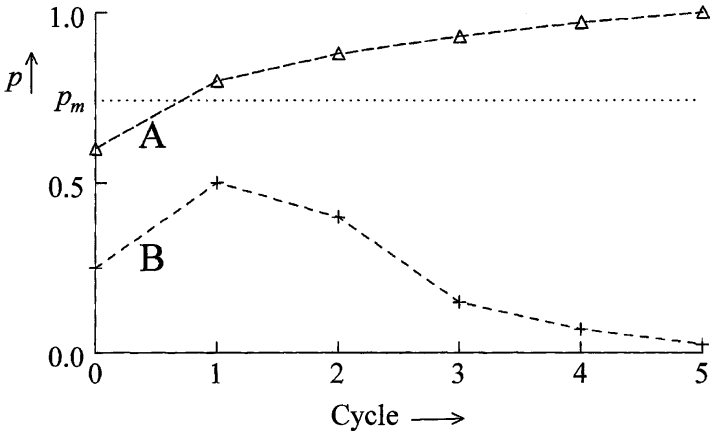


Fig. 11.7 The presumed frequency of allele B in successive cycles of reciprocal recurrent selection in populations A and B, for a locus with strongly different initial allele frequencies (but both smaller than p_m)

progeny testing resembles intrapopulation progeny testing. The selection will then, in both populations, induce p to approach p_m. (This is illustrated in Fig. 11.8 for $p_A \approx p_B$, where both are less than p_m). The result of continued selection will then be two populations with the Hardy-Weinberg genotypic composition, thus two populations with $\mathrm{E}\underline{\mathcal{G}}$ being equal to its maximum, *i.e.*

$$m + \frac{a^2 + d^2}{2d}$$

(Equation (9.10)). For loci with $d > a$ this maximum is less than $m + d$.

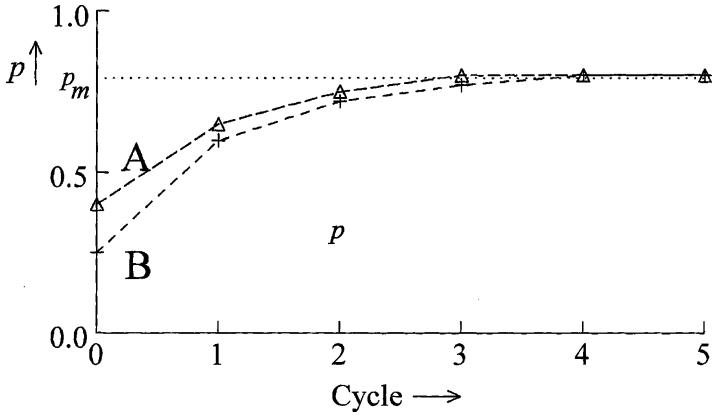


Fig. 11.8 The presumed frequency of allele B in successive cycles of reciprocal recurrent selection in populations A and B, for a locus for which the initial allele frequencies are very similar

The ultimate goal of reciprocal recurrent selection is plant material obtained by a bulk cross of the improved populations. The expected genotypic value of that plant material is, due to the presence of genetic variation, less than the highest possible genotypic value $m+d$, *i.e.* the genotypic value of the heterotic heterozygous genotype.

11.4 Choice of Parents and Prediction of the Ranking of Crosses

Prior to actual selection among evaluated candidates, the breeder selects among conceivable crosses. Parents will only be crossed if the progeny to be obtained are expected to be promising enough to be rewarding for the efforts of the crossing work. It is, of course, very attractive to be able to determine beforehand which crosses have the highest chance of producing a commercially desirable cultivar. This allows valuable time and efforts to be concentrated on crosses with a higher probability of producing desirable genotypes. A **cross prediction method** is, of course, only useful to a plant breeder if it is effective in handling large numbers of crosses.

Crops differ considerably with regard to the amount of work involved in a pollination. A single pollination of a cucumber flower, for instance, may yield hundreds of seeds. In contrast, the efforts required for the pollination of a single wheat ear, for instance, are considerable. A single pollination requires emasculation, in time, of the flowers alongside the ear to be pollinated, bagging of the ear, collection of the pollen and its transfer to the stigma of the flowers to be pollinated, and bagging again. Additionally the breeder should administrate

the parents involved in the pollination. All this work will, hopefully, result in only one seed per pollinated flower. It should be clear that it may be wise to consider seriously the crosses to be made.

Often crosses are made on the basis of implicit expert knowledge, but the choice may be supported by explicit information. Schut (1998) distinguished five sources of such information:

1. Information about *the phenotypes* of the potential parents.
2. Information about *the genotypes* of the potential parents with regard to traits with known genetic control.
3. Information about *differences between the potential parents* with regard to:

   - their geographic origin,
   - their pedigrees
   - their values for a set of traits.

The size of the difference is thought to indicate the number of heterozygous loci in the F_1. This number is, in its turn, thought to determine the heterosis in the F_1 and/or the genetic variance in the segregating generations. Crossing of distantly related lines with desired genotypic values for the relevant traits, which are due to different genotypes, is expected to increase the probability of **transgression** in the segregating populations.

> *N.B.* Transgression occurs if the segregating population contains with regard to some trait one or more lines with a phenotypic value outside the range given by the parental phenotypic values.

Pedigree data offer an opportunity to calculate the degree of relatedness of related parents. Such data are, however, often incomplete or unreliable.

The pedigree information can be quantified by a measure of relatedness of two potential parents, for instance by the coefficient of coancestry (Falconer, 1989).

The traits information may concern:

- agronomic traits,
- morphologic traits,
- biochemical traits (like isozymes, storage proteins) or
- molecular markers.

For agronomic and morphologic traits expressed in a continuous or ordinal scale one can quantify the difference between parents by calculating the Euclidean distance or the generalized distance (Snedecor and Cochran, 1980). For biochemical and molecular marker data one may use the following measure for genetic similarity of genotypes i and j:

$$gs_{if} = \frac{2N_{ij}}{N_i + N_j}$$

where

N_{ij} = number of bands present in both i and j

N_i = number of bands present in i

N_j = number of bands present in j

Transgression may occur at a large genetic distance between potential parents. The greater the distance (up to a certain limit), the larger the number of segregating loci and the larger the probability of transgression.

4. Information about *the performance as a parent of the pursued genotype(s).* Such information is obtained from earlier breeding cycles or earlier test crosses (for example a diallel cross yielding information about general combining ability and about specific combing ability (Section 11.5.2)).
5. Information about *the performance of early generation progenies* from crosses involving the potential parents. From these one can estimate the mean and the variance as expected to apply to later generations.

Sources 1 and 2 deal with qualitative traits, such as growth habit of barley lines, viz. *erectoides* versus *nutans*. Sources 3–5 deal with information about quantitative traits. Parents are crossed in such a way that weaknesses of one parent are compensated for by the other parent.

Jensen (1988, pp. 423–444, 449–469) reviewed the topic of choosing parents extensively. Indeed, the association between genetic distance and probability of transgression has often been studied. A number of scientists advocated the crossing of parents with a low genetic similarity, but experimental evidence supporting this advice is scarce (Example 11.16). Crossing of divergent lines often yields populations with a low mean performance due to one of the parents involved. Linkage groups of favourable genes are broken at meiosis of the heterozygous plants. Such groups are difficult to recover in later generations.

Brown and Caligari (1989) studied cross prediction based on evaluation of parental genotypes, or their offspring obtained after selfing. Thus mid-parent phenotypic values, *i.e.*

$$\tfrac{1}{2}(\overline{p}_{\mathrm{P1}} + \overline{p}_{\mathrm{P2}})$$

and mid-line phenotypic values, *i.e.*

$$\tfrac{1}{2}(\overline{p}_{\mathrm{L(P1)}} + \overline{p}_{\mathrm{L(P2)}})$$

were used as predictions.

In Section 9.1 it was shown that the latter two procedures may be expected to be reliable for traits where dominance does not play a role in the genetic control. Example 11.15 provides some results.

Example 11.15 Brown and Caligari (1989) analysed data from an experiment with potatoes. According to the rank correlation coefficient, cross rank – in the second clonal year – for breeder's preference and for total

yield appeared to be best predicted by seedling performance ($r = 0.48$ and 0.95, respectively). For mean tuber weight and number of tubers (these are the two yield components), the predictions based on mid-line values turned out to be the best (with $r = 0.68$ and 0.80, respectively). This may indicate the presence of an additive mode of inheritance for yield components. (This phenomenon underlies the explanation of hybrid vigour by the theory of recombinative heterosis (Section 9.4.1).)

Example 11.16 presents some results of a study to procedures for cross prediction based on relationship measures.

It has to be emphasized that information sources 4 and 5 do, in fact, not provide information with regard to crosses still to be made. They merely indicate which already existing segregating populations are most promising.

Example 11.16 In order to be able to draw general conclusions, Schut (1998) studied 20 cross populations resulting from crosses involving 18 European two-row spring barley varieties. Each population was represented by 48 pure lines, developed by continued selfing applied in the absence of selection. (Such sets of lines are called **recombinant inbred lines**; RILs). The RILs were tested along with their parents by means of 10-row plots in each of 7 environments, distributed over two years. Four traits were studied: plant height, flowering time, thousand kernel weight and grain yield.

For each pair of parents underlying the cross populations four relationship measures were calculated

- Genetic similarity (gs) based on marker data (Section 12.3.2)
- Coefficient of coancestry (f) based on pedigree data
- Morphologic distance (md)
- Agronomic distance(ad) based on multi-environment data for several agronomic traits

The study resulted into the following correlations, estimated across the 18 pairs of parents and the 18 cross populations, between the relationship of the parents and the variance between the RILs with regard to the studied traits:

- The correlations between $1-gs$ and the variances were generally positive, but rarely significant. This disappointing result was said to be due to a poor genomic representation of the genes affecting the traits by the markers.
- The correlations between $1-f$ and the variances were positive but non-significant. (This concerned only those ten crosses for which reliable pedigree data were available).
- The correlations between md and the variances were non-significant.

- The correlations between *ad* and the variances were mainly positive and sometimes significant. The correlations between *ad* for just height or just flowering date and RIL variance for height, respectively flowering time were significant.

Combined relationship measures generally had the highest correlations with RIL variance. Schut concluded, altogether, that the studied correlations were not high enough to be useful for practical breeding.

With regard to that topic, crosses, in fact segregating populations, may be ranked according to some criterion. In a self-fertilizing crop crosses may, for instance, be ranked according to

- Their ability to give rise to entries with a genotypic value exceeding some minimum, say $\mathcal{G}_{min}$. This may involve ranking of crosses with regard to $\mathrm{P}(\underline{\mathcal{G}} > \mathcal{G}_{min})$, *i.e.* the probability that the genotypic value of some obtained genotype exceeds $\mathcal{G}_{min}$. The probabilities are then predicted on the basis of estimates of m and $\sum_i a_i^2$.
- The observed proportion of (F_3) lines with a mean phenotypic value exceeding $\mathcal{G}_{min}$.

Reliability of the prediction of the performance of the progenies to be obtained when crossing parents is, of course, very desirable. Genotype by environment interaction is, of course, a disturbing phenomenon. If such interaction occurs, predictions on the basis of data collected in a certain macro-environment (year and/or location) will be of little value for other macro-environments. Furthermore the reliability of cross prediction methods is questionable in as far as the estimators of the statistical parameters are biased and/or inaccurate.

In the case of a normal probability distribution of the genotypic values, *i.e.*

$$\underline{\mathcal{G}} = N(\mathrm{E}\underline{\mathcal{G}}, \sigma_g^2),$$

one can predict $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$ on the basis of estimates of $\mathrm{E}\underline{\mathcal{G}}$ and σ_g^2. This is elaborated for plant material with identical reproduction (Section 11.4.1) and for self-fertilizing crops (Section 11.4.2).

Cross prediction with regard to several traits deserves attention because selection is rarely focussed on only a single trait. The probability that an inbred line has a satisfactory genotypic value for two or more traits simultaneously cannot be calculated as the product of the probabilities for the separate traits, unless the traits are not correlated. Multivariate cross prediction procedures require, in addition to knowledge of m and of $\sum_i a_i^2$ for each character, also knowledge of the genetic correlation coefficient, ρ_g (Section 12.2), between each pair of characters. Powell *et al.* (1985b) present an application of multivariate cross prediction methods.

11.4.1 Plant Material with Identical Reproduction

This section gives attention to the prediction of the ranking of crosses dealing with plant material with identical reproduction, *e.g.* clones, pure lines (especially DH-lines). The conditions required for a reliable prediction of the probability that the genotypic value of some genotype exceeds some minimum, *i.e.* $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$, are

1. A normal distribution of the genotypic values
2. Absence of genotype × environment interactions

When estimating $\mathrm{E}\underline{\mathcal{G}}$ by $\overline{p}$ and $\mathrm{var}(\underline{\mathcal{G}})$ on the basis of a completely randomized experiment or randomized (complete) blocks (Section 11.2.1), one may predict $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$ by:

$$P\left(\frac{\underline{\mathcal{G}} - \overline{p}}{\sqrt{\hat{var}(\underline{\mathcal{G}})}} > \frac{\mathcal{G}_{min} - \overline{p}}{\sqrt{\hat{var}(\underline{\mathcal{G}})}}\right) = P\left(\underline{\chi} > \frac{\mathcal{G}_{\min} - \overline{p}}{\hat{\sigma}_g}\right) = 1 - \Phi\left(\frac{\mathcal{G}_{min} - \overline{p}}{\hat{\sigma}_g}\right) \tag{11.49}$$

This probability can be read from a table presenting values of the standard normal distribution. The probability can be predicted for each of a number of families ('crosses') and this allows ranking of the crosses. The coefficient of correlation between predicted rank and actual rank indicates the reliability of the prediction. Examples 11.17 and 11.18 give illustrations.

Example 11.17 In 1981, Caligari and Brown (1986) raised, for each of eight potato crosses, seedlings in 10 cm square pots in a glasshouse. In 1982 each genotype that produced sufficient tubers was grown in a field experiment. In 1983, *i.e.* the second clonal year, each cross was represented by 70 randomly chosen clones. These were grown in a field in Blythbank in two randomized complete blocks consisting of three-tuber plots. Both in 1981 and 1983 potato breeders assigned, on the basis of visual assessment of tubers, to each clone a phenotypic value for 'preference score'. From these data values for $\overline{p}$ and $\hat{\sigma}_p$ (for 1981) and for $\overline{p}$ and $\hat{\sigma}_g$ (for 1983) were obtained for each cross.

For the 1981 data of cross C_1, for instance, these values were: $\overline{p} = 4.36$ and $\hat{\sigma}_p = 1.52$. Thus for the minimal acceptable preference score $\mathcal{G}_{min} = 5$ one can calculate

$$P\left(\underline{\chi} > \frac{5 - 4.36}{1.52}\right) = P(\underline{\chi} > 0.421) = 0.337$$

For the seven other crosses the following probabilities were estimated:

$$\mathrm{C}_2 : 0.274, \mathrm{C}_3 : 0.176, \mathrm{C}_4 : 0.251, \mathrm{C}_5 : 0.015, \mathrm{C}_6 : 0.192,$$
$$\mathrm{C}_7 : 0.281, \text{ and } \mathrm{C}_8 : 0.117.$$

For the glasshouse conditions of 1981 the crosses could thus be ranked as:

$$C_5 < C_8 < C_3 < C_6 < C_4 < C_2 < C_7 < C_1$$

For the 1983 data of $C_1, P(\underline{\mathcal{G}} > \mathcal{G}_{min})$ can likewise be predicted to amount to 0.119. The actual proportions of clones with a preference score of at least 5 amounted to 0.217 in 1981 (the average of the estimated probabilities amounted then to 0.205) and to 0.157 in 1983.

The coefficient of correlation, across the eight crosses, between the predicted probabilities and the observed proportions were 0.96 in 1981 (the average of the estimated probabilities amounted then to 0.205) and 0.91 in 1983. The coefficient of correlation between probabilities predicted on the basis of the 1981 data (which were obtained from seedlings raised in a glasshouse) and the proportions observed in 1983 was as high as 0.59.

It was concluded that $\overline{p}$ and $\hat{\sigma}$ estimated from the data in any environment provided a good prediction of the number of clones in each cross that would exceed some defined minimum preference score.

Example 11.18 Fifty-two *Solanum tuberosum* crosses were chosen deliberately to represent the range, in commercial breeding material, with regard to their preference scores. In the spring of 1984, eighty seedlings from each cross were sown into seed pans and later transplanted into 10 cm square pots (Brown *et al.*, 1988). Two tubers were taken from each genotype to be used in 1985, the first clonal year.

In 1985 the 52 crosses were grown in each of four completely randomized blocks in Blythbank and in Murrays. Each plot contained 15 genotypes, together representing the involved family. After assessment, the produce from each of the $52 \times 4 \times 15 = 3{,}120$ genotypes was used in 1986, the second clonal year.

In 1986 each cross was represented by 40 clones at Blythbank and by 20 clones, a subsample of the 40 clones evaluated at Blythbank, at Murrays. At each site each clone was grown as a four-plant, single-row plot.

Each year the mean value per clone for the visually assessed breeder's preference score of the tubers was determined. The minimal acceptable score was 5.

For Blythbank the coefficient of correlation between the mean score for each of the 52 families in 1985 and those in 1986 amounted to 0.91; the correlation between the results from Blythbank (1985 data) and Murrays (1986 data) was 0.70. From the $52 \times 40 = 2,080$ clones that were grown in Blythbank in both years, 222 scored at least 5 in 1985, 181 did so in 1986, but only 69 did so in both years. Thus $181 - 69 = 112$ (*i.e.* 62%) of the second clonal year selections would have been discarded in the first year. This implies that a high proportion of potentially desirable clones would have been lost if individual clone selection was practised in 1985!

For each site/year combination the following quantities were determined per family: $\overline{p}, \hat{\sigma}_p$ and the prediction of $P(\underline{\mathcal{G}} > 5)$. The coefficient of correlation, across the 52 crosses, between site/year combinations ranged for $\overline{p}$ from 0.70 to 0.89. For the prediction of $P(\underline{\mathcal{G}} > 5)$ it ranged from 0.59 to 0.76. All correlations were highly significant and it should thus be possible to identify the 'better' crosses on the basis of data from seedlings grown in pots.

11.4.2 Self-fertilizing Plant Material

If the genotypic values of the homozygous genotypes in an F_∞ population of a self-fertilizing crop have a normal distribution, the probability distribution of $\underline{\mathcal{G}}$ is completely specified by $E\underline{\mathcal{G}}$ and σ_g^2. Under the conditions specified below, one may predict these parameters from data collected from the parents and from a random sample of F_3 lines. Then one may predict the probability that the genotypic value of an F_∞ plant exceeds $\mathcal{G}_{min}$.

The conditions required for a reliable prediction are the following:

1. A normal distribution of the genotypic values
2. Absence of epistasis
3. Absence of linkage
4. Absence of genotype × environment interactions

If condition 1 applies the probability distribution of the genotypic values of the plants in population F_∞ is given by

$$\underline{\mathcal{G}} = N(m, \mathrm{var}(\underline{\mathcal{G}}_{F\infty}))$$

Condition 2 is required to estimate parameter m by means of Equation (11.46):

$$\hat{m} = \tfrac{1}{2}\left(\overline{p}_{P_1} + \overline{p}_{P_2}\right)$$

If conditions 2 and 3 are satisfied, $\mathrm{var}(\underline{\mathcal{G}}_{F\infty})$ is equal to $\sum_i a_i^2$ (Table 10.3). A biased but relatively accurate estimate of this quantity is $2\hat{\mathrm{var}}(\underline{\mathcal{G}}_{LF3})$ (Equation (11.47)). The probability distribution of F_∞ can thus be predicted.

An interesting application, *i.e.* prediction of $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$, requires condition 4. If the condition applies, the probability that some F_∞ plant to be obtained in the future has a genotypic value exceeding $\mathcal{G}_{min}$, is predicted by:

$$P\left(\frac{\underline{\mathcal{G}} - \hat{m}}{\sqrt{\hat{var}(\underline{\mathcal{G}}_{F_\infty})}} > \frac{\mathcal{G}_{min} - \hat{m}}{\sqrt{\hat{var}(\underline{\mathcal{G}}_{F_\infty})}}\right) = P\left(\underline{\chi} > \frac{\mathcal{G}_{\min} - \hat{m}}{\hat{\sigma}_g}\right) = 1 - \Phi\left(\frac{\mathcal{G}_{min} - \hat{m}}{\hat{\sigma}_g}\right) \tag{11.50}$$

Calculation of this probability may be rewarding. When for two segregating populations the means m_1 and m_2 and the genetic variances $\hat{\mathrm{var}}_1(\underline{\mathcal{G}}_{F\infty})$ and

$\hat{\text{var}}_2(\underline{\mathcal{G}}_{\text{F}\infty})$ differ such, that $m_1 > m_2$ and $\hat{\text{var}}_1(\underline{\mathcal{G}}_{\text{F}\infty}) < \hat{\text{var}}_2(\underline{\mathcal{G}}_{\text{F}\infty})$, then it is of interest to calculate $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$ for each population. Example 11.19 illustrates calculation of $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$, Example 11.20 discusses some results.

Example 11.19 It is shown how one may calculate the probability that the genotypic value of some plant, belonging to an F_∞ population to be developed, lies outside the range between the genotypic values of the two parents, *i.e.* $P(\underline{\mathcal{G}} < \mathcal{G}_{\text{P2}}) + P(\underline{\mathcal{G}} > \mathcal{G}_{\text{P1}})$, where $\mathcal{G}_{\text{P2}} < \mathcal{G}_{\text{P1}}$.

In the case of a normal probability distribution of the genotypic values, the probability distribution is symmetric around m. As Equation (11.45)

$$\boldsymbol{m} = \tfrac{1}{2}\left(\mathcal{G}_{P_1} + \mathcal{G}_{P_2}\right)$$

implies

$$\mathcal{G}_{\text{P1}} - m = m - \mathcal{G}_{\text{P2}},$$

i.e. $\mathcal{G}_{\text{P1}}$ is as much larger than m as $\mathcal{G}_{\text{P2}}$ is smaller than m, it follows that

$$P(\underline{\mathcal{G}} < \mathcal{G}_{\text{P2}}) = P(\underline{\mathcal{G}} > \mathcal{G}_{\text{P1}})$$

This means that

$$P(\underline{\mathcal{G}} < \mathcal{G}_{\text{P2}}) + P(\underline{\mathcal{G}} > \mathcal{G}_{\text{P1}}) = 2P(\underline{\mathcal{G}} > \mathcal{G}_{\text{P1}})$$

This probability is equal to

$$2P\left(\frac{\underline{\mathcal{G}} - \hat{m}}{\sqrt{\hat{var}(\underline{\mathcal{G}}_{\text{F}\infty})}} > \frac{\mathcal{G}_{\text{P1}} - \hat{m}}{\sqrt{\hat{var}(\underline{\mathcal{G}}_{\text{F}\infty})}}\right) = 2P\left(\underline{\chi} > \frac{\mathcal{G}_{\text{P1}} - \hat{m}}{\hat{\sigma}_g}\right) = 1 - 2\Phi\left(\frac{\mathcal{G}_{\text{P1}} - \hat{m}}{\hat{\sigma}_g}\right)$$

Jinks and Pooni (1976) present three applications where predicted probabilities and actual proportions coincided fairly well. Their first application concerned a cross of two pure lines of *Nicotiana rustica* L. For plant height, as observed in 1954 and measured in inches, they reported

$$\hat{m} = 43.29,$$
$$\hat{\text{var}}(\underline{\mathcal{G}}_{\text{F}\infty}) = (5.69)^2, \text{and}$$
$$\mathcal{G}_{\text{P1}} = 44.69.$$

This yields for the above probability

$$2P\left(\underline{\chi} > \frac{44.69 - 43.29}{5.69}\right) = 0.81$$

In the same season 20 random inbred lines representing F_{10} were grown. The season's growing conditions were intermediate in a group of 16 growing seasons. The average plant height of the 20 lines amounted to 44.56. Eight lines were shorter than P_2 and 10 lines were taller than P_1. Thus the actual proportion of lines outside the range of parental genotypic values was 0.9.

Example 11.20 Schut (1998) studied the F_4 and F_∞ generation of 20 barley crosses. For each cross both the F_4 and the F_∞ generation were represented by 48 lines tracing back to the same set of 48 F_2 plants. The F_4 lines were tested at two locations in 1994; the related 'recombinant inbred lines' (RILs) were tested at two locations in 1995 and at four locations in 1996. Schut (1998; p. 33) found that the yields of the 20 RIL populations, each averaged over the six environments, were only moderately correlated ($r = 0.42$) with the yields of the 20 F_4 populations. Mid-parent values, based on small plot yield data from the same two trials as the F_4 evaluation showed a similar correlation ($r = 0.45$) with the yields of the RIL populations. Mid-parent values based on 1994 yield data from large plots at the same locations showed, however, a much higher correlation ($r = 0.70$). This correlation is about equal to the correlation between RIL population yields and mid-parent yields based on large plots in the same six environments where the RIL populations were tested ($r = 0.71$).

Schut concluded that a labourious early generation small plot yield assessment offered hardly any perspective for practical breeding, neither for selection within crosses nor for selection between crosses.

Schut predicted for the F_∞ generation of each of the 20 cross populations $P(\underline{\mathcal{G}} > \mathcal{G}_{min})$, with $\mathcal{G}_{min}$ = average yield of three standard cultivars. These probabilities were correlated with the observed proportion of RILs yielding more than $\mathcal{G}_{min}$. The correlations were virtually absent when estimating m on the basis of the small plot trials of 1994, either the mid-parent value or the F_4 population mean (Schut, 1998; p. 37). When estimating m on the basis of mid-parent values of large plot trials in six environments, the average rank correlation was only 0.22. Also directly observed proportions of F_4 lines yielding in the small plot trial more than $\mathcal{G}_{min}$ were not clearly related with the observed proportions in the F_∞ generation.

In addition to the foregoing, one may perhaps wish to predict the genotypic values of the two extreme homozygous genotypes (Jinks and Perkins, 1972). These values are

$$m - \sum_i a_i \text{ and } m + \sum_i a_i$$

Prediction of these values requires estimates of m and $\sum_i a_i$. The latter quantity may be estimated when assuming a constant degree of dominance across all relevant loci, *i.e.*:

$$\frac{d_i}{a_i} = c$$

Then one may derive

$$\sum_i d_i \cdot \sqrt{\frac{\sum_i a_i^2}{\sum_i d_i^2}} = \sum d_i \cdot \sqrt{\frac{\sum a_i^2}{c^2 \sum a_i^2}} = \sum d_i \cdot \frac{1}{c} = \sum d_i \cdot \frac{\sum a_i}{\sum d_i} = \sum_i a_i$$

According to Table 9.1, the quantity $\sum_i d_i$ may be estimated by

$$\hat{\mathcal{G}}_{F1} - \hat{m}$$

The quantity

$$\sum_i a_i^2$$

is estimated as

$$2\hat{\text{var}}(\underline{\mathcal{G}}_{\text{LF3}})$$

(Equation (11.47))
and

$$\sum_i {d_i}^2$$

can, for instance, be estimated on the basis of Equations (10.24), (10.25) or (10.27). The reliability of this approach for estimating $\sum_i a_i$ is questionable. In the case of presence of one or more loci with additive effects, for instance, it yields a false result. Example 11.21 provides an illustration.

Example 11.21 Jinks and Perkins (1972) observed plant height (in inches) of *Nicotiana rustica* plants. They obtained from their data the following estimates:

$$\sum_i \hat{d}_i = 6.11$$

$$\sum_i \hat{a}_i^2 = 30.69$$

$$\sum_i \hat{d}_i^2 = 4.08$$

Thus

$$\sum_i \hat{a}_i = \frac{6.11}{\sqrt{\frac{4.08}{30.69}}} = 16.76$$

implying for the genotypic values a predicted range of 33.5.

Starting with 100 F_2 plants, 82 F_8 lines were obtained with a plant height ranging from 34.53 to 61.49. Thus the actual range amounted to 26.96.

11.5 The Concept of Combining Ability as Applied to Pure Lines

11.5.1 Introduction

The genetic quality of a genotype appears often poorly from the phenotype of the plant(s) representing the genotype, especially when the genotype is represented by only a single or a few plants. An alternative way of assessing the genetic quality of the genotype is by means of evaluation of progeny obtained from it. Indeed, in cross-fertilizing crops the application of selection based on progeny testing, *i.e.* selection for breeding value, is quite common. Candidate genotypes, representing some genetically heterogeneous population, are then pollinated by a tester population producing pollen with a diverse haplotypic composition (Section 11.3). Candidate genotypes yielding the best progenies are selected.

With regard to sets of pure lines something similar may be applied. The genetic quality of a pure line is then assessed on the basis of the progeny obtained by crossing the line with a tester population (in the present case consisting of a set of pure lines). This procedure may be applied to a self-fertilizing crop but also to a cross-fertilizing crop. The latter situation applies when testing pure lines with the goal to develop a hybrid variety. Candidate genotypes producing the best performing offspring are said to have the highest combining ability. The crossing design of the lines to be assessed may consist of a **diallel cross**, sometimes indicated as: a diallel set of crosses. In this case all N pure lines are crossed in pairwise combinations. The diallel cross is said to be **complete** if each line is crossed with all other lines. This will yield N^2 progenies, viz. N S_1-lines due to selfing, and N^2-N FS-families due to pairwise crosses. If selfing is omitted and reciprocal crosses are not made only $\frac{1}{2}N(N\text{-}1)$ FS-families will be obtained.

In this book it is assumed that the N candidate genotypes are pure lines. They may be designated as $P_1, P_2, \ldots, P_N$. The progenies may be coded as F_{ij}, where

- i refers to maternal parent P_i; with $i = 1, \ldots, N$
- j refers to paternal parent P_j; with $j = 1, \ldots, N$

Each progeny may be represented by a single plant or by a number of plants that are either cultivated as individually randomized plants or as J plots each containing K plants. The quantitative genetic interpretation of the observation characterizing the single cross hybrid progeny F_{ij} may thus range from 'the phenotypic value of a single plant representing the hybrid' to 'a precise estimate of the genotypic value of the hybrid'. For this reason the observation will be designated by the general symbol x_{ij}. Table 11.5 presents a summary of the observations derived from all progenies resulting from a complete diallel cross.

Table 11.5 The observation x_{ij} characterizing progeny F_{ij} obtained from a complete diallel cross involving pure lines $\mathrm{P}_1, \ldots, \mathrm{P}_N; i, j = 1, \ldots, N$. The margins of the table provide for each maternal parent as well as for paternal parent the mean progeny performance

		Paternal parent					
		P_1	.....	P_j	.....	P_N	
Maternal parent:	P_1	x_{11}	.....	x_{1j}	.....	x_{1N}	$\bar{x}_{1.}$
	.	.		.		.	.
	.	.		.		.	.
	P_i	x_{i1}	.....	x_{ij}	.....	x_{iN}	$\bar{x}_{i.}$
	.	.		.		.	.
	.	.		.		.	.
	.	.		.		.	.
	P_N	x_{N1}	.....	x_{Nj}	.....	x_{NN}	$\bar{x}_{N.}$
		$\bar{x}_{.1}$	.....	$\bar{x}_{.j}$	.....	$\bar{x}_{.N}$	$\bar{x}_{..}$

The set of progenies occurring in row i, *i.e.* $\{\mathrm{F}_{i1}, \ldots, \mathrm{F}_{iN}\}$, or the set of progenies occurring in column j, *i.e.* $\{\mathrm{F}_{1j}, \ldots, \mathrm{F}_{Nj}\}$, forms an HS-family, which may be designated by $\mathrm{F}_{i.}$ and $\mathrm{F}_{.j}$, respectively. A row as well as a column comprises the observations from all progenies descending from the same maternal parent or the same paternal parent, respectively. The average across row i, say $\bar{x}_{i.}$, or across column j, say $\bar{x}_{.j}$, represents the mean across the single cross hybrids constituting HS-family $\mathrm{F}_{i.}$ or $\mathrm{F}_{.j}$, respectively.

If the total number of $\frac{1}{2}N(N\text{-}1)$ progenies is unmanageably large, or if the breeder fails to produce all of them, for instance due to asynchronous flowering, a **partial diallel cross** (or **incomplete diallel cross**) may be made. This partial diallel cross may produce progenies according to a structured scheme, such as used for a balanced incomplete block design or an α-design, see Example 19.3, or it may produce progenies according to an unstructured ('wild') crossing design. In the former case the maternal parents play the role of the treatments and the paternal parents the role of the incomplete blocks. Care must be taken for a wild crossing design that it is a **connected design** (John, 1971; Breure and Verdooren, 1995).

In this book two reasons for making a diallel cross are elaborated

1. Prediction of the performance of a TC- or a DC-hybrid variety of a cross-fertilizing crop (Section 9.4.2). This application plays an important role in practical plant breeding aiming at the development of a hybrid variety.
2. Determination of the general combining ability of a pure line and/or the specific combining ability of a pair of pure lines. This application occurs rather frequently at research stations, possibly in the framework of the development of a new variety (Section 11.5.2).

11.5.2 General and Specific Combining Ability

It is of interest to know whether or not a pure line possesses a good **general combining ability** (*gca*), with regard to a tester population; or whether two pure lines have a good **specific combining ability** (*sca*) or not. (The precise definitions of these quantities are developed hereafter, see Equations (11.53) and (11.54)). It should thus be clear that the main interest when applying an analysis in terms of *gca* and *sca* is not in the progenies but in their parents. An analysis of a diallel cross in these terms is, indeed, a special way of progeny testing.

When applying a diallel cross the tester population consists of the set of inbred lines involved in the diallel cross. For inbred line i the value obtained for

$$\bar{x}_{i.}\text{-}\bar{x}..$$

where
$\bar{x}..$ designates the overall mean progeny phenotypic value,
may be considered as an estimate of its general combining ability. Thus the general combining ability of a pure line is indeed estimated from the performance of its offspring in comparison to the overall mean performance.

One may subtract from the expected genotypic value, calculated across all progenies descending from pure line i, the expected genotypic value calculated across all progenies. The quantity obtained is similar to the breeding value of line i, except for the factor 2 occurring in Equation (8.24). The variance of the *gca* values is, consequently, similar to the variance of the breeding values. One should, nevertheless, be cautious. The concepts of additive genotypic value, breeding value, additive genotypic variance and variance of the breeding values are applied in the context of panmictic populations. Only in that situation Equation (8.28), *i.e.*

$$\sigma_a{}^2 = \text{var}(\underline{b\nu}),$$

applies. In contrast the concepts of *gca* and *sca* apply to a different context, *viz.* to pure lines involved in a diallel cross.

The concepts of *gca* and *sca* are also used in other contexts than diallel crosses, *e.g.* recurrent selection for gca, recurrent selection for sca, reciprocal recurrent selection. The concepts have, consequently, been defined in different ways. Sprague and Tatum (1942), who introduced the terms gca and sca, used definitions different from those proposed by Griffing (1956). The approach of the latter, which is considered here, is similar to the one used for the statistical analysis of a two-way table. An analysis of the data resulting from a diallel cross in terms of *gca* and *sca* is thus primarily a statistical analysis. A two-way table may be analysed on the basis of a simple linear model

$$\text{E}\underline{x}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Such a model is also used for data obtained from a randomized complete block experiment such as used to compare the performances of a number of genotypes.

Griffing's parametrization of the genotypic value $\mathcal{G}_{ij}$ of the single cross hybrid obtained by pollinating maternal parent i by paternal parent j is:

$$\mathcal{G}_{ij} = \mu + gca_i + gca_j + sca_{ij} \tag{11.51}$$

where

μ = the overall mean

gca_i = the general combining ability of parent P_i

gca_j = the general combining ability of parent P_j

sca_{ij} = the specific combining ability of parents P_i and P_j

In the case of a complete diallel cross yielding N^2 progenies the formulae for estimating the parameters μ, gca_i and sca_{ij} in Equation (11.51) are straightforward:

$$\hat{\mu} = \bar{x}_{..} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}}{N^2} \tag{11.52}$$

$$\hat{gca}_i = \tfrac{1}{2}(\bar{x}_{i.} + \bar{x}_i) - \hat{\mu} = \frac{\sum_{j=1}^{N} x_{ij} + \sum_{j=1}^{N} x_{ji}}{2N} - \hat{\mu} \tag{11.53}$$

$$\hat{sca}_{ij} = \tfrac{1}{2}(x_{ij} + x_{ji}) - \hat{gca}_i - \hat{gca}_j - \hat{\mu} \tag{11.54}$$

It is easily shown that the sum of the gca values is zero, namely

$$\sum_{i=1}^{N} \hat{gca}_i = \tfrac{1}{2}\sum_{i=1}^{N}(\bar{x}_{i.} + \bar{x}_i) - N\hat{\mu} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij} + \sum_{j=1}^{N}\sum_{i=1}^{N} x_{ji}}{2N} - N\hat{\mu} = \frac{2N^2\hat{\mu}}{2N} - N\hat{\mu} = 0$$

This implies that the average gca value is bound to be zero. Likewise it is easily shown that for any line, for instance line i, the sum of the sca values is zero:

$$\sum_{j=1}^{N} \hat{sca}_{ij} = \sum_{j=1}^{N}(\tfrac{1}{2}(x_{ij} + x_{ji}) - \hat{gca}_i - \hat{gca}_j - \hat{\mu}) = \tfrac{1}{2}(x_{i.} + x_{.i}) - N\hat{gca}_i - N\hat{\mu}$$

$$= N\hat{gca}_i - N\hat{gca}_i = 0$$

Griffing (1956) elaborated the appropriate statistical analysis of data characterizing the progenies evolving from four different designs of a diallel cross, *i.e.* data from

1. The N^2 progenies obtained from a complete diallel cross
2. All parental pure lines plus all FS-families, reciprocals excluded, *i.e.* N S_1-lines and $\frac{1}{2}N(N-1)$ FS-families

3. All FS-families, reciprocals included, *i.e.* $N(N-1)$ FS-families
4. All FS-families, reciprocals excluded, *i.e.* $\frac{1}{2}N(N-1)$ FS-families

Both the analysis of variance according to a linear model assuming fixed effects and the analysis according to a linear model assuming random effects were elaborated (Kuehl, 2000, p. 148, 183–190). According to the model assuming fixed effects, the parents involved in the evaluated progenies are the subjects of study, whereas with the model assuming random effects interest is primarily in the population of pure lines represented by the random sample consisting of the N parents whose progenies were evaluated.

Designs 2 and 4 do not allow estimation of reciprocal differences, which may, for instance, be due to maternal effects via plasmagenes.

In Section 11.5.1 it was said that the genetic quality of a genotype might appear from an evaluation of its progeny. In the present section attention is focussed on progeny obtained from a diallel cross. An alternative for such progeny is the progeny obtained by selfing. Indeed, whenever a candidate has a valuable genotype its genetic value will appear from the quality of its offspring. The performance of offspring obtained by selfing is not at all affected by the tester genotype. Deleterious recessive genes hiding in the candidate genotype to be tested will clearly be exposed in the line obtained by selfing the candidate. For this reason, the authors are of the opinion that progeny testing of candidate genotypes by means of progenies obtained from selfing is a good alternative for progeny testing using progenies obtained from a diallel cross: it saves a lot of efforts (less crossing work, fewer progenies to be evaluated) and absence of disturbing tester effects (but possibly disturbing inbreeding effects due to the selfing; selfing might even be impossible due to self-incompatibility). Examples 11.22 and 11.23 support the opinion.

Example 11.22 Kinman and Sprague (1945) collected the grain yield data (in bushel per acre) of the progenies resulting from a maize diallel cross of the pure lines presented in Table 11.6.

Table 11.6 The grain yield (in bu/acre) of 10 pure lines of maize, *i.e.* $\hat{\mathcal{G}}_{\mathrm{P}}$, and the average grain yield of their offspring obtained from a diallel cross, say $\hat{\mathcal{G}}_{\mathrm{HS}}$. The rank, from lowest (1) to highest (10), is given in brackets (source: Kinman and Sprague (1945))

Line	$\hat{\mathcal{G}}_{\mathrm{P}}$		$\hat{\mathcal{G}}_{\mathrm{HS}}$	
CI14	2.7	(1)	61.6	(1)
Oh04	15.1	(2)	69.7	(3)
WV7	20.1	(3)	68.1	(2)
38-11	26.5	(4)	80.5	(8)
WF9	28.5	(5.5)	76.3	(5.5)
Oh07	28.5	(5.5)	78.4	(7)
Hy	31.9	(7)	71.2	(4)
B2	39.0	(8)	82.5	(9)
R46	39.8	(9)	76.3	(5.5)
K159	49.8	(10)	82.7	(10)

The coefficient of correlation of $\hat{\underline{\mathcal{G}}}_{\mathrm{P}}$ and $\hat{\underline{\mathcal{G}}}_{\mathrm{HS}}$ estimated from these data is 0.85, whereas the rank correlation is 0.74. In this example *gca* and performance *per se* are clearly related. Hallauer and Miranda (1981, pp. 281–283) concluded, on the basis of a literature review, that such a positive relation generally exists.

Example 11.23 Genter and Alexander (1962) reported to have been successful in improving *gca* by selection of the best S_1 lines of maize.

N.B. It is rather strange to report that *gca* has been improved as the average *gca* value is equal to zero.

In some cases intercrossing of the best lines yielded an improved population. Therefore, selection for an improved performance of S_1 lines plays a role of some importance in maize breeding (Hallauer and Miranda, 1981, p. 227).

N.B. The described procedure implies selection of the best S_1-lines. It is to be distinguished from so-called **simple recurrent selection**. In the latter procedure many plants are selfed. Only plants that are attractive both for traits expressed before and for traits expressed after pollen distribution are harvested. Thus the best parental plants are selected. In the next generation the S_1 lines tracing back to these plants are intercrossed without paying attention to the trait(s) to be improved.

Horner *et al.* (1973) applied so-called S_2 progeny selection in maize. With regard to ear yield, the 10-12 best S_2 lines were selected out of 60 S_2 lines (first cycle) or out of 100 S_2 lines (later cycles). The selected lines were intercrossed to start a new 'cycle'. Across five cycles, progress of 2% per cycle was obtained. This progress was measured with plant material obtained from crosses with genetically heterogeneous testers.

When selecting with regard to ear yield of families obtained by crossing S_1 plants (first cycle) or S_1 lines (later cycles) with an inbred line, the progress amounted to 4% per cycle.

In Section 11.5.1 it was said that the genetic quality of a pure line can be assessed from the progenies resulting from a diallel cross in a way similar to the assessment of the breeding value of an open pollinating candidate. Indeed, an analysis of the data resulting from a diallel cross in terms of *gca* and *sca* is primarily a statistical analysis. It is, however, interesting to compare the pure line quantities *gca* and *sca* with the open pollinating candidate quantities breeding (bv) value and dominance deviation (δ). For this reason the quantitative genetic interpretation of the concepts *gca* and *sca* is developed (better than the rough quantitative genetic interpretation of *sca* given in Note 9.1).

The concept of breeding value applies to segregating populations of cross-fertilizing crops; the concept of general or specific combining ability applies to sets of pure lines. There is, nevertheless, a rather close relationship between these concepts. In the absence of epistasis the expressions for *gca* and *sca* for a polygenic trait consist of the sum, across the involved loci, of the contributions due to individual loci. This requires the presence of linkage equilibrium when dealing with expressions for the variances of *gca* or *sca*. (Section 10.1). The expressions of interest are thus derived from the expressions for locus *B*-*b*, affecting quantitative variation in a trait of an open pollinating population from which pure lines have been extracted. The relevant genotypic compositions are then

		Genotype		
		bb	Bb	BB
f:	In a panmictic population (RM):	q^2	$2pq$	p^2
	In a set of pure lines (L):	q	0	p

The expected genotypic values are

$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}} = m + (p-q)a + 2pqd$$
$$\mathrm{E}\underline{\mathcal{G}}_{\mathrm{L}} = m + (p-q)a$$

A diallel cross yields FS-families. The genotypic composition of the aggregate of all FS-families is equal to the genotypic composition of the panmictic population. Thus $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{FS}} = \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}$.

The genotypic composition of the HS-family obtained from a line with genotype bb, *i.e.* the set of all FS-families obtained from that line, is

	Genotype		
	bb	Bb	BB
f	q	p	0

The genotypic composition of the HS-family obtained from a line with genotype BB is

	Genotype		
	bb	Bb	BB
f	0	q	p

The general combining abilities of genotypes bb and BB may be designated by gca_0 and gca_2, respectively. They are equal to $\mathrm{E}\underline{\mathcal{G}}_{\mathrm{HS}} - \mathrm{E}\underline{\mathcal{G}}_{\mathrm{RM}}$. Thus

$$gca_0 = q(m-a) + p(m+d) - [m + (p-q)a + 2pqd] = pd - pa - 2pqd$$
$$= -p(a - d + 2qd) = -p[a - (1-2q)d] = -p[a - (p-q)d] = -p\alpha$$

It can likewise be shown that

$$gca_2 = q(m+d) + p(m+a) - [m + (p-q)a + 2pqd] = q\alpha$$

Comparison of the above results with Table 8.6 show very simple relations between the above *gca* values and the *bv* values of the (homozygous) genotypes:

$$\underline{gca} = \tfrac{1}{2}\underline{b\nu} = \tfrac{1}{2}(\underline{\gamma} - \mathrm{E}\underline{\mathcal{G}}) \tag{11.55}$$

and

$$\underline{b\nu} = 2\underline{gca}$$

The expected *gca* value, calculated across all homozygous genotypes, is easily obtained from the genotypic composition of the pure lines schema:

	Genotype	
	bb	BB
f	q	p
gca	$-p\alpha$	$q\alpha$

Thus

$$\mathrm{E}\underline{gca} = q(-p\alpha) + p(q\alpha) = 0 \tag{11.56}$$

Furthermore

$$\mathrm{var}(\underline{gca}) = \mathrm{E}(\underline{gca})^2 - [\mathrm{E}(\underline{gca})]^2 = \mathrm{E}(\underline{gca})^2 = qp^2\alpha^2 + pq^2\alpha^2 = pq\alpha^2 = \tfrac{1}{2}{\sigma_a}^2 \tag{11.57}$$

N.B. The results expressed by Equations (11.56) and (11.57) may not be derived, via Equation (11.55), from $\mathrm{E}\underline{b\nu}$ and $\mathrm{var}(\underline{b\nu})$ as the latter quantities apply to panmictic populations. Equation (11.55) would, for instance, yield: $\mathrm{var}(\underline{gca}) = \frac{1}{4}\mathrm{var}(\underline{b\nu}) = \frac{1}{4}{\sigma_a}^2$.

In the scheme below, the margins provide the relative frequencies of the maternal and paternal pure lines involved in the diallel cross (and their genotypes); the central part provides the relative frequencies of the various FS-families resulting from the diallel cross (and their genotypic compositions):

	$q(bb)$	$p(BB)$
$q(bb)$	$q^2(1,0,0)$	$pq(0,1,0)$
$p(BB)$	$pq(0,1,0)$	$p^2(0,0,1)$

The genotypic value of (genetically uniform!) FS-families with genotypic composition (1,0,0) is $m - a = \mathcal{G}_0$. It is $m + d = \mathcal{G}_1$ for FS-families with genotypic composition (0,1,0) and $m + a = \mathcal{G}_2$ for FS-families with genotypic composition (0,0,1).

The specific combining ability of genotypes *bb* and *bb*, of genotypes *bb* and *BB*, and of genotypes *BB* and *BB* are now designated by sca_{00}, sca_{02} and sca_{22}, respectively. According to Equation (11.51), they are equal to

$$sca_{ij} = \mathcal{G}_{ij} - \mu - gca_i - gca_j,$$

i.e. to

$$\underline{\mathcal{G}}_{\mathrm{FS}ij} - \mu - gca_{\mathrm{P}i} - gca_{\mathrm{P}j}$$

According to Equation (8.8) the dominance deviation of a genotype belonging to a panmictic population is equal to the difference between its genotypic value and its additive genotypic value, where the additive genotypic value is equal to $\mu + bv$ (Equation (8.18)). Thus

$$\delta = \mathcal{G} - \gamma = \mathcal{G} - \mu - bv$$

This implies

$$\begin{aligned} sca_{00} &= \mathcal{G}_0 - \mu - 2gca_0 = \mathcal{G}_0 - \mu - bv_0 = \delta_0 \\ sca_{02} &= \mathcal{G}_1 - \mu - \tfrac{1}{2}bv_0 - \tfrac{1}{2}bv_2 = \mathcal{G}_1 - \mu - bv_1 = \delta_1 \\ sca_{22} &= \mathcal{G}_2 - \mu - 2gca_2 = \mathcal{G}_2 - \mu - bv_2 = \delta_2 \end{aligned}$$

The *sca* value of a pair of homozygous genotypes appears thus to be equal to the dominance deviation of the corresponding F_1 genotype. Alternatively, the other way around – the dominance deviation of a genotype is equal to the *sca* value of its homozygous parents.

The variance of the *sca* values of pairs of lines is calculated from the probability distribution of the various pairs of lines and their *sca* values, *i.e.*

	Pair of lines		
	(bb, bb)	(bb, BB)	(BB, BB)
f	q^2	$2pq$	q^2
sca	δ_0	δ_1	δ_2

This means that

$$\mathrm{E}\underline{sca} = \mathrm{E}\underline{\delta} = 0$$

and

$$\mathrm{var}(\underline{sca}) = \mathrm{var}(\underline{\delta}) = {\sigma_d}^2$$

(see Section 8.3.3 and Equation (10.5)). Furthermore Equation (11.51) implies that the variance of the genotypic values of the progenies obtained from the complete diallel cross is equal to

$$\mathrm{var}(\underline{\mathcal{G}}) = \mathrm{var}(\underline{gca}_{\mathrm{M}}) + \mathrm{var}(\underline{gca}_{\mathrm{P}}) + \mathrm{var}(\underline{sca}) = {\sigma_a}^2 + {\sigma_d}^2 \qquad (11.58)$$

where M and P refer to the maternal and paternal lines, respectively.

In conclusion, the quantitative genetic interpretation of the statistical quantities *gca* and *sca* is in terms of breeding values, additive genotypic values and dominance deviations. In the absence of overdominance one may state that the *gca* value of a line will be high if it has, for many loci, the homozygous genotype BB, giving rise to a good performance. Then lines with a good *gca* will tend to have a good performance *per se*. Improvement of *gca* can then simply be pursued by elimination of undesired recessive alleles, *e.g.* by line selection (see Examples 11.22 and 11.23). This means that a diallel cross, made with the single goal to evaluate *gca* values, is a waste. The observation that a cross

between certain inbred line yields an unexpectedly good performing offspring is, nevertheless, of direct significance when developing a SC-hybrid variety.

The *gca* of a pure line and the *sca* of a pair of pure lines depend on the set of pure lines used as a tester. Thus, estimates of *gca* and *sca* derived from a particular diallel cross do not apply to other sets of pure lines. In this sense estimation of *gca* and *sca* is of minor significance. For an incomplete diallel cross one may, however, predict the genotypic value $\mathcal{G}_{ij}$ of any FS-family F_{ij} which was not actually generated, by

$$\hat{\mathcal{G}}_{ij} = \overline{x}_{..} + g\hat{c}a_i + g\hat{c}a_j$$

If the *sca* effects, *i.e.* the dominance deviations, are of minor importance, this approach may save considerable efforts otherwise to be dedicated to crossing and testing. It is speculated that this possibility of predicting progeny performance is insufficiently exploited.

The timing of the estimation of the combining ability of inbred lines deserves attention. In maize breeding it is still current procedure to develop pure lines by selfing for 5-7 generations. Until this stage only some visual selection is applied, but – because it has often been observed that the performances of inbred lines do not predict precisely enough the performance of the SC-hybrid to be obtained from these lines – the selection is useless with regard to the performances of the hybrids to be made. Thereafter the combining abilities of the more or less pure lines are determined.

Effort-saving shortcuts are, of course, attractive. Consequently, it is of interest to check how well the performances of progenies obtained by crossing 'young' inbred lines predict the performances of the hybrids obtained by crossing pure lines tracing back to these young lines. The limits of the potentials of the inbred lines derived from some S_0 plant are *a priori* determined by the genotype of the S_0-plant. Thus a reliable procedure for early assessment of the potentials of lines under development would be of great value. It would allow breeders to devote more efforts to selection among lines from S_0 plants that appeared to be promising.

Jenkins (1935) came to the conclusion that the 'genetic values' of inbred lines, evaluated by testing progenies obtained from **top-crosses**, are determined early in the inbreeding process. This led to the evaluation procedure called **early testing**. It was aimed at the identification of young lines deserving further development. Example 11.24 provides some results.

Example 11.24 Hallauer and Lopez- Perez (1979) studied the reliability of early testing on the basis of 50 S_1 lines and derived S_8 lines. As a yardstick, the coefficient of correlation of the performances of progenies obtained from the S_1 lines and the performances of corresponding progenies obtained from the S_8 lines was used. These coefficients of correlation were estimated when using four different types of testers. This yielded

- $r = 0.17 - 0.20$ with tester I, a genetically heterogeneous population related to the tested lines,
- $r = 0.35$ with tester II, an unrelated inbred line,
- $r = 0.42$ with tester III, an related low yielding inbred line; and
- $r = 0.56$ with tester IV, a related high yielding line.

The rather low coefficients of correlation imply that early testing is not very reliable. In a few cases only three of the top six S_1 lines were related with the top six S_8 lines. The progeny from the S_1 line related to the S_8 line producing the best progeny performed worse than the average calculated across the progenies from all S_1 lines.

As expected, the variation among the progenies was greater when using tester III or IV than when using tester I. Furthermore, the variation among the progenies from the S_8 lines was greater than the variation among the progenies from the S_1 lines. Progenies from the unrelated tester tended to be the best.

One may conclude as follows: an unrelated elite inbred line, which could be used as parent of a hybrid, may be a good tester. Inbred lines having a good specific combining ability with regard to this tester will then be identified. Possibly a hybrid variety may be developed on the basis of test-crosses between the tested lines and this tester.

Chapter 12
Selection for Several Traits

In the preceding chapter only selection with regard to a single trait was considered. One may say that, in practice, selection generally involves several traits. An inexperienced breeder might assume that he is selecting with regard to just a single quantitatively varying trait, for instance biomass yield of maize (Example 11.1), whereas (s)he is, in fact, selecting with regard to a set of mutually correlated traits (see end of Section 11.1). Selection, indeed, is often indirect.

With regard to traits with quantitative variation breeders always apply indirect selection. They select among candidates on the basis of observed phenotypic values, whereas the trait of interest concerns the genotypic values underlying the observed phenotypic values. Recently, indirect selection based on molecular markers has become an important new tool to improve the efficiency of selection with regard to traits with quantitative variation.

The smallest set of mutually correlated traits consists of two traits. The selected trait is the trait as observed under the macro-environmental conditions applying to the population subjected to selection, and the other trait is the same trait but then as expressed under different macro-environmental conditions.

This chapters deals with various aspects related to selection for several traits.

12.1 Introduction

In practice breeders generally select with regard to several traits. These may involve qualitative as well as quantitative variation. Procedures for selection with regard to several traits, **multiple selection**, may be classified according to several criteria. We consider here two criteria for classifying methods of multiple selection:

1. The timing of the multiple selection: successively or simultaneously and
2. The motive to apply multiple selection: unintentional or intentional.

Successive or simultaneous multiple selection

If the selection concerns different traits in the first few generations than in later generations, so-called **tandem selection** is applied. This common approach is applied because initially the number of candidates, each represented by a small number of plants, is very high. Thus in the first generations selection is focussed on:

(i) Traits having a relatively high heritability with the number of plants available per candidate

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 289–323.

(ii) Traits which are reasonably easily assessed

In later generations the number of candidates is considerably smaller. Each candidate may then be represented by such a high number of plants that the heritability is high enough to make the selection efforts rewarding. Example 12.1 specifies for a few crops traits selected in earlier and in later generations.

Example 12.1 In cereal breeding attention is initially focussed on traits like disease resistance or plant habit. With regard to the latter either seedlings with a prostrate or seedlings with an erect growth habit are selected. Thereafter candidates are subjected to selection for grain yield, a trait with a relatively low heritability. In potato breeding selection may start with simultaneous selection for eye depth and colour of the tuber. Later on, and especially in the latest stage, tuber yield is considered.

With **simultaneous selection** several traits are considered in the same generation. This approach is also commonly applied. A specific procedure, called independent-culling-levels selection, is elaborated in Section 12.5.

Unintentional or intentional multiple selection

Unintentional multiple selection may occur even if the breeder intends to select for just one trait. The response to the pursued single-trait selection may then be associated with so-called **correlated responses to selection** with regard to other traits. This is due to associations between the trait considered by the breeder and other traits (see Example 12.2).

Example 12.2 In the long-lasting selection programme of maize described in Example 8.4, the direct selection for either high or low oil or protein content implied unintentional **indirect selection** with regard to many other traits. A correlated response to selection was observed for grain yield, earliness, plant height, tillering, *etc.*

Intentional multiple selection is applied in various ways. **Visual selection** for an abstract trait like 'general impression' or 'breeder's preference' is characteristic for the non-formal way. In Section 12.5 two formal forms of intentional multiple selection are considered:

- **Index selection:** With index selection some index value is assigned to each candidate. This index value indicates the aggregate value of each candidate across several traits. The selection itself consists of truncation selection among the candidates with regard to their index values.
- **Independent-culling-levels selection** (ICL-selection): With truncation selection all plants performing – with regard to some trait – better than a certain minimum phenotypic value are selected (Section 11.1). ICL-selection is an extension of truncation selection. It implies simultaneous application of minimum phenotypic values for several traits.

Unlike the treatment in Chapter 6 of selection for variation determined by a single qualitative locus, it is virtually impossible to describe the process of multiple selection in algebraic expressions. The process differs from crop to crop, for a given crop from stage to stage, and for a given stage from breeder to breeder. It is, in fact, impossible to present a general description of genetic progress. Thus the present chapter deals predominantly with the introduction of two new concepts, *viz.* **genetic correlation** (Section 12.2) and **indirect selection** (Section 12.3).

12.2 The Correlation Between the Phenotypic or Genotypic Values of Traits with Quantitative Variation

A clear linear association of the phenotypic values for trait X and the phenotypic values for trait Y implies a high value for the **phenotypic correlation** $\rho_p(X, Y)$. Indeed, the coefficient of correlation measures the degree of linear relationship between two traits. In fact, the commonly experienced association of phenotypic values for different characters is one of the characteristic features of traits with quantitative traits. This association may be due to

1. A functional relationship
2. Pleiotropy and/or linkage
3. Variation in environmental conditions

A functional relationship between different traits

In Example 8.3 the functional relationship between phenotypic values for grain yield (Y) of cereals and phenotypic values for its components X_1, X_2, X_3 and X_4 was described by:

$$\underline{p}_Y = \underline{p}_{X_1} \cdot \underline{p}_{X_2} \cdot \underline{p}_{X_3} \cdot \underline{p}_{X_4}$$

Such relationship implies an association between, for example, the phenotypic values for traits X_1 and Y. The question may be raised as to whether a complex trait such as Y is directly affected by specific loci or whether its expression is due to loci affecting the components.

Pleiotropy and/or linkage

An allele with **pleiotropic** effects affects the genotypic value of, sometimes, apparently unrelated traits. This phenomenon gives rise to a genetic syndrome. Pleiotropy and linkage are genetic causes for the occurrence of association of phenotypic values for different quantitative traits.

If some plants have a genotype for a pleiotropic locus affecting traits X and Y both in a favourable way and others a genotype affecting both traits in an unfavourable way, then the genotypic values for X and Y will be positively correlated.

In the case of linkage disequilibrium, the probability distribution of the genotypes for locus B_1-b_1 affecting trait T_1 and the probability distribution of the genotypes for locus B_2-b_2 affecting trait T_2 are not independent. This implies correlation of the genotypic values for traits T_1 and T_2 (in as far as affected only by these loci). In the presence of linkage equilibrium with regard to these loci, there will be no genotypic correlation, unless the involved loci have pleiotropic effects with regard to the considered traits.

Example 12.3 considers these two causes for traits to be associated.

Example 12.3 In Fig. 12.1 locus *B*-*b* has pleiotropic effects with regard to traits X_1 and X_2. Locus *H*-*h* is pleiotropic with regard to traits X_1 and X_3 and loci *D*-*d* and *G*-*g* are pleiotropic with regard to traits X_2 and X_3. These pleiotropic effects induce phenotypic correlation of traits X_1 and X_2, X_1 and X_3 and X_2 and X_3. Trait X_4 is controlled by the non-pleiotropic loci *I*-*i*, *J*-*j* and *K*-*k*.

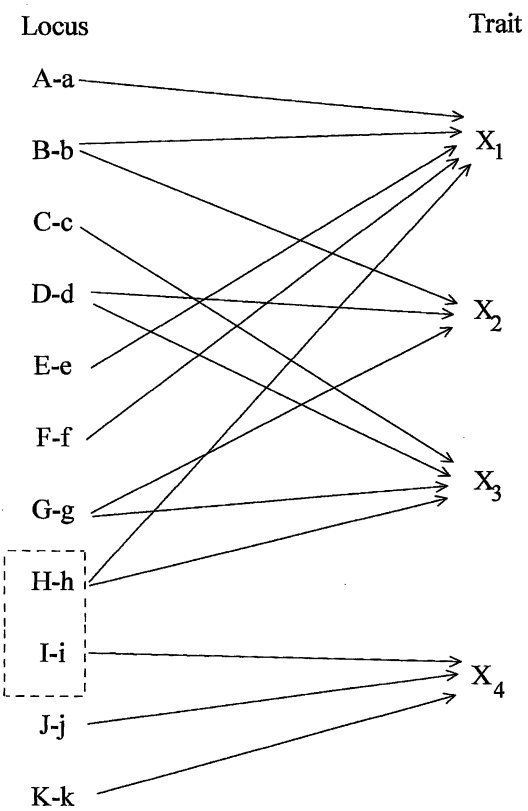


Fig. 12.1 The genetic control of the quantitative traits X_1, X_2, X_3 and X_4 by the loci *A*-*a*, ..., *K*-*k*. The dashed box encloses linked loci

Variation in environmental conditions

Variation in the quality of growing conditions induces correlation of the phenotypic values for different traits. Such variation induces covariance of the environmental deviations: certain plants grow under favourable conditions for traits X and Y and others under unfavourable conditions.

In genetically homogeneous plant material the coefficient of phenotypic correlation between traits X and Y has a special interpretation. The correlation of $\underline{p}_{\mathrm{X}} = \mathcal{G}_{\mathrm{X}} + \underline{e}_{\mathrm{X}}$ and $\underline{p}_{\mathrm{Y}} = \mathcal{G}_{\mathrm{Y}} + \underline{e}_{\mathrm{Y}}$ is then equal to the correlation of the environmental deviations:

$$\rho_p = \frac{cov(\underline{p}_{\mathrm{X}}, \underline{p}_{\mathrm{Y}})}{\sigma_{p_{\mathrm{X}}} \cdot \sigma_{p_{\mathrm{Y}}}} = \frac{cov(\underline{e}_{\mathrm{X}}, \underline{e}_{\mathrm{Y}})}{\sigma_{e_{\mathrm{X}}} \cdot \sigma_{e_{\mathrm{Y}}}} = \rho_e$$

The parameter ρ_e is called the **environmental correlation**. Example 12.4 describes an interesting cause for environmental correlation, namely interplant competition.

Example 12.4 In a genetically uniform variety of a cereal crop, the coefficient of correlation of grain yield and plant height of separate plants tends to be positive. This might be due to variation in seed size. Some plants originate from large kernels giving rise to early emergence and/or large seedlings. These plants tend to have a higher grain yield and to be taller than plants originating from small seeds. This cause for a positive correlation applies especially in the presence of interplant competition, *i.e.* at high plant density. However, whatever the plant density may be, variation in soil fertility will always induce a positive correlation: tall and high-yielding plants will develop at good positions, whereas short and low-yielding plants will occur at poor positions.

The relationship between $\rho_p(\mathrm{X}, \mathrm{Y})$, the **genetic correlation** $\rho_g(\mathrm{X}, \mathrm{Y})$ and the **environmental correlation** $\rho_e(\mathrm{X}, \mathrm{Y})$ will now be derived. In statistics ρ, the coefficient of correlation of the random variables $\underline{x}$ and $\underline{y}$, is defined as

$$\rho := \frac{cov(\underline{x}, \underline{y})}{\sigma_x \cdot \sigma_y}$$

Thus

$$\mathrm{cov}(\underline{x}, \underline{y}) = \rho \sigma_x \sigma_y$$

This is applied to an elaborated expression for ρ_p:

$$\rho_p(\mathrm{X}, \mathrm{Y}) = \frac{cov(\underline{p}_{\mathrm{X}}, \underline{p}_{\mathrm{Y}})}{\sigma_{p_{\mathrm{X}}} \cdot \sigma_{p_{\mathrm{Y}}}} = \frac{cov(\underline{\mathcal{G}}_{\mathrm{X}} + \underline{e}_{\mathrm{X}}, \underline{\mathcal{G}}_{\mathrm{Y}} + \underline{e}_{\mathrm{Y}})}{\sigma_{p_{\mathrm{X}}} \cdot \sigma_{p_{\mathrm{Y}}}}$$

If, due to randomization, the covariance of the genotypic value and the environmental deviation is zero, $\rho_p(\mathrm{X}, \mathrm{Y})$ is equal to

$$\frac{cov(\underline{\mathcal{G}}_{\mathrm{X}}, \underline{\mathcal{G}}_{\mathrm{Y}}) + cov(\underline{e}_{\mathrm{X}}, \underline{e}_{\mathrm{Y}})}{\sigma_{p_{\mathrm{X}}} \cdot \sigma_{p_{\mathrm{Y}}}}$$

This is rewritten into

$$\frac{\rho_g \sigma_{g_X} \sigma_{g_Y} + \rho_e \sigma_{e_X} \sigma_{e_Y}}{\sigma_{p_X} \cdot \sigma_{p_Y}} = \rho_g h_X h_Y + \rho_e e_X e_Y \tag{12.1}$$

where

$$h = \frac{\sigma_g}{\sigma_p}$$

and

$$e = \frac{\sigma_e}{\sigma_p}$$

Thus

$$e^2 = \frac{{\sigma_e}^2}{{\sigma_p}^2} = \frac{{\sigma_p}^2 - {\sigma_g}^2}{{\sigma_p}^2} = 1 - h^2$$

and

$$e = \sqrt{1 - h^2}$$

(see also Equation (11.24)). If $h_X = h_Y = 0$, *i.e.* $e_X = e_Y = 1$, Equation (12.1) yields $\rho_p = \rho_e$. Thus, as shown before, the coefficient of phenotypic correlation occurring in genetically uniform plant material is to be interpreted as the coefficient of environmental correlation.

The environmental variance for some trait may differ from genotype to genotype (Example 8.9). Likewise, the environmental correlation of two traits may vary across genotypes.

The phenotypic correlation in a genetically heterogeneous population depends on both the genetic and the environmental correlation. These may have very different values, even values of opposite signs.

Estimation of ρ_p, ρ_g or ρ_e may require considerable effort. In Section 12.4 several procedures for obtaining estimates, designated by r_p, r_g and r_e, respectively, are elaborated.

12.3 Indirect Selection

In the case of genetic correlation between traits X and Y, the mean phenotypic value with regard to trait Y of the candidates selected for trait X will differ from the mean phenotypic value of all candidates. The difference is called **correlated selection differential** (see Equation (11.4)). The selection for trait X will thus not only yield a selection response with regard to trait X itself but, due to the correlated selection differential, also a **correlated response** (CR) with regard to trait Y. The response to such **indirect selection** is the topic of the present section. It will be compared to the response to direct selection for Y.

Indirect selection is in fact always applied as the selection for some trait involves phenotypic values, whereas the target of the selection is improvement with regard to genotypic values. Application of indirect selection is thus unavoidable.

When applied deliberately, indirect selection may be defined as selection with regard to some trait X with the target to attain some selection response with regard to trait Y. Trait X serves then as the so-called auxiliary trait; trait Y is the target trait, often yield. To be able to compare the response to indirect selection with the response to direct selection the concept of **relative selection efficiency** has been developed (Section 12.3.1).

Indirect selection may be applied deliberately. A specific application is index selection (Section 12.5). It may also be applied because of economic reasons, especially the saving of time. Three examples are given:

1. A breeder might select among inoculated seedlings in order to improve adult plant resistance.
2. Woody crops, such as coffee or oil palm, have a long lasting juvenile phase. Yield is only expressed after a number of years. Selection among juvenile plants with regard to juvenile plant traits related to yield, may then be considered. Thus juvenile girth width at breast height may indicate adult plant production.
3. The breeder might select among seedlings on the basis of observation of markers predicting adult plant performance. This is specifically pursued when applying marker-assisted selection (Section 12.3.2). Such selection may be applied, not just because of saving time but also because of its high relative selection efficiency.

Indirect selection is also applied when the selection occurs under conditions deviating from the conditions provided in plant production practice (Section 12.3.3).

12.3.1 Relative selection efficiency

Equation (11.13) indicates how the response to selection for trait X, say R_{X}, to be expected at a certain selection differential with regard to this trait, say S_{X}, can be predicted, *viz.*

$$R_{\mathrm{X}} = \beta S_{\mathrm{X}},$$

where the quantitative genetic meaning of β depends on the situation. In the case of selecting candidates with identical reproduction β is equal to the heritability of X in the wide sense $\left({h_w}^2\right)$, in the case of selection of candidates belonging to a cross-fertilizing crop (non-identical reproduction) β is equal to the heritability of X in the narrow sense $\left({h_n}^2\right)$.

We now consider, both for the case of identical reproduction of the selected candidates and for the case of non-identical reproduction by means of cross-fertilization of the selected candidates:

1. The correlated response, with regard to trait Y, say CR_Y, to be expected at a selection differential, amounting to S_{X}, with regard to trait X. Analogous

to Equation (11.13) we write

$$CR_{\mathrm{Y}} = \beta' S_{\mathrm{X}}, \tag{12.2}$$

The quantitative meaning of β' is derived for both situations.

2. The ratio

$$\frac{CR_{\mathrm{Y}}}{R_{\mathrm{Y}}} \tag{12.3}$$

This ratio is called **relative selection efficiency** (RSE). If $RSE > 1$ one may consider application of indirect selection for Y instead of direct selection. The selection is then for the **auxiliary trait** X in order to improve **target trait** Y. Indirect selection may thus be applied because it offers better prospects than direct selection.

Identical reproduction of the selected candidates

At identical reproduction of the selected candidates the quantitative genetic meaning of β' is

$$\beta' = \frac{cov(\underline{\mathcal{G}}_{\mathrm{Y}}, \underline{p}_{\mathrm{X}})}{var(\underline{p}_{\mathrm{X}})} = \frac{cov(\underline{\mathcal{G}}_{\mathrm{Y}}, \underline{\mathcal{G}}_{\mathrm{X}})}{var(\underline{p}_{\mathrm{X}})} = \frac{cov(\underline{\mathcal{G}}_{\mathrm{Y}}, \underline{\mathcal{G}}_{\mathrm{X}})}{\sigma_{g_{\mathrm{X}}} \cdot \sigma_{g_{\mathrm{Y}}}} \cdot \frac{\sigma_{g_{\mathrm{X}}}}{\sigma_{p_{\mathrm{X}}}} \cdot \frac{\sigma_{g_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}} = \rho_g \cdot h_{w_{\mathrm{X}}} \cdot \frac{\sigma_{g_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}}$$

This yields

$$CR_Y = \rho_g \cdot h_{w_{\mathrm{X}}} \cdot \frac{\sigma_{g_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}} \cdot S_{\mathrm{X}} = i_{\mathrm{X}} \cdot \rho_g \cdot h_{w_{\mathrm{X}}} \cdot \frac{\sigma_{g_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}} \cdot \sigma_{p_{\mathrm{X}}} = i_{\mathrm{X}} \rho_g h_{w_{\mathrm{X}}} \sigma_{g_{\mathrm{Y}}} \tag{12.4}$$

The relative selection efficiency is thus

$$RSE = \frac{i_{\mathrm{X}} \rho_g h_{w_{\mathrm{X}}} \sigma_{g_{\mathrm{Y}}}}{i_{\mathrm{Y}} h_{w_{\mathrm{Y}}} \sigma_{g_{\mathrm{Y}}}} = \frac{i_{\mathrm{X}}}{i_{\mathrm{Y}}} \cdot \rho_g \cdot \frac{h_{w_{\mathrm{X}}}}{h_{w_{\mathrm{Y}}}} \tag{12.5}$$

Cross-fertilization of the selected candidates

At cross-fertilization of the selected candidates the quantitative genetic meaning of β' is

$$\beta' = \frac{cov(\underline{\gamma}_{\mathrm{Y}}, \underline{p}_{\mathrm{X}})}{var(\underline{p}_{\mathrm{X}})} = \frac{cov(\underline{\gamma}_{\mathrm{Y}}, \underline{\gamma}_{\mathrm{X}})}{var(\underline{p}_{\mathrm{X}})} = \frac{cov(\underline{\gamma}_{\mathrm{Y}}, \underline{\gamma}_{\mathrm{X}})}{\sigma_{a_{\mathrm{Y}}} \cdot \sigma_{a_{\mathrm{X}}}} \cdot \frac{\sigma_{a_{\mathrm{X}}}}{\sigma_{p_{\mathrm{X}}}} \cdot \frac{\sigma_{a_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}} = \rho_a \cdot h_{n_{\mathrm{X}}} \cdot \frac{\sigma_{a_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}}$$

where $\underline{\gamma}$ represents the additive genotypic value (Equation (8.6)) and where $\rho_a(\mathrm{X}, \overline{\mathrm{Y}})$ is the so-called **additive genetic correlation** of traits X and Y. This parameter can be related to a parameter called coheritability of traits X and Y, see Note 12.1.

Note 12.1 We define now a parameter, called **co-heritability in the wide sense of traits X and Y** ($coh^2{}_{\mathrm{w}}(\mathrm{X},\mathrm{Y})$), for the case of identical reproduction, *viz.*

$$coh_w{}^2(\mathrm{X},\mathrm{Y}) := \frac{cov(\underline{g}_{\mathrm{Y}},\underline{g}_{\mathrm{X}})}{\sigma_{p_{\mathrm{X}}}\cdot\sigma_{p_{\mathrm{Y}}}} = \frac{cov_g(\mathrm{X},\mathrm{Y})}{\sigma_{p_{\mathrm{X}}}\cdot\sigma_{p_{\mathrm{Y}}}},$$

as well as a parameter, called **co-heritability in the narrow sense of traits X and Y** ($coh^2{}_{\mathrm{n}}(\mathrm{X},\mathrm{Y})$), for the case of the non-identical reproduction occurring in a cross-fertilizing crop, *viz.*

$$coh_n{}^2(\mathrm{X},\mathrm{Y}) := \frac{cov(\underline{\gamma}_{\mathrm{Y}},\underline{\gamma}_{\mathrm{X}})}{\sigma_{p_{\mathrm{X}}}\cdot\sigma_{p_{\mathrm{Y}}}} = \frac{cov_a(\mathrm{X},\mathrm{Y})}{\sigma_{p_{\mathrm{X}}}\cdot\sigma_{p_{\mathrm{Y}}}} \tag{12.6}$$

Thus

$$cov(\mathrm{X},\mathrm{Y}) = coh^2(\mathrm{X},\mathrm{Y})\cdot\sigma_{p_{\mathrm{X}}}\cdot\sigma_{p_{\mathrm{Y}}}$$

As

$$cov(\mathrm{X},\mathrm{Y}) = \rho(\mathrm{X},\mathrm{Y})\cdot\sigma_{\mathrm{X}}\cdot\sigma_{\mathrm{Y}}$$

the above definitions imply

$$coh_w^2(\mathrm{X},\mathrm{Y}) = \rho_g(\mathrm{X},\mathrm{Y})\cdot h_{w_{\mathrm{X}}}\cdot h_{w_{\mathrm{Y}}} \tag{12.7a}$$

and

$$coh_n^2(\mathrm{X},\mathrm{Y}) = \rho_a(\mathrm{X},\mathrm{Y})\cdot h_{n_{\mathrm{X}}}\cdot h_{n_{\mathrm{Y}}}. \tag{12.7b}$$

respectively.

The correlated response to selection amounts thus to

$$CR_{\mathrm{Y}} = i_{\mathrm{X}}\cdot\rho_a\cdot h_{n_{\mathrm{X}}}\cdot\frac{\sigma_{a_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{X}}}}\cdot\sigma_{p_{\mathrm{X}}} = i_{\mathrm{X}}\rho_a h_{n_{\mathrm{X}}}\sigma_{a_{\mathrm{Y}}} \tag{12.8}$$

The relative selection efficiency is thus

$$RSE = \frac{i_{\mathrm{X}}\rho_a h_{n_{\mathrm{X}}}\sigma_{a_{\mathrm{Y}}}}{i_{\mathrm{Y}}h_{n_{\mathrm{Y}}}\sigma_{a_{\mathrm{Y}}}} = \frac{i_{\mathrm{X}}}{i_{\mathrm{Y}}}\cdot\rho_a\cdot\frac{h_{n_{\mathrm{X}}}}{h_{n_{\mathrm{Y}}}} \tag{12.9}$$

Equation (12.9) resembles Equation (12.5) very closely.

The conditions yielding $RSE > 1$ are

1. $\rho_g > \frac{h_{\mathrm{Y}}}{h_{\mathrm{X}}}$ at $i_{\mathrm{X}} \approx i_{\mathrm{Y}}$
   This condition applies with a strong genetic correlation of traits X and Y and when $h_{\mathrm{X}}{}^2 \gg h_{\mathrm{Y}}{}^2$, *i.e.* when the target trait has a very low heritability compared to the heritability of the auxiliary trait.
2. $i_{\mathrm{X}} > i_{\mathrm{Y}}$ at $\rho_g \approx \frac{h_{\mathrm{Y}}}{h_{\mathrm{X}}}$

This condition may apply when dealing with a dioecious crop. The auxiliary trait X may be expressed by both male and female plants, whereas the target trait Y is only expressed by female plants, *e.g.* seed or fruit yield (see Example 12.5).

Example 12.5 Breure (1986) considered improvement of oil palm yield per ha by selecting palms with a high bunch index (BI), *i.e.* the proportion of the above-ground dry matter per palm used for fruit bunches (Y). In fact he considered indirect selection for Y. It appeared that the heritability of both BI and Y was quite low in the material tested. An additional problem is that *pisifera* palms, *i.e.* the male parents of the presently cultivated *tenera* palms, can not be selected for BI and/or Y as they are mostly female sterile. *Pisifera* selection concerns therefore general impression based on visual observations. Other selection criteria are therefore desired. Breure studied a few potential auxiliary traits:

- Magnesium content of the leaves of *pisifera* palms. In magnesium deficient areas the Leaf Magnesium status (LMG) was found to be positively correlated with yield, whereas it also has a high heritability.
- Sex ratio (SR), *i.e.* the ratio of the number of female inflorescences to the total number.
- Leaf are ration (LAR), *i.e.* the ratio of new leaf are produced to new dry matter used for vegetative growth.

Breure applied multiple linear egression of data for Y, as observed for *tenera* palms on parental data for LMG, SR and LAR. He found that 80% of the variance for Y in the offspring was exclusively accounted for by LMG of both parents, with LMG of *pisifera* being most important (66% of the variance explained). The use of LMG values of effectively male *pisifera* palms looked thus promising for indirect selection.

In the case of dioecy we have

$$i_{\mathrm{X}} = \tfrac{1}{2}(i_{m_{\mathrm{X}}} + i_{f_{\mathrm{X}}})$$

and, because $i_{m_{\mathrm{Y}}} = 0$:

$$i_{\mathrm{Y}} = \tfrac{1}{2}(i_{m_{\mathrm{Y}}} + i_{f_{\mathrm{Y}}}) = \tfrac{1}{2} i_{f_{\mathrm{Y}}}$$

Example 12.6 gives, for a dioecious crop, a theoretical illustration of a situation with $i_{\mathrm{X}} > i_{\mathrm{Y}}$.

Example 12.6 We consider a population of a dioecious crop consisting of 500 male and 500 female plants. Trait Y is the target trait which is expressed by female plants after pollen distribution; X is an auxiliary trait which is expressed by all plants before pollen distribution. One may select 50 plants with regard to trait Y. These plants, *i.e.* 10% of the female plants, have already been pollinated in the absence of selection among the male plants. According to Falconer (1989; Appendix Table A) this implies $i_{\mathrm{Y}} = \frac{1}{2} i_{f_{\mathrm{Y}}} = \frac{1}{2}(1.755) = 0.8775$. Selection of 50 plants with regard to X, *i.e.* 5%, implies $i_{\mathrm{X}} = 2.063$. In this situation $\frac{i_{\mathrm{X}}}{i_{\mathrm{Y}}} = 2.35$, which may imply that RSE > 1.

The situation $RSE > 1$ may of course especially occur if both of the former conditions apply. Example 12.7 summarizes some practical results of application of indirect selection.

Example 12.7 For five seasons Lonnquist (1967) applied indirect selection with regard to grain yield by selecting for prolificacy in the open-pollinating maize variety Hays Golden. In each season a selection field comprising 4000 to 5000 plants was grown. The plant density was only 2 plants per m^2. This promotes the expression of prolificacy. From each of the circa 200 selected prolific plants, *i.e.* about 5%, one ear was harvested. The result of each selection cycle was established by means of a yield trial with at least 10 replicates and including the original variety as a check. Each yield trial lasted 3 years and was grown at a plant density of 3.45 plants/m^2.

Regression of the relative yield, *i.e.* the grain yield expressed as percentage of the grain yield of Hays Golden, on the rank of the selection cycle showed a progress of 6.3% per cycle. The progress due to direct selection of 10% of the plants, measured in the same way, was 3.8% per cycle. (This favourable result of indirect selection may have been due to the higher selection intensity as well as to the low plant density applied in the yield trial).

In oat indirect selection for grain yield via selection for harvest index, *i.e.* grain yield/biomass, was 43% as effective as direct selection (Rosielle and Frey, 1975). However, indirect selection was expected to retain lines with a more favourable combination of yield, plant height and heading date than the lines expected to be retained with direct selection for yield.

Indirect selection may even be attractive if $RSE < 1$. It may be applied to save time and/or effort. Time is saved if selection for a trait, expressed in an early ontogenetic phase, is applied in order to get improvement with regard to an adult plant trait. In resistance breeding this form of indirect selection is common practice. In many cases it has been established that seedling resistance and adult plant resistance are strongly correlated. Barley seedlings may, for instance, be selected for partial resistance to barley leaf rust (*Puccinia hordei*) in order to improve the resistance of adult plants.

Especially for crops with a long-lasting juvenile phase, breeders are interested in juvenile plant traits correlated with the target trait(s) expressed by adult plants. For woody crops, such as apple, coffee or oil palm, often the girth width of the stem at breast height is used as an auxiliary trait. Effort is saved if the auxiliary trait is easier to assess than the target trait.

12.3.2 The use of markers

One may generalize that direct selection tends to be inefficient with regard to traits with quantitative variation. Chapter 17 summarizes causes for this challenging situation. As a way-out breeders may consider indirect selection

by selecting for marker phenotypes. Such selection is, of course, only of interest if it gives rise to a rewarding correlated response with regard to the target trait.

A **marker** with regard to some quantitative trait is a trait such that different phenotypic values/classes of the marker trait are associated with different mean phenotypic values of the quantitative trait of interest. In the present context markers are auxiliary traits used for indirect selection with regard to a target trait. The association requires linkage between the locus (or the loci) controlling the marker and the locus (or the loci) affecting the target trait. (For random mating populations even the more demanding condition of linkage disequilibrium is required). The probability distribution for the genotypes for the locus controlling the marker and the probability distribution for any locus affecting the target trait should thus be interdependent. Only in that case a (positive or a negative) covariance, *i.e.* an association, between marker and target trait may occur (Section 10.1).

The marker may be a plant trait that is visually observed, for instance flower colour. It may also be the product of a genotype for a certain locus, for instance a polypeptide or a protein. An important category of markers are the so-called **molecular markers**. In this case the marker is neither a plant trait nor a gene product; the marker consists of (cloned parts of) the DNA itself. The presence or the absence of a certain band in the lane obtained by gel electrophoresis involving some genotype characterizes the studied entry.

With the aid of molecular marker techniques it has become possible to identify individual loci affecting quantitative traits (Stam, 1998). This greatly improves the understanding of the genetic control of quantitative traits. It permits the assessment of the degree to which related traits are controlled by the same or by distinct loci. (Thus a locus affecting kernel size may or may not coincide with a locus affecting grain yield.) Or it may appear, when growing a certain population in a range of environments, that some of the loci affecting a trait are expressed in all environments, whereas other loci are only expressed in specific conditions. The latter loci are responsible for genotype $\times$ environment interaction (Manneh, 2004).

If polymorphic, a molecular marker reflects small differences in the DNA sequence that are observed as the presence or the absence of a band at a certain position in the lane. This implies that molecular markers have a heritability which is equal to one: the presence or the absence of the band is completely determined by the genotype. A further advantage is that the marker phenotypes (or genotypes; $h^2 = 1$!) can already be determined from DNA extracted from seedlings. It is tempting to assume that the relative efficiency of so-called **marker-assisted selection**, often indicated as MAS, tends to be larger than one: $RSE > 1$.

It was already emphasized that a polymorphism, appearing when a set of genotypes segregates with regard to the presence or the absence of a band at a certain position in a gel alongside the lanes, can only be used as a marker if the genotypes where the band is present have a higher or a lower mean phenotypic

value for one or more target traits than the genotypes where the band is absent. This requires that the involved population is in linkage disequilibrium. For the sake of illustration such associations are here only elaborated for an F_2 population, as well as for sets of pure lines obtained in the absence of selection, either by some procedure to generate doubled haploids (DH) or by continued selfing (F_∞). Weber and Wricke (1994) consider associations occurring in some other populations: F_3 populations, backcross families, backcrosses selfed, F_1 top cross.

Let locus X-x designate the locus controlling variation in a marker, *i.e.* variation with regard to the auxiliary trait X, and locus Y-y, a locus affecting variation with regard to the target trait Y. Locus Y-y is often called a quantitative trait locus (QTL). These two loci are linked with recombination value r, where $0 < r \leq \frac{1}{2}$. The genotypic compositions of the considered populations, as obtained from the initial cross $xxyy \times XXYY$, are derived from Tables 2.2 and 3.2:

	Genotype								
	$xxyy$	$xxYy$	$xxYY$	$Xxyy$	$XxYy$	$XxYY$	$Xxyy$	$XXYy$	$XXYY$
$\mathcal{G}-m$	$-a$	d	a	$-a$	d	a	$-a$	d	a
$f:\mathrm{F}_2$	$\frac{1}{4}(1-r)^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{2}(1-r)^2 + \frac{1}{2}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}r^2$	$\frac{1}{2}r(1-r)$	$\frac{1}{4}(1-r)^2$
DH	$\frac{1}{2}(1-r)$	0	$\frac{1}{2}r$	0	0	0	$\frac{1}{2}r$	0	$\frac{1}{2}(1-r)$
F_∞	$\frac{1}{2(1+2r)}$	0	$\frac{2r}{2(1+2r)}$	0	0	0	$\frac{2r}{2(1+2r)}$	0	$\frac{1}{2(1+2r)}$

The plants/lines are classified according to their genotype for locus X-x and the expected genotypic value with regard to trait Y is determined for each class. Association, *i.e.* different classes have different (conditional) expected genotypic values, will be shown to be present if locus X-x is linked with locus Y-y, *i.e.* if $r < 1/2$.

F_2 population

The probability that an F_2 plant belongs to marker class xx is $\frac{1}{4}$. The (conditional) expected genotypic value of such plants amounts to:

$$\begin{aligned}\mathrm{E}(\underline{\mathcal{G}}|xx) &= (1-r)^2(m-a) + 2r(1-r)(m+d) + r^2(m+a)\\ &= m - a[(1-r)^2 + r^2] + 2r(1-r)d\\ &= m - (1-2r)a + 2r(1-r)d\end{aligned}$$

Likewise

$$\mathrm{E}(\underline{\mathcal{G}}|Xx) = m + (1 - 2r + 2r^2)d$$

and

$$\mathrm{E}(\underline{\mathcal{G}}|XX) = m + (1-2r)a + 2r(1-r)d$$

The (conditional) expected genotypic values of the three marker classes are equal if loci X-x and Y-y are unlinked, *i.e.* if $r = \frac{1}{2}$:

$$\mathrm{E}(\underline{\mathcal{G}}|xx) = \mathrm{E}(\underline{\mathcal{G}}|Xx) = \mathrm{E}(\underline{\mathcal{G}}|XX) = m + \tfrac{1}{2}d$$

They are different if loci X-x and Y-y are linked, *i.e.* $r < \frac{1}{2}$. For genotypes XX and xx the expected difference is

$$\mathrm{E}(\underline{\mathcal{G}}|XX) - \mathrm{E}(\underline{\mathcal{G}}|xx) = 2(1-2r)a = (1-2r)(\mathcal{G}_{YY} - \mathcal{G}_{yy}) \qquad (12.10)$$

Example 12.8 shows for an F_2 population how different marker genotypes give rise to different expected genotypic values with regard to trait Y because of linkage between the marker locus X-x and some locus Y-y affecting trait Y.

Example 12.8 An F_2 population segregates for locus Y-y, affecting a quantitative trait (with $m = 80$, a $= 20$ and $d = 0$), as well as for locus X-x, controlling a marker. In the homozygous parental genotypes these loci were linked (with recombination value $r = 0.2$) in coupling phase. According to Table 2.2 the genotypic composition of the F_2 is:

	Genotype								
	xxyy	*xxYy*	*xxYY*	*Xxyy*	*XxYy*	*XxYY*	*Xxyy*	*XXYy*	*XXYY*
f	0.16	0.08	0.01	0.08	0.34	0.08	0.01	0.08	0.16
$\mathcal{G}$	60	80	100	60	80	100	60	80	100

Thus:

$$\begin{aligned}\mathrm{E}(\underline{\mathcal{G}}|xx) &= 4(0.16 \times 60 + 0.08 \times 80 + 0.01 \times 100) = 68\\ \mathrm{E}(\underline{\mathcal{G}}|Xx) &= 2(0.08 \times 60 + 0.34 \times 80 + 0.08 \times 100) = 80\end{aligned}$$

and

$$\mathrm{E}(\underline{\mathcal{G}}|XX) = 4(0.01 \times 60 + 0.08 \times 80 + 0.16 \times 100) = 92$$

It can easily be verified that these conditional expected genotypic values are equal to $m - (1-2r)a$, m, and $m + (1-2r)a$, respectively. The difference between the expected genotypic value of plants in marker class XX and plants in marker class xx is equal to $92 - 68 = 24$, *i.e.* to $2(1-2r)a$.

DH lines

Among DH lines of marker class xx the expected genotypic value is

$$\mathrm{E}(\underline{\mathcal{G}}|xx) = m + (1-r)(-a) + r(a) = m + (1-2r)(-a)$$

and likewise

$$\mathrm{E}(\underline{\mathcal{G}}|XX) = m + r(-a) + (1-r)a = m + (1-2r)a$$

Thus

$$\mathrm{E}(\underline{\mathcal{G}}|XX) - \mathrm{E}(\underline{\mathcal{G}}|xx) = 2(1-2r)a = (1-2r)(\mathcal{G}_{YY} - \mathcal{G}_{yy}) \qquad (12.11)$$

F_∞ lines

For F_∞ lines it can be derived that

$$E(\underline{\mathcal{G}}|xx) = m - \frac{(1-2r)a}{1+2r}$$

and

$$E(\underline{\mathcal{G}}|XX) = m + \frac{(1-2r)a}{1+2r}$$

This implies that

$$E(\underline{\mathcal{G}}|XX) - E(\underline{\mathcal{G}}|xx) = \frac{2(1-2r)a}{1+2r} = \frac{1-2r}{1+2r}(\mathcal{G}_{YY} - \mathcal{G}_{yy}) \qquad (12.12)$$

For any marker the expected contrast between the genotypic values of classes xx and XX as obtained for DH lines is equal to the expected contrast as obtained for F_2 plants. This contrast is expected to be larger than the corresponding contrast for F_∞ lines. However, when comparing a set of DH lines with a set of F_∞ lines it depends on the marker, *i.e.* on r, which set of lines gives rise to the larger contrast between the considered marker classes.

Linkage, *i.e.* $0 < r < \frac{1}{2}$, is shown to be present if the mean phenotypic values of plants representing different marker classes differ significantly. Equations (12.10) to (12.12) show that both r and a (or $\mathcal{G}_{YY} - \mathcal{G}_{yy}$) affect the size of the difference between marker classes XX and xx.

Knowledge about linkage between a marker and a QTL requires that a marker linkage map is available. Such a map is constructed by studying the co-segregation of pairs of markers in the offspring generation(s) obtained after crossing two genotypes. The estimated recombination values serve as a basis to assign each marker to a linkage group and to determine its best-fitting position within the group. Computer programs have been developed to assist with the determination of the best-fitting position among other markers within the group; see e.g. Stam and Van Ooijen (1995).

The position on the linkage map assigned to a QTL affecting the considered quantitative trait depends on the degree of association between genotypes of markers closely linked to the QTL with trait values. By scanning the markers alongside an ordered map for their association with the trait values a likely map position is assigned to each QTL (Van Ooijen and Maliepaard, 1995). Simultaneously the effects of the genes at the QTL are estimated. Indeed, the contrasts like those specified by Equations (12.10) to (12.12) depend both on the parameters a and d for locus Y-y and on r, the recombination value of the marker locus and the involved QTL. In Note 12.2 it is shown how one may obtain separate estimates for both the position of a QTL and its genetic effect.

Note 12.2 Separate estimation of r and a or d is possible by considering two linked marker loci X_1-x_1 and X_2-x_2, with known recombination value r, which embrace locus Y-y. The recombination value of loci X_1-x_1 and Y-y

is designated by r_1 and the recombination value of loci X_2-x_2 and Y-y is designated by r_2. Here only the situation of absence of chiasma interference (Section 2.2.4) is elaborated; thus: $r = r_1 + r_2 - 2r_1r_2$.

The determination of the position of locus Y-y relative to the positions of the flanking marker loci is called **interval mapping**. The procedure is illustrated for DH lines as obtained from the initial cross $x_1x_1yyx_2x_2 \times X_1X_1YYX_2X_2$. The genotypic composition of the set of DH lines follows from the haplotypic composition of the gametes produced by the F_1:

Genotype	f	$\mathcal{G}$
$x_1x_1YYx_2x_2$	$\frac{1}{2}r_1r_2$	$m+a$
$x_1x_1yyx_2x_2$	$\frac{1}{2}(1-r_1)(1-r_2)$	$m-a$
$X_1X_1YYx_2x_2$	$\frac{1}{2}(1-r_1)r_2$	$m+a$
$X_1X_1yyx_2x_2$	$\frac{1}{2}r_1(1-r_2)$	$m-a$
$x_1x_1YYX_2X_2$	$\frac{1}{2}r_1(1-r_2)$	$m+a$
$x_1x_1yyX_2X_2$	$\frac{1}{2}(1-r_1)r_2$	$m-a$
$X_1X_1YYX_2X_2$	$\frac{1}{2}(1-r_1)(1-r_2)$	$m+a$
$X_1X_1yyX_2X_2$	$\frac{1}{2}r_1r_2$	$m-a$

The above genotypes have been ordered according to their (homozygous) marker genotypes. The frequencies of the marker genotypes are:

Genotype	f
$x_1x_1x_2x_2$	$\frac{1}{2}r_1r_2 + \frac{1}{2}(1-r_1)(1-r_2) = \frac{1}{2}[1-(r_1+r_2-2r_1r_2)] = \frac{1}{2}(1-r)$
$X_1X_1x_2x_2$	$\frac{1}{2}(1-r_1)r_2 + \frac{1}{2}r_1(1-r_2) = \frac{1}{2}(r_1+r_2-2r_1r_2) = \frac{1}{2}r$
$x_1x_1X_2X_2$	$\frac{1}{2}r_1(1-r_2) + \frac{1}{2}(1-r_1)r_2 = \frac{1}{2}r$
$X_1X_1X_2X_2$	$\frac{1}{2}(1-r_1)(1-r_2) + \frac{1}{2}r_1r_2 = \frac{1}{2}(1-r)$

The conditional expected genotypic values of each marker class amount then to:

$$\mathrm{E}(\underline{G}|x_1x_1x_2x_2) = m + \frac{[r_1r_2 - (1-r_1)(1-r_2)]a}{1-r} = m - \frac{(1-r_1-r_2)a}{1-r}$$

$$\mathrm{E}(\underline{G}|X_1X_1x_2x_2) = m + \frac{[(1-r_1)r_2 - r_1(1-r_2)]a}{r} = m - \frac{(r_1-r_2)a}{r}$$

$$\mathrm{E}(\underline{G}|x_1x_1X_2X_2) = m + \frac{[r_1(1-r_2) - (1-r_1)r_2]a}{r} = m + \frac{(r_1-r_2)a}{r}$$

$$\mathrm{E}(\underline{G}|X_1X_1X_2X_2) = m + \frac{[(1-r_1)(1-r_2) - r_1r_2]a}{1-r} = m + \frac{(1-r_1-r_2)a}{1-r}$$

The position of the QTL can be estimated, *i.e.* r_1 and, due to the equation $r = r_1 + r_2 - 2r_1r_2$, implicitly r_2, can be estimated by applying linear regression. For each of a number of tentative values for r_1 the regressor values are calculated. These are the values for the coefficients of a in the above expressions, *i.e.* for $-\frac{1-r_1-r_2}{1-r}$, $-\frac{r_1-r_2}{r}$, $\frac{r_1-r_2}{r}$ and $\frac{1-r_1-r_2}{1-r}$. Then, indeed, for

each tentative value for r_1 the trait values of the DH-lines are regressed onto these regressor values. Among all regression analyses the one yielding the smallest residual sum of squares of deviations is taken to indicate the most appropriate value for r_1 (Kearsey and Pooni, 1996). The values estimated, according to this regression analysis, for the intercept is an estimate for m and the value estimated for the slope is an estimate for a. This procedure is repeated for other adjacent pairs of markers, *viz.* X_2-x_2 and X_3-x_3, X_3-x_3 and X_4-x_4 *etc.*, in order to find the most likely position(s) of the QTL.

A QTL identified in the offspring of a particular cross may or may not be expressed in the offspring of another cross. Two reasons for this can be brought forward (Stam, 1998):

1. The QTL detected in the offspring of the first cross does not segregate in the offspring of the second cross because the parents involved in this second cross have the same homozygous genotype.
2. The expression of the QTL depends on the genetic background, which may vary from one pair of parents to the next.

When growing a certain population in a range of environmental conditions it may appear that some of the QTLs are expressed in all environments, whereas others are only expressed in specific conditions. The latter QTLs are responsible for genotype × environment interaction.

Since the major part of the nuclear DNA is non-coding, most markers are phenotypically neutral: presence or absence of a band has no effect on the phenotype. A polymorphic band may thus only be identified as a marker *in sensu stricto, i.e.* the mean phenotypic values of plants/lines belonging to different marker classes differ significantly. The following two factors determine the level of significance:

1. The accuracy of the estimates of the expected genotypic values of the marker classes.
   The expected genotypic value of a marker class is more accurately estimated by its mean phenotypic values as the number of entries sampled within the marker class is higher. Additionally the accuracy depends on the heritability of the trait, *i.e.* on the extensiveness of the evaluation (in the field or in the glasshouse) of the involved genotypes. The heritability will be higher when evaluating the genotypes in larger plots and/or in a higher number of replications. DH and F_∞ lines are therefore of special interest for QTL mapping
2. The size of the contrast between marker classes.
   It has been shown that a contrast depends on the strength of the linkage between the marker locus and the involved QTL: a contrast tends to be larger as r is closer to 0, *i.e.* as $(1-2r)$ is larger. Furthermore the contrast depends on the involved QTL: the larger its values for a (and d), the larger the contrast.

Marker-assisted selection

When the genetic control of the target trait is monogenic (or oligogenic) marker-assisted selection consists of selection of the candidates belonging to the marker class associated with the most attractive mean phenotypic value for trait Y. Alternatively, one may try to incorporate the attractive allele for trait Y by means of repeated backcrossing focussed on the introduction of the allele at 'locus' X-x linked to this attractive allele. This is only manageable if

1. There is strong linkage between 'locus' X-x and locus Y-y
2. The expression for trait Y is controlled by a small number of loci (Dudley, 1993)

When dealing with a polygenic target trait, the breeder may select for markers tagging favourable QTL alleles (and not with regard to the phenotype). Favourable alleles tagging different QTLs are then accumulated. In this way fruit size increasing QTL alleles occurring in a wild species of tomato *(Lycopersicon pimpinellifolium* Mill.) were efficiently transferred to lines of cultivated tomato (*L. esculentum* Mill.).

Marker-assisted selection may involve prediction of the genotypic values of the candidates on the basis of marker data. The prediction might be based on multiple linear regression of trait data on the markers, which – in this case – serve as regressors. The trait data should be derived from an evaluation of plant material under growing conditions identical to those applied in commercial cultivation. Thus, preferentially, the evaluation should involve large, replicated plots. In the regression analysis several, or even many, polymorphic bands should be evaluated as a marker with regard to the target trait. Example 12.9 illustrates such linear regression approach and suggests how marker data can assist in choosing pairs of parents to be crossed.

Example 12.9 Bos and Qi (1997) studied a set of 103 pure lines of barley. These were obtained after crossing cultivar Vada (P_1) and L94 (an exotic line, P_2) and developed by continued selfing in the absence of selection. They assigned numerical values to the regressors in the following way: if a genotype showed a band at a position identical to the position of a band in parent P_1, whereas that band did not occur in parent P_2, the genotype got the score 1 for this marker; otherwise it got the score 0. If the genotype showed a band at a position identical to the position of a band in P_2, whereas that band did not occur in P_1, the genotype got the score 0; otherwise it got the score 1. Altogether this coding rule simply implies that a genotype got score 1 for the considered band position in the case of similarity to P_1 and score 0 in the case of similarity to P_2.

When regressing (with backward elimination of markers) data for date of anthesis ($\underline{y}$) on 74 polymorphic AFLP-markers ($\underline{x}_1, \ldots, \underline{x}_{74}$) the following regression function was obtained:

$$\hat{y} = -4.37 - 3.29x_1 + 4.7x_2 + 4.09x_3 + 3.15x_4$$

The adjusted coefficient of multiple determination amounted to ${R_{adj}}^2 = 0.68$.

The regressors $x_1, \cdots, x_4$ represent four markers. The coefficient -3.29 of the first marker implies that a pure line with a marker genotype like Vada is expected to have a 3.29 days earlier anthesis than a pure line with a marker genotype like L94. A breeder dealing with the described plant material and pursuing a later date of anthesis, might select lines with marker genotype 0, 1,1 and 1, resp. for the four markers in the regression function. The expected date of anthesis of a line with such a complex marker genotype is 7.57 days; *i.e.* 7.57 days later than the overall mean date of anthesis. If the set of pure lines would not contain a line with the genotype coded as 0111, the breeder might try to generate it by crossing lines that are selected such that the line with the pursued genotype might occur in the set of pure lines obtained after crossing them. Thus cross 1000×0111 is expected to produce a line with genotype 0111 with a probability of $(\frac{1}{2})^4 = 0.0625$ if indeed the four markers would segregate independently.

12.3.3 Selection under Conditions Deviating from the Conditions Provided in Plant Production Practice

This section gives attention to reasons for applying indirect selection as discussed before. Breeders aim to develop plant material performing better under the conditions applied by professional farmers or horticulturists. The improvement with regard to the target trait is pursued by means of selection for that trait as expressed in the growing conditions required for efficient selection. For example, the selection may occur at a plant density that is low compared to the plant density applied by growers. Such a difference in growing conditions means that the actual selection should be considered as indirect selection for the target trait. Examples 12.10 and 12.11 give illustrations.

Example 12.10 Arboleda-Rivera and Compton (1974) applied mass selection in maize under three different conditions:

1. Selection in the rainy season
   When evaluating under rainy season conditions the response to direct selection is measured. It amounted to 10.5% per cycle for yield and to 8.8% per cycle for number of ears per plant. When testing in the dry season the improvement due to 'indirect' selection was only 0.8% for yield and 1.0% for number of ears per plant
2. Selection in the dry season
   When evaluating in the dry season the response per cycle amounted to 2.5% for yield and to 4.4% for number of ears per plant. When testing in

the rainy season the progress per cycle was 7.6% for yield and 11.4% for number of ears per plant.

3. Selection in both seasons
   When evaluated in the rainy season the increase of yield was 5.3% and that of number of ears 7.0%. In the dry season the progress for yield was 1.1% per cycle, whereas for number of ears per plant it was 3.3%.

Example 12.11 Ceccarelli, Grando and Impiglia (1998) studied the efficiency of direct selection of barley in stress environments in comparison with indirect selection in near-optimum environments followed by testing under stress conditions. They classified a certain environment, *i.e.* year-location combination, as a stress or a non-stress environment depending on whether the average grain yield of all lines tested in the particular environment was one or more standard deviations lower or higher than the average grain yield across all 8–10 (this depended on the set of lines) studied year-location combinations.

Lines were selected for high yield under stress (the YS set of lines) or non-stress (the YNS set of lines) during three growing seasons. All selected lines together with six checks were grown during four successive growing seasons in a total of 21 year-location combinations with average grain yield ranging from 0.35 to 4.86 $\mathrm{t.ha}^{-1}$.

The YS-lines yielded under stress 27% to 54% higher than the YNS-lines, with the top YS-lines yielding under stress between 16% and 30% more than the top YNS-lines. Under stress, the best YNS line ranked only 19^{th} for yield.

The study showed that the most effective way to increase grain yield under less-favourable conditions was to select in the target environment. Direct selection in the target environment may, however, be difficult or costly to implement if the target environment is remote or in an area with little infrastructure. One way to reach such areas is through decentralized-participatory breeding, an approach which brings genetic diversity to farmers (before it is reduced by selection in an environment very different from farmers' fields such as experiment stations).

The relative efficiency of selection under conditions deviating from the conditions provided by growers is now considered. The phenotypic value $\underline{p}_{\mathrm{Y}}$ represents an observation under grower's conditions, whereas $\underline{p}_{\mathrm{X}}$ represents the phenotypic value for the same trait (of the same genotype) but now observed under the growing conditions provided by the breeder. We consider first the correlation of $\underline{p}_{\mathrm{X}}$ and $\underline{p}_{\mathrm{Y}}$ across a set of genotypes. In the case of absence of covariance of genotypic value and environmental deviation one can derive:

$$\rho_p(\mathrm{X},\mathrm{Y}) = \frac{cov(\underline{p}_{\mathrm{X}},\underline{p}_{\mathrm{Y}})}{\sigma_{p_{\mathrm{X}}} \cdot \sigma_{p_{\mathrm{Y}}}} = \frac{cov(\underline{G}_{\mathrm{X}},\underline{G}_{\mathrm{Y}})}{\sigma_{p_{\mathrm{X}}} \cdot \sigma_{p_{\mathrm{Y}}}} = \rho_g \cdot \frac{\sigma_{g_{\mathrm{X}}}}{\sigma_{p_{\mathrm{X}}}} \cdot \frac{\sigma_{g_{\mathrm{Y}}}}{\sigma_{p_{\mathrm{Y}}}} = \rho_g h_{w_{\mathrm{X}}} h_{w_{\mathrm{Y}}} \tag{12.13}$$

This equation coincides with Equation (12.1) if $\rho_e = 0$, *i.e.* if $\text{cov}(\underline{e}_X, \underline{e}_Y) = 0$ (as well as with Equation (12.7a)). It is substituted into Equation (12.5), describing the ratio of the response to indirect selection for Y, via selection for X, and the response to direct selection for Y:

$$RSE = \frac{i_X}{i_Y} \cdot \rho_g \cdot \frac{h_{w_X}}{h_{w_Y}} = \frac{i_X}{i_Y} \cdot \rho_p \cdot \frac{h_{w_X}}{h_{w_X} h^2_{w_X}} = \frac{i_X}{i_Y} \cdot \rho_p \cdot \frac{1}{h^2_{w_Y}} \tag{12.14}$$

At $i_X = i_Y$ the relative loss in potential selection response, due to selection under conditions deviating from the conditions prevailing in practice, amounts to

$$\frac{R_Y - CR_Y}{R_Y} = 1 - RSE = 1 - \left(\rho_p \cdot \frac{1}{{h_{w_Y}}^2}\right) \tag{12.15}$$

It amounts thus to $100(1 - RSE)\%$ and it will be large if RSE is small, *i.e.* if ρ_p is small and/or ${h_Y}^2$ is high. Example 12.12 gives an illustration.

Example 12.12 Kramer, van Ooijen and Spitters (1982) considered the grain yield of spring wheat genotypes. These genotypes were evaluated on the basis of two plot types:

1. 2 m long single-row-plots, interrow distance 20.8 cm
2. 6 m long six-row-plots, interrow distance 25 cm

Because of the virtual absence of intergenotypic competition it was assumed that the six-row-plots provide commercial growing conditions. For the six-row-plots the heritability was estimated to be 0.88. Furthermore the estimate for ρ_p amounted to 0.31. Then the relative selection efficiency amounts to

$$RSE = \frac{i_X}{i_Y} \cdot \frac{0.31}{0.88}$$

i.e. for $i_X = i_Y$ to 0.35. In this experiment the loss in potential selection response due to selection on the basis of single-row-plots was 65%. Apparently interplot competition is an important cause for a bias when evaluating candidates on the basis of single-row-plots.

The relative selection efficiency measures the quality, with regard to the response to selection, of the growing conditions provided by a breeder in comparison to the quality of the growing conditions in the target environment. The 'deficit' in selection response, due to the use of sub-optimal growing conditions, can be calculated if $\rho_p(X, Y)$ and ${h_w}^2(Y)$ are known. To be able to estimate these parameters the considered set of genotypes should be evaluated under both types of environmental conditions. The efforts required for this are rarely made and the criterion measuring the relative efficiency of selection under conditions deviating from the conditions provided by growers is rarely used. In fact, statements about optimum conditions for selection (Section 11.1 and hereafter) are often merely opinions. This is partly due to the fact that a

breeder aims to develop plant material well adapted to a wide range of conditions instead of a single well-defined target environment. Furthermore, one should realize that optimum conditions may differ from trait to trait.

The opinions about the optimum conditions for selection to be discussed here concern

1. Plant density
2. Quality of the growing conditions

Optimum plant density for selection

There is no general agreement about the optimum plant density for selection. Spitters (1979, p. 117) advocated selection at high plant density. Fasoulas (1981, p. 58), however, preferred a density so low that interplant competition does not occur.

Both points of view are, in fact, merely opinions and are not based on experimental evidence. Bos (1981, p. 150) re-analysed data of Spitters' experiments with barley and concluded in favour of a low plant density. Fasoulas and Tsaftaris (1975, p. 29) and Kyriakou and Fasoulas (1985) concluded without reserve that a very low density, e. g. only 1.43 plants per m^2 for wheat, is to be preferred. However, on the basis of experiments with spring rye, Pasini and Bos (1990a,b) were very reservedly in favour of a very low density. Bussemakers and Bos (1999) concluded that mass selection should be applied at the high plant density used in commercial practice (Example 15.6).

It is concluded that a clear-cut advice with regard to the plant density to be applied in selection cannot yet be given. The topic is considered further in Section 15.2.1.

Optimum quality of growing conditions for selection

It is admitted that plant density is an aspect of the growing conditions. It was, however, thought to be appropriate to consider plant density separately in a discussion on optimum conditions for selection.

Also with regard to the quality of the growing conditions there is no general agreement about what conditions are optimal for selection. Fasoulas (1973, p. 23) concluded that the growing conditions in the selection field should permit unrestricted growth and development of the plant material. McVetty and Evans (1980), on the other hand, stated that for selection in wheat that it did not matter whether selection occurred under optimum, *i.e.* non-stress, conditions or not. Rosielle and Hamblin (1981) do not generalize. They followed Equation (12.5) when stating: 'The situation most favourable to plant breeders would be one in which genetic variances in stress environments are greater than those in non-stress environments and genetic correlations between yields in stress and non-stress environments are highly positive'. A well-known

application of this statement is selection for disease resistance under the stress conditions due to artificial inoculation.

12.4 Estimation of the Coefficient of Phenotypic, Environmental, Genetic or Additive Genetic Correlation

This section gives attention to the estimation of coefficients of correlation. Some of these coefficients may be interpreted in quantitative genetic terms. This can only be justified if a number of assumptions apply (see Section 10.2.1). When estimating, on the basis of different procedures, one may encounter variable, strange and/or unreliable estimates. Indeed, the estimation procedures may differ with regard to their bias. They may also differ with regard to their accuracy (which mainly depends on the sample size), see Example 12.20.

Estimation of $\boldsymbol{\rho_p}$

The coefficient of phenotypic correlation of traits X and Y concerns the phenotypic values for these traits. It is estimated on the basis of a sample of plants representing the population of interest. Because phenotypic values can easily be obtained, the estimation of ρ_p is straightforward (see Example 12.13).

Example 12.13 For individual cereal plants the phenotypic values for grain yield and plant height can easily be obtained.

Bos (1981, p.35 and p. 78) studied in winter rye the phenotypic correlation of grain yield and plant height. Plants belonging to different generations of an open pollinating population subjected to selection for high grain yield and reduced plant height were observed. These plants were grown in the seasons 1974–75 and 1977–78. The estimates amounted to $r_p = 0.52 (n = 57)$ and $r_p = 0.30 (n = 200)$.

Estimation of $\boldsymbol{\rho_e}$

The environmental correlation of traits X and Y is the correlation of the environmental deviations of the considered candidates with regard to these traits.

In genetically homogeneous plant material the phenotypic correlation is to be interpreted as the environmental correlation (Section 12.2). For such plant material ρ_e can be estimated in the same way as ρ_p (Example 12.13).

In other plant material ρ_e may be estimated from Equation (12.1). This is illustrated in Example 12.14.

Example 12.14 Van der Vossen (1974, p. 28, 45) studied, in oil palm, the genetic control of number of bunches (trait X) and mean single bunch weight (trait Y).

He estimated the narrow sense heritability of these traits by means of offspring-midparent regression. This yielded $\hat{h}^2_{n_X} = 0.512$ and $\hat{h}^2_{n_Y} = 0.206$. The genetic correlation coefficient was estimated to be $r_g = -0.584$ (see Example 12.15) and the phenotypic correlation coefficient was estimated to be $r_p = -0.59$. When assuming ${h_n}^2 = {h_w}^2$, the environmental correlation coefficient can be estimated from Equation (12.1):

$$-0.59 = -0.584 \times 0.716 \times 0.454 + r_e \times 0.699 \times 0.891$$

This yields $r_e = -0.643$.

Estimation of $\boldsymbol{\rho_g}$

The genetic correlation of traits X and Y is the correlation of the genotypic values for X and Y. Reliable information about genotypic values is rarely available. Thus ρ_g is mostly evaluated in an indirect way. The procedure to be applied is dictated mainly by the nature of the plant material processed by the breeder or the researcher. The following procedures to estimate ρ_g are elaborated:

1. Estimation on the basis of the genotypic values
2. Estimation by using genetically uniform plant material
3. Estimation from the relative selection efficiency

Estimation of $\boldsymbol{\rho_g}$ *from genotypic values*

It may be demanding to obtain unbiased and accurate estimates of genotypic values. Kearsey and Pooni (1996, p. 288) described the following procedure.

Each of I genotypes is represented by J plants. (This implies that one is dealing with clones, pure lines or F_1 hybrids). Each of these IJ plants is assigned a position in the field by means of single plant randomization. For each set of J plants half the number of plants are sampled to be observed with regard to trait X and the remainder is observed with regard to trait Y. Then the genotypic values $\mathcal{G}_X$ and $\mathcal{G}_Y$ are estimated for each genotype (by means of $\overline{p}_X$ and $\overline{p}_X$, respectively). The correlation of $\overline{p}_X$ and $\overline{p}_Y$ is then exclusively due to genetic variation among the J genotypes. The estimate $r(\overline{p}_X, \overline{p}_Y)$ of these correlation may then be interpreted as an estimate of $\rho_g(X, Y)$, especially for $J \to \infty$.

This procedure is not only demanding but may yield an estimate biased with regard to the coefficient of genetic correlation applying to the growing conditions of commercial practice. Estimation of ρ_g on the basis of estimates

of genotypic values may thus tend to be unreliable. Example 12.15 describes estimation of ρ_g according to this procedure.

Example 12.15 Van der Vossen (1974, p. 45) established for 52 palms the genotypic values for number of bunches (trait X) and mean single bunch weight (trait Y). These values were obtained by applying a procedure outlined in Example 9.2. This procedure is based on the assumption of absence of dominance. Thus the genetic correlation coefficient could be estimated directly from the estimates of the genotypic values. This yielded $r_g = -0.584$.

Estimation of $\boldsymbol{\rho_g}$ by using genetically uniform plant material

One may estimate for some genetically heterogeneous population the phenotypic variances for traits X and Y as well as the phenotypic covariance of these traits. Additionally genetically uniform plant material may be used to estimate the environmental variances of these traits and the environmental covariance. By subtracting the latter estimates from the corresponding estimates for the former parameters one obtains estimates for the genetic variances and the covariance. From these one can calculate an estimate for ρ_g. The procedure is illustrated by Example 12.16.

Example 12.16 Weber and Moorthy (1952) studied the relation between 100-grain-weight (trait X) and oil content (trait Y) in soybean.

For a genetically heterogeneous F_2 population they obtained

$$\hat{\text{var}}(\underline{p}_X) = 2.28,$$
$$\hat{\text{var}}(\underline{p}_Y) = 0.54$$

and

$$\hat{\text{cov}}(\underline{p}_X, \underline{p}_Y) = -0.26,$$

yielding

$$r_p = \frac{-0.26}{1.51 \times 0.73} = -0.23$$

For genetically homogeneous plant material they got the following estimates;

$$\hat{\text{var}}(\underline{e}_X) = 1.05,$$
$$\hat{\text{var}}(\underline{e}_Y) = 0.24$$

and

$$\hat{\text{cov}}(\underline{e}_X, \underline{e}_Y) = -0.09.$$

Thus ρ_e was estimated to be

$$\boldsymbol{r}_e = \frac{-0.09}{1.025 \times 0.490} = -0.18$$

Because

$$\text{var}(\underline{p}) = \text{var}(\underline{\mathcal{G}}) + \text{var}(\underline{e})$$

and

$$\text{cov}(\underline{p}_\text{X}, \underline{p}_\text{Y}) = \text{cov}(\underline{\mathcal{G}}_\text{X} + \underline{e}_\text{X}, \underline{\mathcal{G}}_\text{Y} + \underline{e}_\text{Y}) = \text{cov}(\underline{\mathcal{G}}_\text{X}, \underline{\mathcal{G}}_\text{Y}) + \text{cov}(\underline{e}_\text{X}, \underline{e}_\text{Y}),$$

they also got

$$\begin{aligned} \hat{\text{var}}(\underline{\mathcal{G}}_\text{X}) &= 2.28 - 1.05 = 1.23, \\ \hat{\text{var}}(\underline{\mathcal{G}}_\text{Y}) &= 0.54 - 0.24 = 0.30 \end{aligned}$$

and

$$\hat{\text{cov}}(\underline{\mathcal{G}}_\text{X}, \underline{\mathcal{G}}_\text{Y}) = -0.26 - (-0.09) = -0.17.$$

This yields

$$\boldsymbol{r}_g = \frac{-0.17}{1.11 \times 0.55} = -0.28$$

Estimation of $\boldsymbol{\rho_g}$ from the relative selection efficiency

One may estimate ρ_g when knowing the relative selection efficiency and the heritability values in the wide sense of traits X and Y (see Equation (12.5)). Example 12.19 illustrates a similar procedure for the estimation of ρ_a.

Estimation of $\boldsymbol{\rho_a}$

The additive genetic correlation (ρ_a) was defined in Section 12.3.1. The following procedures to estimate ρ_a are elaborated:

1. Estimation by regressing offspring data on maternal parent data
2. Estimation on the basis of an analysis of covariance
3. Estimation from the relative selection efficiency

Estimation of $\boldsymbol{\rho_a}$ on the basis of regression of the performance of offspring on the performance of maternal plants

According to Equation (10.10), which applies to cross-fertilizing species, the additive genetic variance for trait X is equal to

$$2cov(\underline{p}_{M_\text{X}}, \underline{p}_{HS_\text{X}})$$

Analogous to this, $\mathrm{cov}(\underline{\gamma}_{\mathrm{X}}, \underline{\gamma}_{\mathrm{Y}})$, *i.e.* the **additive genetic covariance** of traits X and Y (say $\mathrm{cov}_a(\mathrm{X}, \mathrm{Y})$), of a cross-fertilizing species is taken to be equal to

$$cov_a(\mathrm{X}, \mathrm{Y}) = cov(\underline{p}_{M_{\mathrm{X}}}, \underline{p}_{HS_{\mathrm{Y}}}) + cov(\underline{p}_{M_{\mathrm{Y}}}, \underline{p}_{HS_{\mathrm{X}}}) \tag{12.16}$$

when calculating the arithmetic mean across the two covariances, and to

$$cov_a(\mathrm{X}, \mathrm{Y}) = \sqrt{cov(\underline{p}_{M_{\mathrm{X}}}, \underline{p}_{HS_{\mathrm{Y}}}) \cdot cov(\underline{p}_{M_{\mathrm{Y}}}, \underline{p}_{HS_{\mathrm{X}}})} \tag{12.17}$$

when calculating the geometric mean across the 2 covariances.

The additive genetic correlation, *i.e.*

$$\rho_a(\mathrm{X}, \mathrm{Y}) = \frac{cov_a(\mathrm{X}, \mathrm{Y})}{\sigma_{a_{\mathrm{X}}} \cdot \sigma_{a_{\mathrm{Y}}}} \tag{12.18}$$

is then estimated on the basis of estimates for $\mathrm{cov}_a(\mathrm{X}, \mathrm{Y})$, $\sigma^2_{a_{\mathrm{X}}}$ and $\sigma^2_{a_{\mathrm{Y}}}$.
N. B. When estimating $\mathrm{cov}_a(\mathrm{X}, \mathrm{Y})$ on the basis of Equation (12.18) it gets the sign (+ or −) obtained when applying Equation (12.17) or the sign obtained for the estimate of ρ_p. Example 12.17 illustrates the procedure.

Example 12.17 Bos (1981, Table 9 and p. 35) studied the inheritance of plant height (trait X) and grain yield (trait Y) in an open-pollinating population of winter rye, consisting of 5260 plants, This was done on the basis of a random sample of 57 plants and their offspring. The parents were grown during the season 1974–75 and their offspring during the season 1975–76. The regression of the mean performance of HS-families on the performance of their maternal parents was calculated.

For trait X the following was obtained:

$$\hat{\mathrm{cov}}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) = 31.67\,\mathrm{cm}^2, i.e. \hat{\sigma}^2_a = 63.34\,\mathrm{cm}^2$$

and

$$\hat{\mathrm{var}}(\underline{p}_{\mathrm{M}}) = 163.58\,\mathrm{cm}^2.$$

Equation (11.39) yields thus

$$\hat{\boldsymbol{h}}^2_{n_{\mathrm{X}}} = \frac{2(31.67)}{163.58} = 0.387$$

For trait Y it was derived that

$$\hat{\mathrm{cov}}(\underline{p}_{\mathrm{M}}, \underline{p}_{\mathrm{HS}}) = 107.7\,\mathrm{dg}^2, i.e. \hat{\sigma}^2_a = 215.5\,\mathrm{dg}^2$$

and

$$\hat{\mathrm{var}}(\underline{p}_{\mathrm{M}}) = 4869.1\,\mathrm{dg}^2$$

Thus

$$\hat{h}^2_{n_{\mathrm{Y}}} = 0.044$$

The covariances of X and Y were estimated to be

$$\hat{cov}(\underline{p}_{M_X}, \underline{p}_{HS_Y}) = 9.9\,\text{cm} \cdot \text{dg},$$

and

$$\hat{cov}(\underline{p}_{M_Y}, \underline{p}_{HS_X}) = 69.1\,\text{cm} \cdot \text{dg}.$$

The estimate of ρ_a according to Equation (12.16) is then

$$\boldsymbol{r}_a = \frac{9.9 + 69.1}{\sqrt{63.34 \times 215.5}} = 0.68$$

and that according to Equation (12.17) is

$$\boldsymbol{r}_a = \frac{2\sqrt{9.9 \times 69.1}}{\sqrt{63.34 \times 215.5}} = 0.45$$

Estimation of $\boldsymbol{\rho_a}$ on the basis of an analysis of (co)variance

The additive genetic variance of trait X, say $\sigma^2_{a\mathrm{X}}$ can be estimated on the basis of an analysis of variance of HS-families obtained from open pollination in the sampled population (Table 11.3). It is estimated according to Equation (11.37).

In a similar way, one may estimate the additive genetic covariance of traits X and Y ($\mathrm{cov_a(X, Y)}$) on the basis of an analysis of covariance. Instead of calculating sums of squares for each of the traits X and Y, as in the analysis of variance, one should calculate similar sums of products. Table 12.1 presents the analysis of covariance that applies when I HS-families are tested in each of J blocks.

The additive genetic covariance of X and Y is thus estimated by

$$\hat{cov}_a(\mathrm{X}, \mathrm{Y}) = 4\hat{cov}(\underline{\mathcal{G}}_{HS_X}, \underline{\mathcal{G}}_{HS_Y}) = \frac{4(MP_f - MP_r)}{J} \tag{12.19}$$

Table 12.1 The structure of the analysis of covariance of data obtained from a randomized complete block experiment with I HS-families, each evaluated, for two traits, in each of J blocks. The columns headed by SP, MP and E(MP) present sums of products, mean products and expectations of the mean products.

Source of covariation	df	SP	MP	E(MP)
Blocks	$J-1$	$\mathrm{SP_b}$	$\mathrm{MP_b}$	$\mathrm{cov_r} + I\mathrm{cov_b}$
HS-families	$I-1$	$\mathrm{SP_f}$	$\mathrm{MP_f}$	$\mathrm{cov_r} + J\mathrm{cov_f}$
Residual	$(J-1)(I-1)$	$\mathrm{SP_r}$	$\mathrm{MP_r}$	$\mathrm{cov_r}$

The procedure is illustrated by Example 12.18.

Example 12.18 Bos (1981, p. 94) estimated for a population of winter rye plants the additive genetic correlation of plant height (trait X) and grain yield (trait Y). HS-families were obtained by harvesting of a random sample of 102 winter rye plants taken from an open-pollinating population of 5, 111 plants grown in the season 1977–78. Each HS-family was grown in the next season as a single-row plot in each of two complete blocks.

The analyses of variance yielded

$$\hat{\sigma}^2_{\boldsymbol{a}}(\mathrm{X}) = 56.56\,\mathrm{cm}^2,$$

and

$$\hat{\sigma}^2_{\boldsymbol{a}}(\mathrm{Y}) = 54.12\,\mathrm{dg}^2.$$

The analysis of covariance yielded

$$\hat{\mathrm{cov}}_a(\mathrm{X},\mathrm{Y}) = -15.76\,\mathrm{cmdg}.$$

Thus:

$$\boldsymbol{r}_a = \frac{-15.76}{\sqrt{56.56 \times 54.12}} = -0.28$$

Estimation of $\boldsymbol{\rho_a}$ from the relative selection efficiency

When knowing the relative selection efficiency and the heritability values in the narrow sense of traits X and Y one may estimate ρ_a on the basis of Equation (12.9). Example 12.19 illustrates the procedure.

Example 12.19 Bos (1981, Tables 34 and 30) grew, during the season 1976–77, a diploid population of winter rye. Plants with a high grain yield (trait X) were selected. The response to selection was measured according to the second procedure mentioned in Section 11.1. This yielded:

$$R_{\mathrm{X}} = 69.45 - 65.66 = 3.79\,\mathrm{dg}$$

With regard to plant height (trait Y), a correlated response was observed:

$$CR_{\mathrm{Y}} = 147.1 - 146.42 = 0.68\,\mathrm{cm}.$$

The heritability values in the narrow sense of these two traits were estimated to be

$$\hat{h}^2_n(\mathrm{X}) = 0.034$$

and

$$\hat{h}^2_n(\mathrm{Y}) = 0.43$$

According to Equation (12.3) the relative selection efficiency amounted to

$$\frac{CR_{\mathrm{Y}}}{R_{\mathrm{X}}} = \frac{0.68}{3.79} = 0.179\,\mathrm{cm/dg}$$

The additive genetic coefficient of correlation (ρ_a) can then be estimated from Equation (12.9), *i.e.*

$$RSE = \rho_a \cdot \frac{h_{n_{\mathrm{X}}}}{h_{n_{\mathrm{Y}}}}$$

It was estimated to be

$$\boldsymbol{r}_a = \frac{0.179}{\sqrt{\frac{0.034}{0.43}}} = 0.64$$

In the first paragraph of this section it was remarked that estimation of ρ_g or ρ_a may yield rather different values. This is, in fact, a general experience, even when estimating for the same population. The phenomenon is illustrated by Example 12.20. It is due to inaccuracy and/or bias of the involved estimators, as well as to differences between the estimators with regard to their accuracy and/or bias. Thus estimates obtained for ρ_g or ρ_a should only be used as rough indications when considering the efficiency of indirect selection.

Example 12.20 In the preceding examples the following estimates for the correlation of plant height and grain yield of winter rye plants belonging to the same population (be it under selection) were obtained:

Example 12.13: $r_p = 0.52$ or 0.30

Example 12.17: $r_a = 0.68$ or 0.45

Example 12.18: $r_a = -0.28$

Example 12.19: $r_g = 0.64$

These are rather different estimates for more or less the same parameters estimated for more or less the same populations.

12.5 Index Selection and Independent-Culling-Levels Selection

Index selection is a form of indirect selection for a complex trait. It aims to realize a correlated response to selection with regard to some complex target trait Y, *e.g.* financial yield, by selecting candidates which are superior with regard to an abstract trait I, the index. For each candidate the (phenotypic)

index value p_{I} is calculated from a linear function of the phenotypic values $\underline{p}_1, \underline{p}_2, \ldots, \underline{p}_n$ for auxiliary traits $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$. Thus:

$$\underline{p}_{\mathrm{I}} = \beta_1 \underline{p}_1 + \beta_2 \underline{p}_2 + \ldots + \beta_n \underline{p}_n = \sum_{i=1}^{n} \beta_i \underline{p} = \boldsymbol{\beta'} \underline{\boldsymbol{p}} \tag{12.20}$$

where

$\boldsymbol{\beta'}$ = a $1 \times n$ vector of unknown coefficients and

$\underline{\boldsymbol{p}}$ = a $n \times 1$ vector of phenotypic values for the n auxiliary traits

In fact the target of the selection is improvement of $\mathcal{G}_{\mathrm{Y}}$, the genotypic value with regard to the complex target trait Y. The quantity $\mathcal{G}_{\mathrm{Y}}$ is defined as

$$\mathcal{G}_{\mathrm{Y}} = w_1 \mathcal{G}_1 + w_2 \mathcal{G}_2 + \ldots + w_n \mathcal{G}_n \tag{12.21}$$

where

$\mathcal{G}_i$ = the genotypic value for trait X_i and

w_i = the the relative economic weight of X_i.

The coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_n$, in Equation (12.20) have to be determined in such a way, that $\mathcal{G}_{\mathrm{Y}}$ is maximally increased. Smith (1936) derived for the case of a linear relation between $\mathcal{G}_{\mathrm{Y}}$ and p_{I} that maximum progress with regard to $\mathcal{G}_{\mathrm{Y}}$, when selecting for $\underline{p}_{\mathrm{I}}$, is attained at a maximum value for $\rho(\underline{p}_{\mathrm{I}}, \mathcal{G}_{\mathrm{Y}})$, *i.e.* the coefficient of correlation of $\underline{p}_{\mathrm{I}}$ and $\mathcal{G}_{\mathrm{Y}}$. This occurs when substituting in Equation (12.20) values for β_i equal to be the solution to the equation

$$\boldsymbol{P\beta} = \boldsymbol{Gw} \tag{12.22}$$

where

$\boldsymbol{P}$:= the $n \times n$-matrix of phenotypic variances and covariances

$\boldsymbol{G}$:= the $n \times n$-matrix of genotypic variances and covariances

$\boldsymbol{w}$:= the $n \times 1$-vector of relative economic weights

When having determined $\boldsymbol{P}$, $\boldsymbol{G}$ and $\boldsymbol{w}$, the solution for the βs is given by

$$\mathbf{b} = \boldsymbol{P}^{-1} \boldsymbol{Gw}$$

Thus the value to be assigned to β_i depends on

- the relative economic weight of trait X_i
- the phenotypic and genetic variance of X_i and
- the phenotypic and genetic covariances of X_i with the other traits.

When knowing the population parameters an **optimum index** can be determined. Generally, of course, one has to estimate, on the basis of an evaluation of a random sample from the plant material to be improved, the variances and covariances. Determination of the **estimated index** is then rather demanding

The assignment of economic weights may also offer problems because the financial yield of a crop depends on the prices of yield components which vary from site to site and from year to year. These yield components can be the yield of tomatoes or cucumbers in different weight or size classes. For sugar beets the financial yield depends on gross yield as well as sugar content. For cereals it depends on grain and straw yield. It is still more complicated when the traits concern different aspects. In wheat, for instance, protein content and flour yield are related to grain quality, whereas grain yield concerns grain quantity.

An alternative approach for determining relative economic weights proceeds as follows. The breeder determines values for $w_1, \ldots, w_n$ on the basis of multiple regression of the economic value of an entry on the phenotypic values for traits $\mathrm{X}_1, \ldots, \mathrm{X}_n$ (Cotterill and Jackson, 1985). Thus for each trait involved in the regression one obtains an estimate for the corresponding regression coefficient. This regression coefficient indicates the increase in the economic value of a candidate expected at a one-unit increase for the considered trait. When breeding woody crops or fruit-producing perennial crops, one may tend to include traits that can be observed in an early ontogenetic stage, whereas the economic value is determined for adult plants.

If reliable estimates for the phenotypic and genetic variances and covariances are not available the estimated relative economic weights may be used as coefficients in Equation (12.20):

$$\underline{p}_{\mathrm{I}} = \sum_{i=1}^{n} w_i \underline{p} = \mathbf{w}' \underline{\boldsymbol{p}} \tag{12.23}$$

This is the so-called **base index**. It requires only determination of the economic weights (of the genotypic values) of the considered traits.

Because of the problems mentioned above, the assignment of relative economic weights or of coefficients occurs often in a different, sometimes rather intuitive way. A few examples are mentioned.

A weight-free multiplicative index is the so-called **Elston-index** (see Baker, 1986), *i.e.*

$$\underline{p}_{\mathrm{I}} = \prod_{i=1}^{n} (\underline{p}_i - p_{min,i})$$

where

$p_{min,i}$:= the minimally acceptable phenotypic value for trait X_i.

When applying the method of 'desired genetic gains' one should indicate for each trait included in the index the desired relative progress. If $\boldsymbol{p}_d$

designates the vector of relative progresses desired, the vector of coefficients used in Equation (12.20) in order to calculate the phenotypic index values is taken to be

$$\boldsymbol{b} = \boldsymbol{G}^{-1}\boldsymbol{p_d}$$

According to Equation (12.22) this implies

$$\boldsymbol{w} = \boldsymbol{G}^{-1}\boldsymbol{P}\mathbf{b} = \boldsymbol{G}^{-1}\boldsymbol{P}\boldsymbol{G}^{-1}\boldsymbol{p_d}$$

The last approach mentioned here is the method of 'equal emphasis on each trait'. In this method the phenotypic standard deviations of the traits are taken into consideration by taking:

$$w_i = \frac{1}{\sigma_{P_i}}$$

Many different procedures for index selection have thus been proposed. The evaluation of the relative merits of these procedures is a demanding task. In this section the topic of index selection is therefore only briefly introduced. For an extensive treatment the reader is referred to Baker (1986). Because of the problems encountered with optimum and estimated indices, Baker seems to suggest that application of a base index or a weight-free index is to be preferred. Example 12.21 presents some results of application of index selection.

Genetic improvement due to index selection may be negligible due to inadequacy of the estimates of phenotypic and genetic variances and covariances. Thus results may be obtained that are not better than those obtained when applying visual selection for general impression.

Example 12.21 Brim, Johnson and Cockerham (1959) illustrated the application of index selection in soybean, where oil and protein content are the most important components of yield. They calculated, for a 1.0 (for oil yield in g/plot) : 0.6 (for protein yield in g/plot) price ratio, the genetic improvement expected from selecting the 5% top of the F_3 lines. Only expected responses were reported.

Elgin, Hill and Zeiders (1970) applied, for five generations, four selection procedures in alfalfa. The procedures were: tandem selection (Section 12.1), ICL-selection, estimated index and base index. Selection was for resistance against four foliar diseases and for good regrowth after harvest. At the end of the study an evaluation trial was conducted. Twenty four entries were evaluated:

- each of the five generation for each selection procedure;
- the original population developed from an intercross of 45 randomly selected plants each of variety DuPuits and variety Vernal;
- a hybrid of DuPuits and Vernal; and
- the two parental varieties.

In addition to measuring the traits, the total merit of each entry was calculated. With regard to this trait, the response to base index selection was clearly better than the results of the other selection procedures.

De Wolff (1972, p. 51, 42) derived for maize the index

$$\underline{p}_{\mathrm{I}} = 1.53\underline{p}_1 - 0.398\underline{p}_2 + 0.416\underline{p}_3 + 6.14\underline{p}_4$$

The traits involved in this index are

X_1: number of days from planting until the appearance of the tassel;

X_2: number of leaves;

X_3: length (in cm) of the 8th leaf from the top; and

X_4: largest width (in cm) of the 8th leaf from the top.

This index contains traits that can be observed before pollen distribution. This allows indirect selection for yield via traits observed in both male and female parents.

The estimates of the heritability values and of the genetic correlations with yield amounted to:

		h^2	r
Trait:	X_1	0.75	0.40
	X_2	0.84	0.18
	X_3	0.46	0.57
	X_4	0.32	0.85

Independent-culling-levels selection is a form of intentional simultaneous selection. It consists of truncation selection with regard to each of several traits. Thus for each trait, a phenotypic value minimally required for selection (p_{min}) is determined. A candidate is rejected if its phenotypic value does not exceed p_{min} for one or more traits, whatever its quality for all other traits. In contrast to index selection, independent-culling-levels selection does not allow mutual compensation of favourable and unfavourable phenotypic values for different traits. Example 12.22 reports about an application.

Example 12.22 In a study to the efficiency of cross-prediction procedures in potato breeding (Brown and Caligari, 1988), one of these procedures considered the frequency, in each of eight subsamples, of clones satisfying independent-culling-levels for:

- total tuber weight;
- number of tubers;

- mean tuber weight; and
- (the visually assessed score for) regularity of tuber shape.

(For each of the eight crosses a subsample consisting of 25 random clones was studied). Independent-culling-levels selection was applied with $p_{min} = \bar{p}$, for each of the four traits across all 200 clones. For each cross population, the frequency of 'surviving' clones was the basis for the prediction of the number of superior clones expected to occur in (much) larger samples.

Some suggestions for choosing p_{min} values are

- Choose the p_{min} values in such a way that their standardized values are the same for all traits. Then the different traits are subjected to the same intensity of selection (if indeed each trait has a normal probability distribution).
- Choose the p_{min} values in such a way that the ratios of their standardized values are equal to the ratios of their heritability values.
- Choose the intensities of selection in such a way that their ratios are equal to the ratios of the heritability values.

Until now the merits of suggestions with regard to independent-culling-level selection, such as the preceding ones, have not been studied.

Chapter 13
Genotype × Environment Interaction

The genotypic value of some candidate was defined to be equal to the expected phenotypic value at given macro-environmental conditions. This means that the genotypic value of a given candidate depends on the macro-environmental growing conditions. It also means that differences between candidates depend on the macro-environmental conditions. This phenomenon is called genotype by environment interaction ($g \times e$ interaction). It may even mean that ranking of candidates according to their genotypic value depends on the macro-environmental growing conditions. The latter is a disturbance to breeders, who generally want to select candidates performing, under diverse conditions, in a superior way. The phenomenon is also disturbing when testing varieties developed for hopefully a wide range of conditions. This chapter elaborates some relevant aspects of $g \times e$ interaction.

13.1 Introduction

Until now it was consistently assumed that all plants of all generations are exposed to the same macro-environmental growing conditions. This is, of course, only appropriate if one is interested in the performance of genotypes in a specific macro-environment, say macro-environment k. The partitioning of the phenotypic value of some candidate in this macro-environment was given by Equation (8.1), *i.e.*

$$\underline{p} = \mathcal{G} + \underline{e},$$

The genotypic value of the candidate was defined as

$$\mathcal{G} = \mathrm{E}(\underline{p}|gt, E_k)$$

were the genotype of the candidate (gt) and the macro-environmental conditions (E_k) of the evaluation are specified. A particular macro-environment is characterized by the growing conditions of a particular location, a particular growing season (or year) or the combination of a particular location and a particular year. Additionally it may be characterized by the growing conditions due to some temperature regime, *e.g.* in a glasshouse, some amount of fertilizer, some plant density, *etc.*

The quality of the macro-environmental growing conditions is thus reflected by the genotypic value. Indeed, the genotypic value of the considered genotype depends not only on the macro-environmental conditions, but possibly also on the effect of interaction of the considered genotype and the considered

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 325–338.

macro-environment. An appropriate linear model for the genotypic value of genotype j grown in macro-environment k is thus

$$\mathcal{G}_{jk} = \mu + G_j + E_k + (GE)_{jk} \tag{13.1}$$

where

μ := the mean across all considered genotypes and macro-environments
G_j := the main effect of genotype $j; j = 1, 2, \ldots, J$
E_k := the main effect of macro-environment $k; k = 1, 2, \ldots, K$ (this quantity is sometimes indicated as **environmental index**) and
$(GE)_{jk}$:= the effect of the interaction of genotype j and macro-environment k

The model above is similar to the linear model for partitioning of the genotypic value in terms of general and specific combining ability (Equation (11.51)). Similarly, the model given by Equation contains a parameter μ, defined such that the means of the contributions G_j, E_k and $(GE)_{jk}$ to $\mathcal{G}_{jk}$ are zero. Thus the mean value of G_j across the J genotypes, the mean value of E_k across the K macro-environments and the mean value of $(GE)_{jk}$ across all JK combinations of a genotype and a macro-environment are all zero.

The model given by Equation (13.1) implies that the difference between the genotypic values of genotypes j and j' in macro-environment k does not only depend on the main effects of the considered genotypes but also on the effect of their interactions with the considered macro-environment:

$$\mathcal{G}_{jk} - \mathcal{G}_{j'k} = [G_j + (GE)_{jk}] - [G_{k'} + (GE)_{j'k}] \tag{13.2}$$

This implies that a genotypic value is due to confounding of a main genotype effect and an effect of $g \times e$ interaction. One can only estimate G_j and $(GE)_{jk}$ separately when testing genotype j in a set of macro-environments. It also implies that estimates of $\text{var}(\underline{\mathcal{G}})$ based on data from a single macro-environment overestimate $\text{var}(\underline{G})$, *i.e.* the variance of genotypic values across macro-environments. Comstock and Moll (1963) indicated that $\text{var}(\underline{G})$ tends to be smaller if the macro-environments are more diverse.

In macro-environment k the difference between the genotypic values of genotypes j and j' will be different from that in macro-environment k'. Interaction may thus give rise to different rankings of the genotypes (of candidates or of established varieties) in different macro-environments. This is illustrated by Examples 13.1, 13.2 and 13.3.

Example 13.1 Cuany *et al.* (1970), in Frey (1971), observed the yield of three maize cultivars in two macro-environments, viz. application of irrigation or not (Table 13.1).

Table 13.1 The grain yield of three maize cultivars in the presence and absence of irrigation (source: Cuany *et al.* (1970) in Frey, 1971)

		Grain yield (kg/ha)	
Type of cultivar	Cultivar	Under irrigation	Rainfed
open pollinating	Phillips 67	7,965	2,069
hybrid	Pioneer 3579	12,105	1,756
hybrid	Nebr. 501D	13,305	2,132

Under irrigation the grain yields of the hybrids were much higher than that of the open pollinating variety. In the absence of irrigation, however, the yields of the two cultivar types were equivalent.

Example 13.2 Table 8.3 showed the ranking for grain yield of 12 barley varieties grown at four macro-environmental conditions. The ranking of the varieties was different at these conditions, showing genotype × environment interaction had considerable effects.

This chapter considers the performances of genotypes in different macro-environments. Different rankings of candidates or established varieties in different macro-environments are of relevance for breeders and growers, respectively. Breeders aim to develop varieties that, averaged across a number of growing seasons, excel at least in one region or in one soil type. Exploitation of GE effects may contribute to the development of a successful varieties. It requires the breeder's imagination with regard to possible target environments of the variety to be developed (*e.g.* reduced application of pesticides). One may generalize that the selection should preferably be carried out under the growing conditions characteristic for this target environment, but the reader is reminded to some problems with respect to the topic of optimal macro-environmental conditions for selection (Section. 11.1).

Example 13.3 teaches us that one should not make statements such as 'the hybrid is heterotic', or 'variety P_1 yields better than variety P_2' without specifying the macro-environment for which the statement is made.

Growers will choose the best variety considering the growing conditions they can provide. However, the seasonal growing conditions are generally unpredictable. Descriptive lists of varieties containing useful information from variety tests assist growers when they make their choice (Example 13.4).

Example 13.3 At each of two locations the grain yield of pure line varieties P_1 and P_2 and their F_1 is considered as a function of the amount of nitrogen fertilizer (Fig. 13.1).

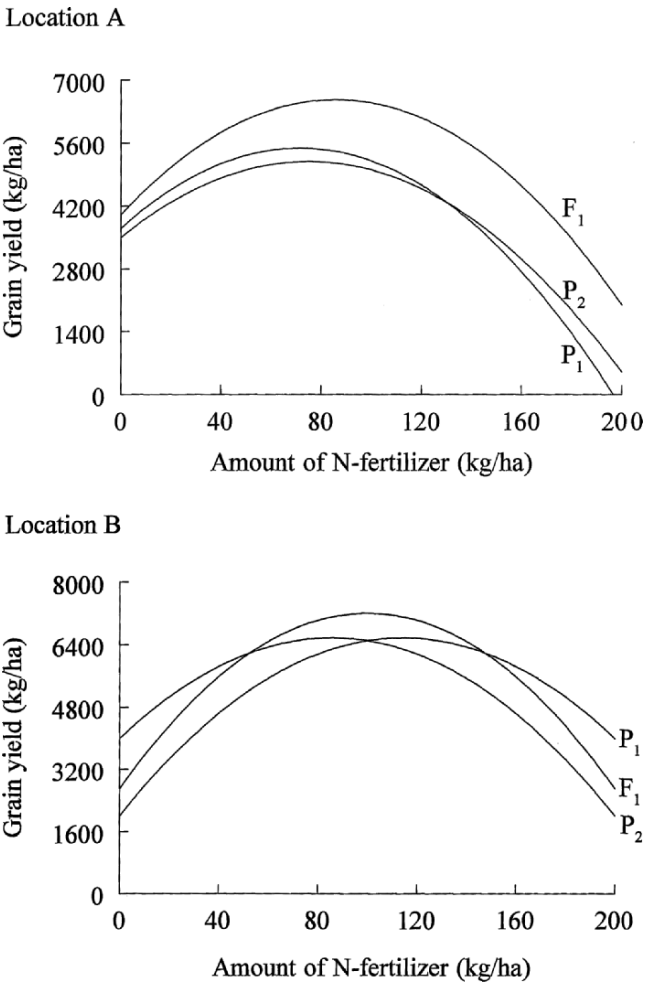


Fig. 13.1 The grain yield of three genotypes, *i.e.* P_1, P_2 and F_1, at a range of nitrogen fertilizer levels both at location A and location B

At location A the yield of the F_1 exceeds that of both parents at each level of nitrogen. At location B the yield of the F_1 exceeds those of both parents only for a small range of nitrogen fertilizer levels. For most of the other amounts it exceeds the parental mean (see also Knight, 1973).

Example 13.4 The annually issued Dutch list of varieties of arable crops presents for winter wheat varieties recommendations per soil type (clay, sand). The lists of varieties of vegetables grown in the open gives, for instance for endive, advice about cultivation period (spring, summer, early autumn, late autumn).

13.2 Stability Parameters

The phenomenon of $g \times e$ interaction is clearly experienced, by plant breeders as well as by growers, when growing the same varieties or candidate varieties ('candivars') in several macro-environments (several years and/or locations). Assume that each of J entries (inbred lines, clones, hybrids, or even open pollinating varieties) is tested in each of K environments. Then the separate observations p_{jk} do not show an clear pattern if they are arranged such that the observations from a certain macro-environment occur in a column and the observations for a certain entry in a row (Fig. 13.2).

A clearer pattern appears when applying the linear regression analysis proposed by Finlay and Wilkinson (1963): first for each macro-environment the mean performance $\overline{p}_{\cdot k}$ across the J entries is calculated. This allows ranking of the K macro-environments according to $\overline{p}_{\cdot k}$.

		Macro-environment						Mean:
		1	2		k		K	
Entry	1	p_{11}	p_{12}				p_{1K}	$\overline{p}_{1\cdot}$
	2	p_{21}	p_{22}					
	j				p_{jk}		p_{jK}	$\overline{p}_{j\cdot}$
	J	p_{J1}			p_{Jk}		p_{JK}	$\overline{p}_{J\cdot}$
Mean:		$\overline{p}_{\cdot 1}$			$\overline{p}_{\cdot k}$		$\overline{p}_{\cdot K}$	$\overline{p}_{\cdot\cdot}$

Fig. 13.2 The phenotypic value p_{jk} for some trait of entry ($j = 1, \ldots, J$) in macro-environment $k (k = 1, \ldots, K)$

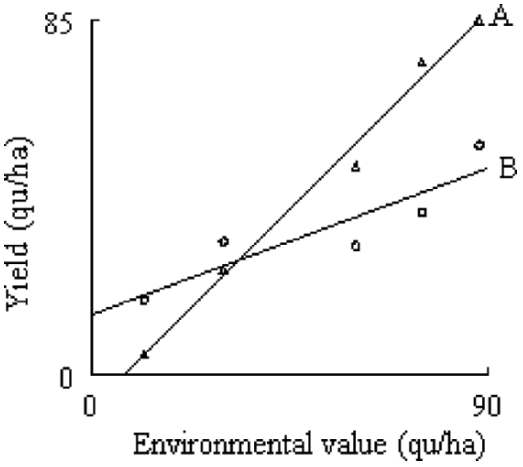


Fig. 13.3 The linear regression of the phenotypic values (p_{jk} and $p_{j'k}$) of genotypes A and B, respectively, on the environmental values ($\overline{p}_{.k}$) of $J = 5$ macro-environments

N. B. Analogous to the definition of the genotypic value one may call $\overline{p}_{.k}$ the **environmental value** of environment k. The difference $\overline{p}_{.k} - \overline{p}_{..}$ estimates E_k, the **environmental index** of environment k; see Equation (13.1). The difference $\overline{p}_{j.} - \overline{p}_{..}$ estimates G_j.

When calculating for entry j (where $j = 1$, or 2, or ..., or J) the linear regression of p_{jk} on $\overline{p}_{.k}$ across the K environments it is usually observed that the performances p_{jk} of entry j do well fit to a linear regression line (Fig. 13.3).

In connection with the preceding regression analysis, Eberhart and Russell (1966) have defined the following **stability parameters** for genotype j:

- $b_j :=$ the regression coefficient in the linear regression function $\hat{\mathcal{G}}_{jk} = a_j + b_j \overline{p}_{.k}$ and
- $s_j^2 :=$ the residual variance when applying the linear regression.

They consider genotype j to be stable if the regression analysis yields a regression coefficient b_j (nearly) equal to 1 and a residual variance ${s_j}^2$ (nearly) equal to 0. According to this 'definition' of stability one can conclude from Fig. 13.3 that genotype A is more stable than genotype B. The preceding is illustrated by Example 13.5.

Example 13.5 The analysis of $g \times e$ interaction by means of linear regression on the environmental value is illustrated on the basis of data presented by Mitchell and Lucanus (1962). These data concern the relative growth rate *(RGR)*, *i.e.* the increase of dry matter per day (in %), of $J = 3$ grass species (*Lolium perenne* L., *Holcus lanatus* L., and *Paspalum dilatatum* Poir.) at each of $K = 6$ temperatures and a photoperiod of 16 hours. The data are

		Temperature (°F)						
		45	55	65	75	85	95	$\overline{p}_{j.}$
Species:	*Lolium perenne*	9	16	18	16	11	2	12
	Holcus lanatus	9	15	19	22	15	1	13.5
	Paspalum dilatatum	5	9	17	22	25	11	14.8
	$\overline{p}_{.k}$	7.7	13.3	18.0	20.0	17.0	4.7	13.4

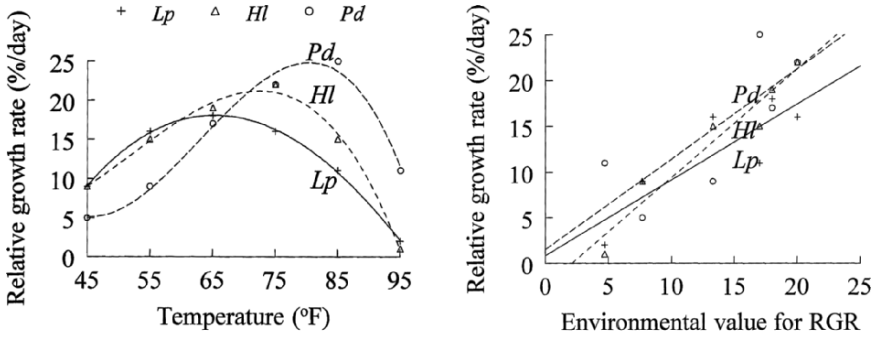


Fig. 13.4 The relation between relative growth rate and temperature (left) or environmental value (right) for the three grass species *Lolium perenne (Lp), Holcus lanatus (Hl)* and *Paspalum dilatatum (Pd)*

Figure 13.4 (left) illustrates the RGR values as a function of the temperature. Linear regression of p_{jk} on $\overline{p}_{.k}$ resulted into

$$\begin{array}{lllll} Lp\text{:} & \hat{\mathcal{G}} = 0.87 + 0.83\overline{p}_{.k} & r = 0.85 & s^2 = 12.49 \\ Hl\text{:} & \hat{\mathcal{G}} = -2.46 + 1.19\overline{p}_{.k} & r = 0.96 & s^2 = 5.12 \\ Pd\text{:} & \hat{\mathcal{G}} = 1.58 + 0.99\overline{p}_{.k} & r = 0.77 & s^2 = 30.85 \end{array}$$

Figure 13.4 (right) shows the regression lines. The intersections indicate different rankings of the grass species with regard to their RGR values at different temperatures: at 55° F *L. perenne* is the species with the greatest RGR, but at 85° F it is the species with the smallest RGR. With regard to the mean RGR across the six temperatures the three species did not differ significantly. *P. dilatatum* is the most stable species when considering the regression coefficient, but *H. lanatus* is the most stable species when considering the residual variance (see Knight, 1970, for further comments).

Genotypes complying with the mean of all tested genotypes are not necessarily stable with regard to another definition of stability, for instance the within genotype variation across the macro-environments.

In the regression analysis presented here the quantity $\overline{p}_{.k}$ is a biological measure of the quality of the macro-environment. It encompasses 'all' environmental factors. Of course, environmental and physiological conditions clearly affecting the trait of interest, *e.g.* presence or absence of a pathogen, should

be studied separately in order to see how different entries respond specifically to them and to check if responses to different factors are correlated (Caligari, 1993). If the stability of genotypes can be measured then it should be possible to study the genetic control of this attribute as a character in its own right. In *Drosophila* and in *Nicotiana* it has, indeed, been shown to be possible to manipulate by selection the expected genotypic value and the stability of the trait of interest independently.

The opinion that high degrees of heterozygosity or genetic heterogeneity induce yield stability is common. Lerner (1958; p. 100) remarked that 'heterozygotes are better buffered, *i.e.* are less responsive to environmental stresses than homozygotes, as far as traits directly related to fitness are concerned'. The usual explanation for this hypothesis is that heterozygosity acts as a buffer against environmental variation (Lynch and Walsh, 1998; p. 116). The idea was given credence by studies concerning SC-hybrids, TC-hybrids and DC-hybrids of the cross-fertilizing crops maize (Hühn and Zimmer, 1983; Schnell and Becker, 1986) and rye (Becker, Geiger and Morgenstern, 1982) and of the mainly self-fertilizing crop sorghum (Reich and Atkins, 1970; Patanothai and Atkins, 1971). However, the differences within types of hybrids were mostly larger than differences among types of hybrids: it appeared to be possible to select within each type of hybrid (very) stable hybrids. See Example 13.6.

Example 13.6 Soliman and Allard (1991) studied the hypothesis that natural selection in genetically heterogeneous populations of barley results in high-yielding, stable plant material. They used the composite cross populations CCII (generations 13, 23 and 45), CCV (generations 5, 10, 21 and 30) and CCXXI (generations 5, 9, 14 and 16); see Example 5.4.

A steady increase in grain yield over generations appeared (*e.g.* 16% increase in population CCII in 11 generations). However, the yield levels of the advanced generations were not commercially attractive and could not justify their release as heterogeneous cultivars.

The study involved also five cultivars. A regression coefficient larger than 1 was obtained for four out of the five cultivars and for only three out of the 11 CC generations. The regression coefficient deviated significantly from 1 for four cultivars and for only two CC generations. The residual variances were much higher for the cultivars than for the CC generations.

The authors assumed that genetic diversity often leads to stability at varying environmental conditions.

Faris, de Araujo and Lira (1981) studied the grain yield of several sorghum varieties in a number of macro-environments. They found a high coefficient of correlation between the mean yield across all macro-environments and the regression coefficient ($r = 0.94$). This phenomenon has been further analysed

by Hardwick (1981). It implies that a genotype with a high mean yield tends to respond better than average to high-input conditions.
Example 13.7 illustrates the phenomenon.

Example 13.7 Powell *et al.* (1986) obtained, in a study dealing with 20 doubled haploid (DH) lines and 40 single-seed-descent (SSD) lines, in some cases, *i.e.* for some traits of some crosses, a positive correlation between the mean performance across the four macro-environments and the regression coefficient or the 'environmental sensitivity'. The latter being measured as the (estimated) phenotypic standard deviation of the performances obtained in the macro-environments. It was concluded that sensitivity could be treated as a character. Univariate and bivariate cross predictions (Section 11.4), based on data from the DH lines and involving mean performance and environmental sensitivity, appeared to agree reasonably good with observed numbers of SSD lines.

When analysing $g \times e$ interaction one should realize that the interaction effects may be larger or smaller depending on the set of genotypes and the set of macro-environments. The individual contribution of each genotype to the total interaction variance or to the mean sum of squares for interaction has therefore been studied. Procedures to do this have been described by Plaisted and Peterson (1959) and by Wricke (1964), respectively.

13.3 Applications in Plant Breeding

This section gives attention to three applications of concepts introduced in the preceding two sections, namely

1. Prediction of the performance of an entry for macro-environmental conditions not earlier experienced by the entry
2. Evaluation of the relative contributions of new varieties and better cultivation practices to yield increases
3. A decision rule for acceptance or rejection of a new candidate variety

Prediction of the performance under macro-environmental conditions new to the entry of interest

If one knows for entry j the linear relationship between p_{jk} and $\overline{p}_{.k}$ *i.e.* if one knows the regression coefficients a_j and b_j, it is possible to predict the entry's genotypic value of in some macro-environment k, provided that one knows the environmental value $\overline{p}_{.k}$ of the macro-environment. The predicted genotypic value is

$$\hat{\mathcal{G}}_j = a_j + b_j \overline{p}_{.k} \tag{13.3}$$

Thus one may calculate whether a new variety is expected to perform better than a standard variety in a macro-environment where it was not yet tested. Example 13.8 illustrates the procedure.

Example 13.8 Hayward and Vivero (1984) studied perennial rye grass (*Lolium perenne* L.). They tested 25 progenies as well as the standard variety S23. Different macro-environmental conditions were provided. These consisted of combinations of two growing seasons and three different degrees of interplant competition, *viz.*

- Spaced plants ($50 \times 50\,\mathrm{cm}^2$/plant)
- Rows ($10 \times 60\,\mathrm{cm}^2$/per plant)
- Miniplots (40 seedlings at an area measuring $38 \times 50\,\mathrm{cm}^2$)

The seventh macro-environment consisted of $1 \times 2\,\mathrm{m}^2$ plots, sown at a standard rate of 25 kg/ha.

For each entry the regression coefficients a_j and b_j were calculated on the basis of data from the first six macro-environments. The yield of S23 from the $1 \times 2\,\mathrm{m}^2$ plot was used as $p_{.7}$, the environmental value characterizing macro-environment 7. The yield of progeny $j(j = 1, \ldots, 25)$ at such $1 \times 2\,\mathrm{m}^2$ plot was predicted by Equation (13.3). For 20 of the 25 progenies the difference between the predicted yield and the actual yield was insignificant.

Although Example 13.8 suggests differently, the prediction of the performance of an entry in a macro-environment where it has not yet been tested may be unreliable. This is certainly the case with a low coefficient of correlation across the entries between the regression coefficients a_j and $b_j(j = 1, \ldots, J)$ calculated for one set of macro-environments and the regression coefficients a'_j and b'_j calculated for another set of macro-environments. This is illustrated by Example 13.9.

Example 13.9 Fatunla and Frey (1976) studied the performance of 180 unselected lines of oats with different levels of phosphorus fertilizer (say: P-environments) as well with different levels of nitrogen fertilizer (say N-environments). Across the lines there was no significant coefficient of correlation between the regression coefficients calculated for the P-environments and the regression coefficients calculated for the N-environments.

Becker (1981) calculated the coefficient of correlation between the coefficients of regression obtained for the year 1979 and those obtained for the year 1980. For maize he obtained $r = 0.65^*(J = 14)$, for barley $r = -0.07(J = 18)$ and for oats $r = 0.16(J = 27)$.

The relative contributions of new varieties and better cultivation practices to yield increases

It is desirable to know whether the efforts to improve plant performance by means of plant breeding have paid off. Determination of the relative contribution of new varieties to yield increases is, therefore, of interest. Example 13.10 gives an illustration.

Example 13.10 Duvick (1992) studied the relative contribution of plant breeding to the improvement of grain yield of maize in Iowa on the basis of hybrids introduced from 1930 to 1989. He estimated the total grain yield improvement to be 100 kg/ha/year. The genetic contribution, adjusted to average on-farm yield levels, was 56 kg/ha/year, *i.e.* 56%. The genetically determined yield improvement was due to improvements in resistance to root lodging, stalk lodging, premature plant death and barrenness. New hybrids responded better to high plant densities. They were consistently superior to the older hybrids in low-yield environments. 'Selection has pre-adapted today's hybrids to lower-input agriculture and harsher growing conditions.' Duvick speculated that yields in the USA will continue to rise in the foreseeable future by about 55 kg/ha/year (≈ 1 % per year).

Due to $g \times e$ interaction it is, in fact, not easy to establish to what extend yields become higher due to improved varieties and to what extend they become higher due to improved agricultural practice. An approach to arrive at an estimate is described in Example 13.11.

Example 13.11 The difference between the value of the regression function for the present standard variety, *i.e.* the value obtained from Equation (13.3), and the value for the former standard variety, at the optimum growing conditions for the former standard variety, may be used to measure the (genetic) contribution to yield improvement. When dividing this difference by the number of years since the former standard variety was grown with the same area as occupied at present by the present standard variety one obtains the mean yearly yield improvement due to variety improvement. Pinthus (1972) calculated for the replacement of wheat variety FA 8193 by new semi-dwarf varieties an yield increase of 55–75 kg/ha/year.

A decision rule for acceptance or rejection of a new candidate variety

Now a **decision rule**, which may play a role when a decision about the acceptance or the rejection of a new candidate variety is to be made, is considered.

In an extensive trial each candidate variety may be tested in each of B blocks per location, at each of L locations per year, for each of Y years. Each 'candivar' is then evaluated on the basis of its average performance, *i.e.* its mean phenotypic value, across BLY plots. This mean phenotypic value is a

random variable with some residual variance σ^2. It can be derived that

$$\sigma^2 = \frac{\sigma_e^2}{BLY} + \frac{\sigma_{gly}^2}{LY} + \frac{\sigma_{gl}^2}{L} + \frac{\sigma_{gy}^2}{Y} \tag{13.4}$$

where

σ_e^2 := the residual variance of a single plot observation;
σ_{gly}^2 := variance due to genotype × location × year interaction effects;
σ_{gl}^2 := variance due to genotype × location interaction effects; and
σ_{gy}^2 := variance due to genotype × year interaction effects.

The optimum combination of values for B, L and Y, *i.e.* the combination of values minimizing σ^2 at a given, fixed value of BLY, has frequently been considered. Of course such an optimum can only be calculated on the basis of estimates of the relevant components of variance. Such estimates apply to a specific trait of a specific crop observed for a specific set of entries and macro-environments. Generalization is not possible. The optimum values mentioned in Example 13.12 should be considered as rough indications.

Example 13.12 Rasmussen and Lambert (1961) studied grain yield data of six commercially grown barley varieties, as obtained when grown in the years 1954, 1956, 1957 and 1958 at eight locations widely scattered in Minnesota. The estimates of the variance components were

$$\hat{\sigma}_e^2 = 42.78, \hat{\sigma}_{gly}^2 = 15.97, \hat{\sigma}_{gl}^2 = 0.22 \text{ and } \hat{\sigma}_{gy}^2 = 3.99$$

For a constant number of plots, *e.g.* $BLY = 54$, the allocation across B, L and Y which results in the smallest residual variance σ^2 will be most efficient ignoring time and costs. Considering time and costs, the most advantageous testing scheme for Minnesota was stated to be $B = 3, L = 6$ and $Y = 3$. It was observed that reduction of B, at fixed values of L and Y, yielded a relatively small increase of σ^2. Thus, practically, the optimum consists of a certain combination of values for L and Y at $B = 2$.

Schutz and Bernard (1967) concluded from a yield test of soybean, in the eastern part of the USA, that one may substitute years by locations. In their opinion, selection in a practical breeding programme is rarely based on testing for longer than two years. For L somewhere between 10 and 15, σ^2 would already be small enough to have a test with a great power, allowing elimination of low-yielding entries.

In extensive tests the yield $\underline{y}_C$ of some candidate variety C is compared to $\overline{\underline{y}}_S$, *i.e.* the mean yield of S standard varieties. The difference

$$\underline{d} = \underline{y}_C - \overline{\underline{y}}_S$$

where

$$\underline{\overline{y}}_{\mathrm{S}} = \left(\frac{\underline{y}_1 + \cdots + \underline{y}_S}{S} \right)$$

is thus considered.

The rule for the decision to accept candidate C is: C is released as a new cultivar if $\underline{d} > D$. What value should then be chosen for D, the so-called **critical difference**?

The following approach may be followed to determine D. It is assumed that $\underline{d}$ is a random variable with the normal distribution $N(\mathrm{E}\underline{d}, \sigma_d^2)$,
where

$$\mathrm{E}\underline{d} := \text{the true, but unknown difference}$$

$$\sigma_d^2 := \mathrm{var}(\underline{d}) = \mathrm{var}(\underline{y}_{\mathrm{C}} - \underline{\overline{y}}_{\mathrm{S}}) = var(\underline{y}_{\mathrm{C}}) + var(\underline{\overline{y}}_{\mathrm{S}}) = \sigma^2 \left(1 + \frac{1}{S}\right)$$

It is reasonable to require the probability $P(\underline{d} > D)$ to be small if $\mathrm{E}\underline{d} = 0$, *i.e.* if $\underline{d} = N(0, \sigma_d^2)$. In that situation the probability $P(\underline{d} > D)$ is equal to

$$P(\sigma_d \underline{Z} > D) = P(\underline{Z} > \frac{D}{\sigma_d})$$

where
Z represents the standard normal variable with the distribution $N(0, 1)$.

This permits calculation of D provided that one knows σ^2.

The requirement that $P(\underline{d} > D)$ would amount only 0.025 (or less) if the true difference is zero, implies the requirement that D/σ_d would amount to $1.96\sigma_d$ (or more). The decision rule is then as follows: C is accepted if $d > 1.96\sigma_d$; C is rejected if $d < 1.96\sigma_d$. Example 13.13 gives an illustration.

Example 13.13 Patterson *et al.* (1977) considered grain yield (in t/ha) of spring barley. They summarized 169 tests involving 26 locations, 8 years and 27 varieties. The relevant components of variance were estimated to be:

$$\hat{\sigma}_e^2 = 0.1101, \hat{\sigma}_{gly}^2 = 0.0561, \hat{\sigma}_{gl}^2 = 0.0084, \hat{\sigma}_{gy}^2 = 0.0322$$

The mean grain yield was 4.96 t/ha.

The critical difference D is calculated for a test involving $B = 3$ blocks at each of $L = 10$ locations during each of $Y = 3$ years. Then σ^2 can, according to Equation (13.4), be calculated to be equal to

$$\frac{0.1101}{90} + \frac{0.0561}{30} + \frac{0.0084}{10} + \frac{0.032}{3} = 0.0146(\mathrm{t/ha})^2$$

More than 2/3 of this, *i.e.* the amount $0.032/3 = 0.0107$, is due to genotype $\times$ year interaction. Thus, unless Y is increased, σ^2 will always be larger than 0.0107; whatever the values for B and L.

For $\mathrm{E}\underline{d} = 0, P(\underline{d} > D) = 0.025$ and $S = 2$ the critical difference is $1.96\sqrt{1.5 \times 0.0146} = 0.290\,\mathrm{t/ha}$, *i.e.* $100(0.290/4.96) = 5.8\%$ of the mean grain yield calculated for all 169 tests. Thus, if the true superiority of C compared to two standard varieties is nil, the probability that $\underline{y}_{\mathrm{C}}$ yields at least 0.29 t/ha higher than the average of the standard varieties (and that C is consequently accepted for release) is only 0.025.

Chapter 14
Selection with Regard to a Trait with Quantitative Variation

In Chapter 6 the topic of selection was introduced. The contrast between natural selection and artificial selection was elaborated. Methods for artificial selection with regard to a trait with qualitative variation were considered in an order of decreasing efficiency. Special attention was given to the question whether the trait under selection is expressed before or after pollen distribution.

The present chapter pays attention to selection with regard to a trait with quantitative variation. The selection is aimed at improving the average genotypic value, but the actual selection among the entries (plants, clones, lines, hybrids, families) is based on the phenotypic values of the candidates. In order to improve the efficiency of such indirect selection the breeder should make efforts to evaluate the candidates in such a way that the coefficient of correlation between the phenotypic value (the auxiliary trait) and the genotypic value (the target trait) is as high as possible. Procedures to promote this are elaborated in the present chapter. Such procedures can be classified according to the way of evaluation of the candidate genotype. Section 14.2 gives attention to procedures concerning the evaluation of entries on the basis of single plants (single plant evaluation); Section 14.3 considers procedures for the evaluation of candidates on the basis of plots such that the plants occurring in a certain plot represent a certain candidate (plot evaluation).

14.1 Disclosure of Genotypic Values in the Case of A Trend in the Quality of the Growing Conditions

In order apply selection successfully a breeder should be able to identify candidates with a superior genotype. This requires disclosure of genotypic values hidden beyond the observable phenotypic values. Variation of the phenotypic values is due to variation of genotypic values as well as variation in the growing conditions. The effects of variation in growing conditions appear from the phenotypic variation exhibited by a clone, a pure line or a single cross hybrid. It appears from the (phenotypic) variation shown in the absence of genetic variation. In Section 8.2 the latter was called environmental variation, measured by the environmental variance. The variation may show up as plant-to-plant variation or as plot-to-plot variation.

An entry is selected when showing an attractive, superior phenotypic value. This may be due to a superior genotype and/or to favourable growing conditions. Variation in the quality of the growing conditions detracts, consequently,

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 339–379.

from the ability of the breeder to identify candidates with a superior genotype. The quality of the growing conditions may vary at random or it may vary with a trend.

In this book it is preferred to use the term 'quality of the growing conditions' rather than the term 'soil fertility'. The reason for this is that the former term is considered to have a broader scope than the second term. It includes, for instance, the presence or the absence of shade or of soil-born pests. **Random environmental variation** includes thus not only plant-to-plant or plot-to-plot variation in soil fertility, but also variation in the strength of the competition experienced. With homogeneous soil fertility the latter depends on the amount of seed sown (both in kg/ha and in number of seeds per ha) as well as seed quality (as determined by seed size and germination ability). Variation for these factors may induce variation in the date of emergence and subsequently give rise to variation in the number and the size of neighbouring plants, occurring within a certain circle around the considered plant(s). The detrimental effects of interplant or interplot competition on the efficiency of selection are considered in Chapter 15.

In some cases the degree of damage caused by nematodes, insects, snails, slugs, birds, mice, *etc.*, could be considered as random environmental variation, but often the degree of such damage will vary with a trend.

Environmental variation with a trend implies a gradient in the field, this may be soil fertility, moisture content or the abundance of soil-borne pathogens and pests (e.g. nematodes). The contribution to the phenotypic value which is due to a trend in the quality of the growing conditions can be eliminated from the phenotypic value, given the right adjustment procedure. The means that the phenotypic value observed for some candidate, say p, is adjusted into a phenotypic value which takes the trend in the quality of the growing conditions into account, say p'.

The type of adjustment depends on whether single plants are evaluated or whether group of plants (clones, lines, families or hybrids) are evaluated by means of plots. In any case an adjustment is sought such that the coefficient of correlation of $\underline{p}'$ and the genotypic value ($\underline{\mathcal{G}}$) of the considered candidates is higher than the coefficient of correlation of $\underline{p}$ and $\underline{\mathcal{G}}$. The adjustment procedure is thus required to yield

$$\rho_{p',g} > \rho_{p,g} \tag{14.1}$$

According to Note 11.2 this means that the adjustment procedure should succeed in yielding an increase of the heritability in the broad sense, *viz.*

$$h_w{}^2(p') > h_w{}^2(p) \tag{14.2}$$

In Section 14.2 procedures that aim to eliminate the contribution due to a plant-to-plant trend in the quality of the growing conditions are considered in connection with single-plant evaluation. In Section 14.3 procedures aimed at eliminating the effect of a plot-to-plot trend in the quality of the growing conditions are considered in connection with plot evaluation (of clones, lines or families tested at ≥ 1 plot).

The environmental variation remaining after adjustment of the phenotypic values with regard to a trend in the quality of the growing conditions is random environmental variation. This cannot be eliminated as a cause for inefficient selection.

14.2 Single-Plant Evaluation

Single-plant evaluation is an essential part of mass selection. In Sections 6.3.1 and 6.3.5 mass selection for qualitative variation in a trait was considered. Here mass selection with regard to quantitative variation in a trait is examined. In the present context the two features of mass selection can be described as follows:

1. Depending on its phenotypic value, possibly in comparison to the phenotypic values of (nearby) other plants, every plant is either selected or rejected. The phenotypic value may be carefully measured, mechanically determined or visually assessed.
2. The selected plants produce seeds by their natural mode of reproduction. These seeds are bulked.

Attractive features of mass selection are (Lonnquist, 1964)

- The practical feasibility of the procedure is unsurpassable.
- Selection can be applied in each generation. The progress per generation obtained when applying mass selection may therefore be high compared to than the progress per generation obtained with selection procedures requiring more than one generation per 'cycle'.
- A high selection intensity (Section 11.1) does not necessarily imply that a large (random) change in allele frequencies, for loci not affected by the selection, have a high probability. This is due to the fact that, for a given cost, the number of candidates that can be evaluated as single plants is much higher than the number of candidates that can be evaluated by means of plots. In other words: with mass selection a high selection intensity does, because of the high number of selected candidates, not necessarily imply a low effective population size (Section 7.2).

A weak point of mass selection is the following. In the case of single-plant evaluation the phenotypic value of an individual plant poorly reflects its genotypic value. The correlation of the phenotypic value $\underline{p}$ of a plants and its genotypic value $\underline{\mathcal{G}}$, *i.e.* $\rho(\underline{p}, \underline{\mathcal{G}})$, tends to be low. The efficiency of mass selection tends thus to be low (h^2 is low) as compared to the efficiency of selection procedures based on evaluation of the candidates by means of plots. This is an important drawback of mass selection. Mass selection does not easily lead to complete exploitation of genetic variation for loci affecting the trait under selection pressure.

The efficiency of mass selection is thus at its minimum if the selection does not involve an attempt to eliminate the effect of a plant-to-plant trend in the quality of the growing conditions. In order to improve its efficiency one may adjust the phenotypic values for this type of variation in the quality of the growing conditions. Thus instead of selection for $\underline{p}$, the selection will be for $\underline{p}'$. As was indicated in the previous section (Equation (14.1)), this adjustment is required to yield $\rho(\underline{p}', \underline{\mathcal{G}}) > \rho(\underline{p}, \underline{\mathcal{G}})$. Some simple adjustment procedures will now be elaborated. The adjustments consist simply of comparison of the phenotypic values of the candidate plants with the phenotypic values of other, unrelated plants, comparison with

- all other candidate plants: **truncation selection** (this section);
- nearby plants representing some standard variety (Section 14.2.1);
- nearby candidate plants: **fixed-grid selection** (Section 14.2.2) and
- direct neighbouring candidate plants: **moving-grid selection** (Section 14.2.3).

Adjustments of phenotypic values aiming at elimination of the contributions due to random environmental variation or to genotype × micro-environment interaction are bound to fail. Random environmental variation comprises variation in the strength of interplant competition (possibly due to variation in seed size). It is also due to irregular occurrence of puddles, clods, stones, leaves, weeds, snails, slugs, mice, rabbits, *etc.*

When considering the response to mass selection, it is useful to distinguish short-term and long-term responses. Due to its low efficiency, short-term responses tend to be of minor importance. The improvement attained in one generation can be annihilated in the next generation. The plants selected in one generation may involve a sample of genotypes different from the genotypes of the plants selected in the previous or the next generation. This is due to the fact that growing conditions vary from one generation to the next. The locations (and thus the soil conditions) may vary, but seasonal conditions will certainly vary. Due to effects of genotype × macro-environment interaction plants selected in one macro-environment (season) may represent genotypes that would be neglected when considering their phenotypes in another macro-environment (season).

Thus mass selection does not tend to give rise to apparent progress from one generation to the next. Long-term responses may, nevertheless, be impressive. Application of mass selection for a large number of generations may yield considerable changes in plant appearance. This is illustrated by Example 14.1.

Example 14.1 Some form of mass selection will have been applied since the onset of plant domestication. Great long-term effects appear from the enormous differences between present-day ears of maize and the eldest, subfossilic ears found in southern Mexico, dating from about 5200 BC. Not only has ear size been increased considerably, but also the area where maize cultivation can be successfully carried out.

Another example is the high sugar content (about 17%) of present-day sugar beet. The sugar beet grown around AD 1825 had a sugar content of only about 7.5%.

These impressive results are due to the long-lasting continuation of the selection. The domestication of the sugar beet is, nevertheless, a relatively recent development. Oil palm is a still more recently developed crop; especially when one counts the number of generations since the onset of the domestication. Responses to long-term selection experiments in maize have been described by Dudley, Lambert and Alexander (1974) (Example 8.4), and Gardner (1978).

14.2.1 Use of Plants Representing a Standard Variety

If the plants belonging to the population subjected to selection are grown in rows, one may insert at certain plant positions within the rows plants representing standard variety S. The adjustment consists then of calculating:

$$\underline{p}' = \underline{p} - \underline{\bar{p}}_{\mathrm{s}} \tag{14.3}$$

where

$\underline{p}$:= the phenotypic value of the considered candidate plant

$\underline{\bar{p}}_{\mathrm{s}}$:= the mean phenotypic value calculated across nearby plants representing the standard variety S.

After having adjusted the phenotypic values in this way, truncation selection for $\underline{p}'$ is applied. Plants with an adjusted phenotypic value exceeding the estimated genotypic value of the standard variety, as represented by plants grown at similar conditions, as much as possible are then selected.

This adjustment requires availability of at least one genetically uniform standard variety. It also requires extra efforts and space. Especially because of the latter requirement, this adjustment procedure is not applied in the case of single-plant evaluation. The authors, at least, are not aware of its application. It is, however, commonly applied when selecting among clones, lines or families (where the candidates are evaluated by means of plots, see Section 14.3.2).

14.2.2 Use of Fixed Grids

The detrimental effect on the efficiency of mass selection of variation in the quality of the growing conditions can be (partly) eliminated by dividing the selection field into parts such that the growing conditions are more uniform

Table 14.1 The analysis of variance of data obtained from the K plants sampled from each of G grids, each containing K' plants

Source of variation	df	SS	MS	E(MS)
Between grids	$G-1$	SS_b	MS_b	${\sigma_w}^2 + K{\sigma_b}^2$
Within grids	$G(K-1)$	SS_w	MS_w	${\sigma_w}^2$

within each of these so-called **grids** than across the whole field. The breeder may then select the most attractive plants within each grid. This selection procedure is called **grid selection**. The improvement of the efficiency of the selection appears from the increase of the heritability. This is elaborated as follows.

Assume that K plants are sampled from each of the G grids containing K' plants. The structure of the analysis of variance of the phenotypic values is then described by Table 14.1. The heritability will now be considered both for truncation selection (selection of the most attractive plants across the whole selection field) and for grid selection (selection of the most attractive plants within each of the grids). It will appear that the heritability can be described in terms of the variance components ${\sigma_w}^2$ and ${\sigma_b}^2$ occurring in the E(MS) column of Table 14.1.

In order to develop some theory it is useful to define the following variances:

- $\mathrm{var}(\underline{\mathcal{G}})$:=the variance of the genotypic values of individual plants
- $\mathrm{var}(\underline{e}_{(G)})$:=the environmental variance of individual plants within a grid
- $\mathrm{var}(\underline{e}_G)$:=the variance of the environmental conditions provided by the grids

The quantitative genetic interpretations of ${\sigma_w}^2$ and ${\sigma_b}^2$ depend on the involved plant material:

(i) if all GK plants would have the same genotype, the quantitative genetic interpretation of ${\sigma_w}^2$ would be: $\mathrm{var}(\underline{e}_{(G)})$; for ${\sigma_b}^2$ it would be: $\mathrm{var}(\underline{e}_G)$;
(ii) if each of the G grids would contain the same sample of K genotypes the quantitative genetic interpretation of ${\sigma_w}^2$ would be: $\mathrm{var}(\underline{e}_{(G)}) + \mathrm{var}(\underline{\mathcal{G}})$; for ${\sigma_b}^2$ it would be: $\mathrm{var}(\underline{e}_G)$;
(iii) if each of the G grids would contain a different sample of K genotypes the quantitative genetic interpretation of ${\sigma_w}^2$ would be: $\mathrm{var}(\underline{e}_{(G)}) + \mathrm{var}(\underline{\mathcal{G}})$; for ${\sigma_b}^2$ it would be: $\mathrm{var}(\underline{e}_G) + \mathrm{var}(\underline{\mathcal{G}})/K$.

Situation (iii) represents, of course, the actual situation in the breeder's practice. If the grids contained many plants, *i.e.* if $K \to \infty$, then the contribution of $\mathrm{var}(\underline{\mathcal{G}})/K$ to ${\sigma_b}^2$ would approach zero, implying that ${\sigma_b}^2$ is then asymptotically equal to $\mathrm{var}(\underline{e}_G)$.

The broad sense heritability as applying to the situation of truncation selection, say ${h_w}^2(T)$, is

$$\frac{\mathrm{var}(\underline{\mathcal{G}})}{\mathrm{var}(\underline{\mathcal{G}}) + \mathrm{var}(\underline{e})} \tag{14.4}$$

where $\mathrm{var}(\underline{e}) = \mathrm{var}(\underline{e}_{(\mathrm{G})}) + \mathrm{var}(\underline{e}_{\mathrm{G}})$.
The broad sense heritability as applying to the situation of grid selection, say ${h_w}^2(\mathrm{G})$, is

$$\frac{\mathrm{var}(\underline{\mathcal{G}})}{\mathrm{var}(\underline{\mathcal{G}}) + \mathrm{var}(\underline{e}_{(\mathrm{G})})} \tag{14.5}$$

If ${\sigma_{\mathrm{b}}}^2 > 0$, *i.e.* if $\mathrm{var}(\underline{e}_{\mathrm{G}}) + \mathrm{var}(\underline{\mathcal{G}})/K > 0$, the within grid heritability ${h_w}^2(\mathrm{G})$ is higher than the across grids heritability ${h_w}^2(\mathrm{T})$ and then grid selection will be more efficient than truncation selection. Testing of the null hypothesis H_0: '${\sigma_{\mathrm{b}}}^2 = 0$' against the alternative hypothesis H_{a}: '${\sigma_{\mathrm{b}}}^2 > 0$' is thus recommended before choosing between truncation selection and grid selection.

Several publications have reported positively on grid selection, *e.g.* Gardner (1961, 1978) dealing with maize (Example 14.2), and Verhalen, Baker and McNew (1975) dealing with cotton. These positive experiences stimulated the application of grid selection, especially in the USA, where – because of the success of hybrid maize – since 1925 no attention had been given to improvement of the efficiency of traditional mass selection, *i.e.* truncation selection.

Example 14.2 Gardner (1961) applied grid selection within the open-pollinating maize variety Hays Golden. Grids consisting of 40 plants were used. In each grid the four best-yielding plants were selected. Application of grid selection for four generations raised the yield from 79.3 bushels/acre to 97.4 bushels/acre (1 bushel/acre corresponds to about 63 kg/ha). The response to the selection varied considerably across the generations. The average response per generation was presented in two ways:

1. As the coefficient of regression of the relative yield, *i.e.* the yield expressed as a percentage of the yield of the original variety, on the generation of selection. This amounted to a response of 3.9% per generation.
2. As the geometric mean response per generation. The total response across four generations amounted to 22.8%; the geometric mean response per generation was thus 5.3%.

Although grid selection tends to be more efficient than truncation selection, if indeed ${\sigma_{\mathrm{b}}}^2 > 0$, it sometimes yielded inferior results (Example 14.3).

Example 14.3 After applying grid selection for five or six generations, Hallauer and Sears (1969) could not establish a significant response with regard to grain yield of maize. They assumed that this was due to the fact that the response was evaluated at three locations, differing from the location where the selection was applied.

Bos (1981, p. 56 and p. 73) applied, in a population of winter rye, grid selection as well as truncation selection. For the selection in 1976 he obtained as response for grain yield 4% and 6.5%, respectively; for 1977 these figures were 0% and 6%.

In the preceding studies the null hypothesis H_0: '$\sigma_b{}^2 = 0$' was not tested. For an experiment with autotetraploid winter rye Bos (1981, p. 153) established highly significant differences among grids when considering culm length, number of ears, grain yield and straw weight.

This may be due to two causes (Bos, 1983b):

1. Selection of a fixed number of plants, *i.e.* k plants, in each grid, or
2. A sub-optimal division of the selection field into grids.

In order to avoid these causes for getting a disappointing result of grid selection it is suggested:

1. To select within the grids a variable number of candidates and
2. To give attention to a proper way of division of the selection field into grids

The number of plants selected in each grid

The number of plants in a grid having a superior genotype, and therefore deserving to be selected, is a random variable $\underline{k}$, possibly with a Poisson distribution. According to the procedure originally suggested for grid selection, however, a fixed number of plants (say k) was to be selected in each grid. For some grids this number will be larger than the number of candidate plants deserving selection and for some other grids it will be smaller. Grid selection may, for this reason, give rise to a lower response than truncation selection, especially if $\sigma_b{}^2 \approx 0$.

To avoid this drawback of selecting of a fixed number of plants in each grid, a modification of the originally described procedure for grid selection is suggested. After division of the selection field into grids, the phenotypic values ($\underline{p}$) of the candidate plants are transformed into adjusted phenotypic values ($\underline{p}'$). This is followed by truncation selection, across the whole population (*i.e.* across all grids), for with regard to $\underline{p}'$. This induces the number of plants selected in a grid to vary from grid to grid.

Two related, grid specific, transformations are suggested:

1.
$$\underline{p}'_{ij} = \underline{p}_{ij} - \overline{\underline{p}}_i \tag{14.6}$$

and

2.
$$\underline{p}'_{ij} = \frac{\underline{p}_{ij} - \overline{\underline{p}}_i}{\underline{s}_i} \tag{14.7}$$

where

$\underline{p}_{ij}$:= the phenotypic value of plant j in grid i
$\overline{\underline{p}}_i$:= the mean phenotypic value across the plants in grid i (This quantity estimates the quality of the growing conditions provided by grid i)

$\underline{s}_i$:= (the estimator of) the phenotypic standard deviation of the plants in grid i

The transformations eliminate, by calculating $\underline{p}_{ij} - \overline{\underline{p}}_i$, the grid-to-grid variation in the quality of the growing conditions as a source of environmental variation.

The effect of variation between grids with regard to ${\sigma_w}^2$, *i.e.* with regard to $\mathrm{var}(\underline{e}_{(G)}) + \mathrm{var}(\underline{\mathcal{G}})$, is now considered. If ${\sigma_w}^2$ is small because of a small value of $\mathrm{var}(\underline{e}_{(G)})$, the efficiency of within-grid selection is high (see Equation (14.5)). If ${\sigma_w}^2$ is large because of a large value of $\mathrm{var}(\underline{e}_{(G)})$ it will be low. Transformation according to Equation (14.6) will, however, result in selection of relatively small numbers of plants in homogeneous grids (${\sigma_w}^2$ small) and in selection of relatively high numbers of plants in heterogeneous grids (${\sigma_w}^2$ large). Indeed, at equal values for $\overline{\underline{p}}_i$, the probability $P(\underline{p}_{ij} - \overline{\underline{p}}_i > p_{min})$ will be larger for a heterogeneous grid, *i.e.* a grid with a large value for $\underline{s}_i$, than for a homogeneous grid. This transformation tends thus to give rise to a positive relation between $\mathrm{var}(\underline{e}_{(G)})$ and the number of plants selected in the involved grid. Transformation according to Equation (14.7) aims to avoid this dependence of the number of plants selected from a grid on the environmental variance within the grid. It is thus appropriate when grids vary with regard to $\mathrm{var}(\underline{e}_{(G)}) + \mathrm{var}(\underline{\mathcal{G}})$.

In a statistical context transformation according to Equation (14.7) implies calculation of t values. Bos (1983b) and Casler (1992) reported more or less positively about selection of plants with high within-grid t values.

The division of the field into grids

In order to promote the efficiency of grid selection the borders of the grids should be chosen such that ${\sigma_b}^2$, in fact especially $\mathrm{var}(\underline{e}_G)$, is as large as possible. However, clear-cut changes in the quality of the growing conditions do rarely show up. Thus it is mostly impossible to choose borders coinciding with such lines of demarcation. On the other hand, it is certain that an arbitrary division of the selection field in grids is sub-optimal. Reliable and easy to apply guidelines for an optimal choice of the size, the shape and the orientation of the grids are as yet not available. Weber and Stam (1988) and Weinbaum *et al.* (1990) considered the problem of how to determine the optimum grid size. Here only a few aspects concerning the optimum grid size, in terms of area as well as the number of plants per grid, are considered.

The environmental variance consists of $\mathrm{var}(\underline{e}_{(G)}) + \mathrm{var}(\underline{e}_G)$. With smaller grids variance component $\mathrm{var}(\underline{e}_G)$ will tend to become smaller and component $\mathrm{var}(\underline{e}_G)$ will tend to become larger. The efficiency of grid selection will thus tend to improve with smaller grids. However, one should not draw the conclusion that the grids should be as small as possible. Smaller grids contain a smaller number of plants. The mean phenotypic value $\overline{\underline{p}}_i$ of the plants

occurring within grid i becomes then a less accurate estimator of the quality of the growing conditions provided by the grid.

When large grids are used $\mathrm{var}(\underline{e}_{(\mathrm{G})})$ will be large. This implies a relatively low efficiency of grid selection. When using grids containing many plants the grids will hardly vary with regard to $\mathrm{var}(\underline{\mathcal{G}})$. Grid to grid variation in ${\sigma_{\mathrm{w}}}^2$ must then be due to variation in $\mathrm{var}(\underline{e}_{(\mathrm{G})})$. The second transformation seems thus appropriate when using grids containing many plants.

14.2.3 Use of Moving Grids

In the preceding section it was remarked that borders of grids seldomly coincide with clear-cut changes of the quality of the growing conditions. An alternative approach to marking out grids, *i.e.* to form groups of candidate plants subjected to more or less identical growing conditions, is to consider each plant in its turn as the centre of a grid. Each of the so-called **moving grids** comprises then a number of plants, say K, subjected to growing conditions similar to those experienced by the central plant.

In connection with the use of moving grids the breeder may select according to two options:

1. The breeder selects the plants with the highest adjusted phenotypic values
2. The breeder selects central plants surpassing all other plants in their grid

Selection of the plants with the highest adjusted phenotypic values

The phenotypic value of candidate plant i is adjusted by calculating

$$\underline{p}_i{}' = \underline{p}_i - \overline{\underline{p}}_i \tag{14.8}$$

where

$\underline{p}_i :=$ the phenotypic value of candidate plant i
$\overline{\underline{p}}_i :=$ the mean phenotypic value of the plants occurring in grid i, *i.e.* the group of plants among which plant i occupies the central position

One may question whether $\overline{\underline{p}}_i$ should be calculated across all K plants in the grid or across the $K' = K - 1$ neighbours of the candidate plant in the centre of the grid. When defining

$S :=$ the sum of the phenotypic values of all K plants in the grid
$S' := S - p$, *i.e.* the sum of the phenotypic values of across the K' neighbours of the central plant

Equation (14.8) implies

$$p' = p - \frac{S}{K} = \frac{Kp - S}{K} = \frac{K'p - (S - p)}{K} = \frac{K'p - S'}{K} = \frac{K'}{K}\left(p - \frac{S'}{K'}\right)$$

Thus, except for the constant coefficient K'/K, which is close to 1 the two ways of calculating p' are equivalent. With regard to the ranking of the adjusted phenotypic values it does not matter whether or not $\overline{p}_i$ comprises the considered central plant.

An alternative to the adjustment according to Equation (14.8) is developed as follows. Linear regression of $\underline{p}_i$ on $\overline{\underline{p}}_i$ is applied to predict the phenotypic value of plant i on the basis of $\overline{p}_i$. The phenotypic value predicted for a conceptual plant occurring in the centre of grid i amounts then to

$$\hat{p}_i = a + b\overline{p}_i \tag{14.9}$$

where

$a :=$ the intercept

$b :=$ the slope of the linear regression line

The adjusted phenotypic value of candidate plant i may then be calculated as

$$\underline{p}_i{}' = \underline{p}_i - \hat{\underline{p}}_i = \underline{p}_i - (a + b\overline{\underline{p}}_i) \tag{14.10a}$$

Because the intercept a is a constant it suffices to calculate

$$\underline{p}_i{}' = \underline{p}_i - b\overline{\underline{p}}_i \tag{14.10b}$$

Adjustment according to Equation (14.8) is, of course, a special case of the adjustment according to Equation (14.10).

N. B. 1. One may argue that the regression of $\underline{p}_i$ on $\overline{\underline{p}}_i$ should be regression through the origin, *i.e.*

$$\hat{\underline{p}}_i = b\overline{\underline{p}}_i,$$

instead of regression with an intercept (Equation (14.9)). Then Equations (14.10a) and (14.10b) can not even contain an intercept.

N. B. 2. Instead of the adjustment given by Expression (14.8) one may consider the transformation

$$\underline{p}_i{}' = 100\left(\frac{\underline{p}_i}{\overline{\underline{p}}_i}\right) \tag{14.11}$$

The transformations aim to adjust the phenotypic values of the candidate plants for the effect of the quality of the growing conditions provided by the grid, estimated by $\overline{\underline{p}}_i$. The so-called **moving mean adjustment** results in detrending, *i.e.* in elimination of the contributions to the phenotypic values which are due to a trend in the quality of the growing conditions. After having

adjusted the phenotypic values, all candidate plants are subjected to truncation selection with regard to $\underline{p}'$.
In the remainder of this section some comments with regard to the efficiency of adjustment procedures are elaborated. It is reminded that the adjustments pursue

$$\rho_{p',g}$$

to be higher than

$$\rho_{p,g}$$

(see Section 14.1).
An adjustment intends to increase the heritability by means of a reduction of the environmental variance. Thus the adjustment aims at

$$\mathrm{var}(\underline{e}') < \mathrm{var}(\underline{e}) \tag{14.12}$$

where

$\mathrm{var}(\underline{e}')$:= the environmental variance after adjustment, and

$\mathrm{var}(\underline{e})$:= the environmental variance in the absence of adjustment.

Yates (1936) remarked with regard to the adjustment according to Equation (14.11) that 'percentages are unlikely to possess any advantages over differences'. He considered adjustment according to Equation (14.8) to be less effective than adjustment according to Equation (14.10).

Spitters (1979, p. 201) showed how one may develop insight with regard to the topic of efficiency of adjustment procedures. He studied the quality of the moving mean adjustment involving plants occurring alongside a row. The adjustment was based on the two nearest neighbours. Thus

$$\underline{p}_i{}' = \underline{p}_i - \tfrac{1}{2}(\underline{p}_{i-1} + \underline{p}_{i+1})$$

where p_{i-1}, p_i and p_{i+1} designate the phenotypic values of three contiguous plants. This means

$$var(\underline{p}') = \tfrac{3}{2}var(\underline{p}) - 2cov(\underline{p}_i, \underline{p}_{i+1}) + \tfrac{1}{2}cov(\underline{p}_{i-1}, \underline{p}_{i+1}) \tag{14.13}$$

In the absence of genetic variation the environmental variance after adjustment amounts thus to

$$\tfrac{3}{2}var(\underline{e}) - 2cov(\underline{e}_i, \underline{e}_{i+1}) + \tfrac{1}{2}cov(\underline{e}_{i-1}, \underline{e}_{i+1}) \tag{14.14}$$

The adjustment leads to an increase of the environmental variance because of

(i) the environmental variances of the neighbours, as well as
(ii) the covariance of the environmental deviations of the neighbours.

It leads to a decrease of the environmental variance in the case of a positive covariance of the environmental deviations of adjacent plants. This covariance may tend to be positive because neighbours are subjected to about the same growing conditions, but it might be negative due to interplant competition.

The adjustment will thus not always give rise to a reduction of the environmental variance.

Spitters also considered the variance of the adjusted environmental deviations in the presence of genetic variation. For that case $\mathrm{var}(\underline{e}')$ is increased by a genetic component arising from the genetic variation among the plants involved in the moving grid. He derived that the pursued goal of adjustment, *i.e.*

$$\mathrm{var}(\underline{e}') < \mathrm{var}(\underline{e}),$$

applies if

$$\rho_{p_i,p_{i+1}} - \tfrac{1}{4}\rho_{p_{i-1},p_{i+1}} > \tfrac{1}{4}$$

This implies that successful adjustment requires a rather strong coefficient of correlation of the phenotypic values of direct neighbours. Otherwise the adjustment gives rise to over-correction of the phenotypic values.

Bos (1981, p. 145) derived, for a regular triangular pattern of plant positions, that the condition $\mathrm{var}(\underline{e}') < \mathrm{var}(\underline{e})$ applies if

$$\rho_{p_i,\overline{p}_i} > \tfrac{1}{2}\left(\frac{\sigma_{\overline{p}_i}}{\sigma_{p_i}}\right)$$

Example 14.4 shows interplant competition to be a disturbing factor with regard to the efficiency of moving grid adjustment.

Example 14.4 Bos and Hennink (1991) performed an experimental verification of the merits of the adjustments according to Equations (14.8) and (14.10). They did so by studying the relationship between the adjusted phenotypic values of candidate plants ($\underline{p}'$) and the mean phenotypic values, calculated across the offspring of these plants obtained by open pollination ($\underline{p}_{\mathrm{HS}}$). If it is true that $\underline{p}$ reveals the underlying genotypic value $\underline{\mathcal{G}}$ worse than $\underline{p}'$, then the coefficient of correlation between $\underline{p}$ and $\underline{p}_{\mathrm{HS}}$ ($\rho_{p,p\mathrm{HS}}$) will tend to be lower than the coefficient of correlation between $\underline{p}'$ and $\underline{p}_{\mathrm{HS}}$ ($\rho_{p',p\mathrm{HS}}$).

This tendency did, however, not show up in the verification for traits like height or grain yield of winter rye plants. The candidate plants were grown in a regular triangular pattern of plant positions with an interplant distance of 15 cm (implying a plant density of 51.3 plants per m^2). For plant height, for instance, Bos and Hennink got $\mathrm{r}_{p,p\mathrm{HS}} = 0.47 (n = 269)$. With Equation (14.8) and (14.10) they got $\mathrm{r}_{p',p\mathrm{HS}} = 0.43$ and 0.45, respectively.

Bos and Hennink (1991) assumed that interplant competition was the main cause for the failure of the adjustments according to Equations (14.8) and (14.10) to attain their goal, *i.e.*

$$\rho_{p',g} > \rho_{p,g}$$

(Section 14.1).
They concluded that these adjustments can only be effective in the virtual absence of intergenotypic competition. This requires, for single plant selection,

an extremely low plant density, whereas selection among candidates tested by means of plots would require evaluation by means of multi-row plots.

Example 14.4 illustrates that the coefficient of correlation between $\underline{p}_i$ and $\overline{\underline{p}}_i$ deserves attention. A moving mean adjustment aims to eliminate the contribution to the phenotypic value that is due to a trend in the quality of the growing conditions. Such a trend may become apparent from estimates for $\rho_{p_i,\overline{p}_i}$. In the absence of interplant competition, *i.e.* at a (very) low plant density, the correlation is expected to be positive (if, indeed, a trend in the quality of the growing conditions is present). At high plant density the actual value of the coefficient of correlation indicates the balance between a positive effect on ρ due to a trend in the quality of the growing conditions and a negative effect due to interplant competition. Example 14.5 presents some estimates for $\rho_{p_i,\overline{p}_i}$.

Example 14.5 Kira, Ogawa and Sakazaki (1953) estimated $\rho_{p,\overline{p}}$ for 1-aureole grids (see Fig. 14.1). This was done for plant weight of soybeans at each of four different plant densities. Table 14.2 presents some of their estimates.

Table 14.2 The coefficient of correlation between the weight of soybean plants and the average weight of their six neighbours (source: Kira, Ogawa and Sakazaki, 1953)

		Plant density (plants/m^2)				
		28.9	51.3	115.5	461.9	Mean
Days after sowing:	12	0.15	0.02	-0.22^{*}	0.12	0.02
	31	0.58	0.16	0.65^{**}	0.21^{*}	0.40
	84	0.29	0.74^{**}	0.55^{**}		0.53
	Mean	0.34	0.31	0.33	0.17	

Many of these estimates were not statistically significant, especially at 12 days after sowing. There was, however, a tendency to get lower coefficients of correlation at higher plant densities, *i.e.* at stronger interplant competition. The stronger competition at a later ontogenetic stage did not give rise to a decrease of the coefficient of correlation.

Fasoulas (1981, p.71) grew maize plants at a density of only 0.8 plants/m^2. He estimated the coefficient of correlation between the yield of the central plant and the mean yield of all seven plants in the 1-aureole grid. Very high estimates were reported, *viz.* $r = 0.64$ and $r = 0.91$. These estimates were, of course, inflated by the inclusion in $\overline{p}$ of the phenotypic value of the central plant.

For winter rye plants grown at a density of 51.3, Bos (1981, p. 145) obtained for 1-aureole grids for culm length $r = 0.29$, and for grain yield $r = 0.28$. The estimates were higher as the considered area, across which $\rho_{p,\overline{p}}$ was estimated, was larger.

Bos and Hennink (1991) estimated $\rho_{p,\overline{p}}$ for grain yield of winter rye plants grown in a regular triangular pattern of plant positions with an interplant distance of 15 cm. They obtained $r_{p,\overline{p}} = 0.06$ (ns, $n = 269$). This low coefficient was explained by assuming an important negative effect due to interplant competition. They estimated the coefficient of correlation between grain yield per plant and the number of plants in the involved 3-aureole grid (thus $K \leq 19$): $r_{p,K} = -0.29$. Plants with a smaller number of neighbours tend thus to perform better than plants with a higher number of neighbours. This implies that the number of neighbours of candidate plants should be taken into account when applying some mass selection method.

Selection of central plants surpassing all other plants in their grid

The breeder may decide to select central plants surpassing all other plants in their grid. Application of this principle is considered in connection with the use of a regular triangular pattern of plant positions. Fasoulas (1973) introduced for this procedure the term **honeycomb selection**: each plant in its turn is compared with neighbours occurring in the grid formed by the central plant and a number of surrounding plants. These surrounding plants occur at plant positions alongside 1, 2, 3, *etc.* **aureoles** around the centre of the grid. Figure 14.1 illustrates the regular triangular pattern of plant positions

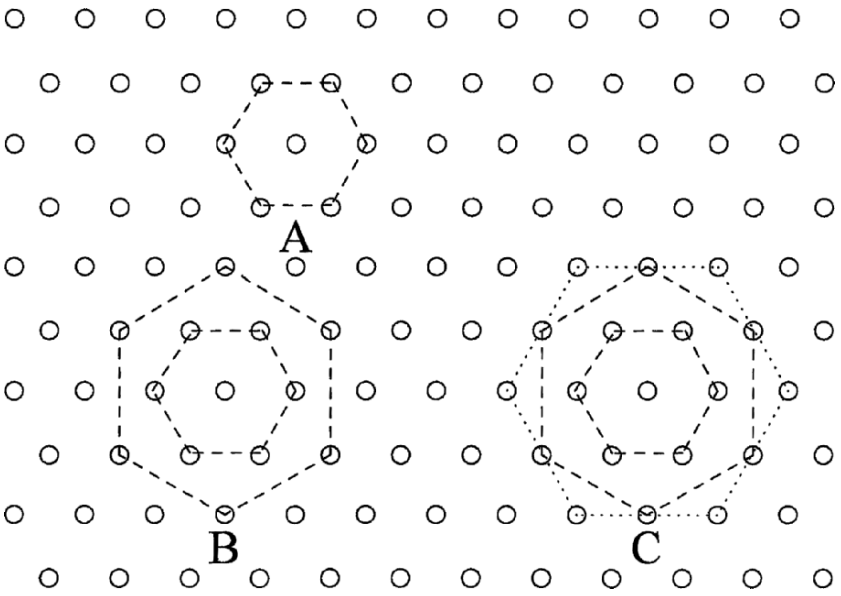


Fig. 14.1 A regular triangular pattern of plant positions. Each plant in its turn is considered as a candidate and compared to the plants occurring alongside one (grid A), two (grid B) or three (grid C) surrounding aureoles

as well as grids consisting of $1 + 6 = 7$ plant positions (grid A, one aureole), $1+(2\times6) = 13$ plant positions (grid B, two aureoles) or $1+(3\times6) = 19$ (grid C, three aureoles) plant positions. These grids have a more or less circular shape. The orientation of the grid is then irrelevant; their size is considered below.

The principle of selecting a central plant, if it performs better than all other plants in its grid, may also be applied at other patterns of plant positions. With a square pattern of plant positions a grid will consist of $1 + 4 = 5$ plant positions (one aureole), $1 + (2 \times 4) = 9$ plant positions (two aureoles), $1 + (3 \times 4) = 13$ plant positions (three aureoles), *etc.*

The idea underlying honeycomb selection is that the plants in a grid experience similar growing conditions. There is a strong indication that the central plant is genetically superior with regard to the considered trait, if it performs better than each of its fellow grid members.

In the first publications on honeycomb selection (Fasoulas, 1973; Fasoulas and Tsaftaris, 1975) grids involving only one aureole were recommended. Bos (1981, p. 144) concluded, on the basis of results of application of honeycomb selection (see Example 14.5a), that 1-aureole grids contain too small a number of plant positions: the great random plant-to-plant environmental variation within such small grids did, apparently, hardly imply that phenotypic superiority of a central plant was due to genetic superiority. Stam (1984) and Kyriakou and Fasoulas (1985) recommended the use of grids involving three aureoles.

Example 14.5a Bos (1981) applied 1-aureole honeycomb selection in winter rye. The intended plant density was 51.3 plants/m^2. Plants with a culm length less than the average culm length of their fellow grid members but yielding better than each of these were selected.

The results confirmed the common experience that the response to mass selection for a single generation is highly affected by incidental circumstances. Thus the cumulative response to honeycomb selection continued for three successive generations was evaluated at an intended plant density of 225 plants/m^2 (which is close to normal). The rye plants descending from the selected plants produced culms with a 6.1% reduced length, whereas their grain yield was 4.3% higher (Bos, 1981; Table 73). It was concluded that, notwithstanding the positive correlation of culm length and grain yield, the selection resulted in changes in the desired directions.

Fasoulas (1973) suggested growing, at specified positions in the triangular pattern of plant positions, plants representing some standard variety. This allows comparison of the phenotypic value of each candidate plant with the average phenotypic value of the three nearest standard plants (Fasoulas and Tsaftaris, 1975). This modification of honeycomb selection is in fact an application of the procedure described in Section 14.2.1.

14.3 Evaluation of Candidates by Means of Plots

14.3.1 Introduction

Clones, lines, hybrids and families are mostly evaluated by means of plots. Such evaluation occurs especially when applying line selection or family selection. In Sections 6.3.2 to 6.3.4 aspects of such selection were considered with regard to traits with qualitative variation. In connection with the topic of evaluation by means of plots, line and family selection are now considered with regard to traits with quantitative variation. The three main features of line or family selection are:

1. Each candidate is judged on the basis of its average performance across all plants representing the candidate. An entry is selected or rejected depending on this average performance. Mostly each candidate is tested at $J(\geq 1)$ plots, each containing $K(\geq 1)$ plants.
2. Within selected families, single-plant selection is either applied or omitted. In the former case the selected plants are the parents of the lines or the families to be evaluated the next generation. They are, in contrast to mass selection, not exclusively selected on the basis of their own phenotypic value. The performance of the line or the family to which they belong plays an important role. **Combined selection** is thus common in connection with line or family selection with regard to quantitative variation. One may, for instance, select the best 10 plants in each of the best 10% of the lines or families.
   In the latter case, seeds produced by a random sample of plants, either or not separately harvested, are used to grow lines or families to be evaluated in the next generation.
3. The next generation is grown in separate plots tracing back to
   - Seed produced by the evaluated plants themselves (this could be individually selected plants)
   - Seed produced by the evaluated (and selected) families
   - Seed produced by sibs of the evaluated (and selected) families (this is called **sib selection**)

New varieties are released continuously because they are superior as compared to the already existing varieties. However, their performance is not always clearly superior. Their attractive performance may merely be due to an improved resistance instead of an improved complex genotype with regard to loci directly controlling the performance, possibly with regard to yield (if such loci do exist at all). Excluding hybrid varieties, spectacular breakthroughs with regard to yield are rare.

In the breeding of self-fertilizing crops, selection for qualitative variation usually starts in the first segregating generations. Selection for resistance

against diseases, pests or abiotic stress factors may start already in populations representing the F_2, the F_3 or the F_4 generation. Later on, from F_4 onwards, selection is focussed on quantitative variation in traits. This strategy is, in fact, a form of tandem selection (Section 12.1).

Within a general framework for the selection in self- or cross-fertilizing crops, many different approaches can be followed. These concern

- The number of evaluated candidates (lines of families)
- The number of plots (J) used to evaluate each candidate
- The number of plants evaluated per candidate
- The number of selected candidates
- The number of plants harvested or selected within the selected candidates
- The procedure for data adjustment aiming at improvement of the efficiency of selection

First some remarks are made with regard to J, the number of plots used to evaluate each candidate. Application of $J = 1$ may be due to

- the small amount of seed produced per (selected?) parental plant and/or
- the (very) large number of candidates.

Adjustment for a plot-to-plot trend in the quality of the growing conditions deserves certainly attention when evaluating the candidates on the basis of non-replicated plots. Such an adjustment may make use of

- Standard plots, *i.e.* plots containing a standard variety (Section 14.3.2)
- A moving mean (Section 14.3.3).

The adjustment aims at improvement of the efficiency of the selection, *i.e.* at an increase of the heritability.

Replicated testing of the candidates ($J > 1$ may be applied when the selected plants produce sufficient amounts of seed and/or when it is permitted due to a reduced number of candidates. Replicated testing within a given macro-environment aims to increase the heritability (Section 11.2.1). Replicated testing across several macro-environments, *i.e.* several growing seasons and/or locations, should be applied when wide adaptation or stability across different macro-environments is pursued. A selection procedure employing replicated testing of families obtained by open pollination in an annual crop, namely **modified ear-to-row selection,** is now described. The procedure has successfully been applied in maize.

Each of the I candidate HS-families is tested at $J \geq 3$ locations. The average performance across these locations is determined for each candidate and used as a measure of the quality of the candidate. With this yardstick effects of genotype $\times$ location are cancelled out. At an additional location the five most attractive open-pollinated plants are preliminary selected in each candidate family. As soon as the average performance across the J locations is known, the top 20% of the families can be identified and then the five most

attractive plants within these families, as identified at the additional location, are eventually selected. This procedure is not only a clear example of combined selection but also of sib selection. A 'cycle' lasts one generation. The procedure is illustrated in Example 14.6.

Example 14.6 Lonnquist (1964) introduced the modified ear-to-row selection procedure in order to improve grain yield of maize (Fig. 14.2).

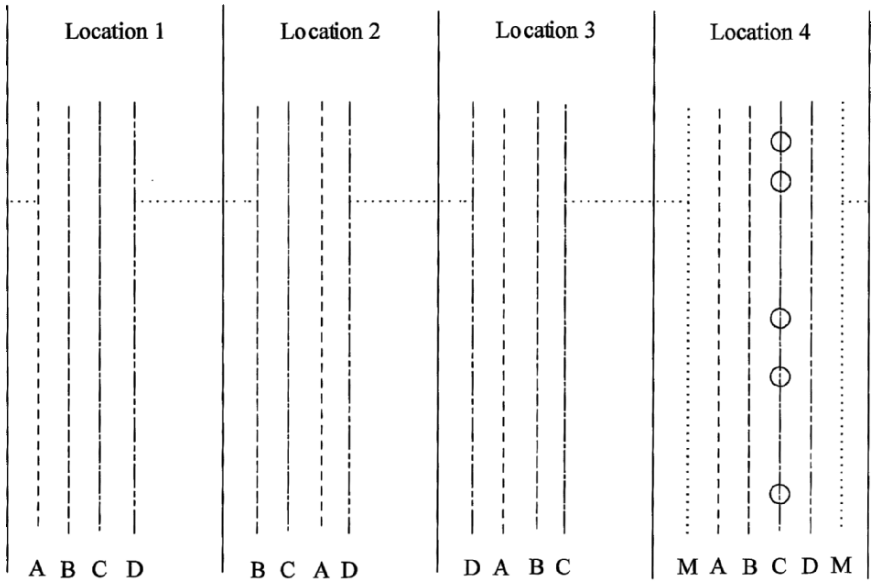


Fig. 14.2 A scheme representing modified ear-to-row selection. At locations 1, 2 and 3 all families (A, B, C, ...) are tested. The families with the best average performance across the three locations are identified as being best (here family C among families A, ..., D). At location 4 all families are grown as rows of plants to be emasculated, alternated by rows consisting of a mixture of all families (indicated by M). The latter rows are the so-called pollinator rows. In the best families the best-performing plants are selected (here five plants in family C)

Each HS-family was tested at three locations. The procedure was applied for four generations to the open-pollinating maize variety Hays Golden. The response to the selection was 9.44% per 'cycle' (Webel and Lonnquist, 1967).

A drawback of modified ear-to-row selection is the participation of inferior families in the open pollination. To avoid this one could adopt a modification similar to the remnant seed procedure (Section 6.3.4). Thus after the first season the best families are identified. In the second season a mixture of remnant seed, representing these best families, is grown in the pollinator rows (Compton and Comstock, 1976).

In Section 16.1 the optimal number of replications when using plots of a fixed size is considered. Replicated testing may, however, impose the use of smaller plots than would be applied in the case of non-replicated testing. Thus the plot size, especially the number of rows per plot, may be affected by the number of replications. Section 16.2 deals with the size and the shape of the plots.

The replications may coincide with (complete) blocks. Then randomization of the entries across each complete block (or in sets of incomplete blocks accommodating all entries) instead of randomization across all available plots should be applied. In Chapter 19 the application of complete and incomplete blocks is considered.

Breeders pursuing the same goal for the same crop tend to apply very different approaches. There is, apparently, no unambiguous guideline to choose the most appropriate procedure. The remainder of this section serves to give an impression of the diversity of the approaches. For an annual crop a 'cycle' of family selection may involve three growing seasons:

- First season: selection of the plants yielding the families to be evaluated. Depending on the trait(s) to be improved the selection may occur both before and after pollen distribution.
- Second season: evaluation of the families and identification of the best families.
- Third season: intercrossing of the best families. These are grown once again from remnant seed. The families may be grown as a mixture.

Mostly the activities of the first and third season take place in the same season. Then a single cycle consists of

- First season: intercrossing of families previously identified to be the best. These families are grown from remnant seed. The families may be grown as a mixture. Among the plants representing these best families the best plants are selected. These yield the families to be evaluated.
- Second season: evaluation of the families and identification of the best families.

Depending on the crop and the trait to be improved a further acceleration may be possible: about 4–8 weeks after the sowing the families to be evaluated, these families are sown again from remnant seed. The sowing occurs ear-to-row in order to be able to intercross plants of families that appear to excel for traits expressed early in the ontogenesis of the crop. For example: by growing two generations per year Singh, Khehra and Dhillon (1986) could complete four 'cycles' of selection with regard to the number of ears of individual maize plants within two years.

The number of entries to be evaluated (I) or the number of plots ($I \times J$) may be determined by the number of plots from which observations can be obtained. Harvesting, weighing and measuring are especially time-consuming. The manageable number of plots can drastically be increased if unbiased and

accurate **visual assessment** of the trait of interest is possible. In that case **visual selection**, *i.e.* selection based on the visual assessment, may successfully be applied. Results of visual selection as applied by breeders and layman (Example 14.7) or by just breeders (Example 14.8) are now given.

Example 14.7 In 1971 Townley-Smith and co-workers tested 251 wheat lines, both in Swift Current and Regina (Saskatchewan, Canada) (Townley-Smith, Hurd and McBean, 1973). In each third plot the standard variety Neepawa was grown. Immediately before harvest the lines were assessed visually for yielding ability. The 25% of the lines presumed to be best were tagged. This was done independently by wheat breeders, by their assistants and by scientists from other disciplines. After the harvest the lines were ranked for their actual grain yield.

It appeared that both in Swift Current and Regina each person had tagged at least one line from each class containing 10% of the ranked lines. For each person the average yield of the tagged lines was higher than the average yield of all lines, but many of the really outstanding lines were not tagged at all.

The breeders' tagging was worse than the assistants'. More lines than were to be expected in the case of tagging at random were tagged in common by several breeders. Their assessment was apparently based on having in mind a similar ideotype for traits such as culm length, uniformity, *etc.*, whereas that ideotype did not appear to imply a high actual grain yield.

Persons tagging both in Swift Current and Regina often tagged the same lines in both locations. This applied especially to the breeders.

It was concluded that variation for morphological traits could be distinguished by visual assessment, but that this did not lead to an unbiased and accurate judgement of grain yield.

14.3.2 Use of Plots Containing a Standard Variety

In the former section it was emphasized that adjustment of data resulting from a non-replicated evaluation of candidates deserves special attention in the presence of a plot-to-plot trend in the quality of the growing conditions. The adjustment may be on the basis of **standard plots**, *i.e.* plots containing a standard variety (this section) or on the basis of **moving means** (Section 14.3.3). Non-replicated testing may be unavoidable because of lack of seed or because replicated testing is not feasible due to the high number of candidates.

These adjustment procedures can, of course, also be incorporated in replicated tests (Example 14.18). The efficiency of an adjustment procedure may

then be measured by comparison of the F value for candidates, calculated in an analysis of variance of the unadjusted phenotypic values, to the F value calculated in an analysis of variance of adjusted phenotypic values.

Example 14.8 Brown *et al.* (1984) studied the efficiency of visual selection in potato. In 1981, 200 seedlings, obtained from each of eight crosses, were grown in an aphid-proof glasshouse (in 10 cm square pots). Each plant/pot was harvested individually. Four potato breeders assessed the 1600 genotypes visually according to a 1–9 scale of increasing desirability. The visual assessment took into account all the features observed to provide an overall measure of commercial suitability.

In the spring of 1982 the harvested tubers were planted in the field. Two sites were used: W (a ware production site) and S (a high-grade seed site). At each site two completely randomized blocks were planted; each plot being a single plant. From the 1600 initial genotypes, 224 failed to produce any tubers and 122 produced only one tuber. The single-tuber clones were only grown at site S. All genotypes represented by two or more tubers were grown in at least one block at each site. The largest tuber was grown in block 1 at site S, the next two largest were grown at site W, and the smallest of the four was planted in block 2 at site S. The S trial contained $1600 - 224 = 1376$ genotypes in block 1, the W trial $1600 - 224 - 122 = 1254$ genotypes in block 1. Only 824 genotypes were grown in all four blocks. Each plant was harvested by hand. The four breeders assessed each plant.

The correlation coefficient of the scores, within each of the three year-site combinations, ranged from 0.34 (breeders 1 and 4 for the seedlings) to 0.84 (breeders 1 and 3 for the W trial). The four breeders were thus in reasonably good agreement as to what should be selected in each environment.

The coefficient of correlation between the mean scores of the four breeders for the glasshouse and the W or the S trial amounted to 0.29 and 0.26, respectively. Of the clones that would have been discarded as a seedling, 24% would have been selected at either S, M or both sites; of the clones that would have been selected as a seedling 41% would have been selected at one or both of the sites. One of the seedlings that was assessed as a 1 by all the breeders had a first-clonal-year score of 8.5 at site S and 7.8 at site W. It was concluded that visual selection of seedlings was not very efficient. (However, significant correlations between yield data (total tuber weight or mean tuber weight) recorded in the glasshouse and in the first-clonal-year suggest that seedling selection for yield characters can be effective (Brown and Caligari, 1986).

In 1983 it was not possible to grow all the material that had been handled previously and in this second-clonal-year each family was represented by 70 randomly chosen clones. These were grown at sites S and W in two completely randomised blocks with each clone in each block represented by a three-tuber plot. The mean breeder's preference score was calculated for

each of the eight families on the basis of 70 clones that were grown in all three years. Brown, Caligari and MacKay (1987) estimated for these means the rank correlation across the 8 families between the five year-site combinations. These correlations suggested that breeders should grow samples of clones representing many families and then identify the families with the highest rankings for mean preference score. The corresponding crosses should then be made again, such that a much larger family size is obtained. This suggestion comes close to cross prediction (Section 11.4): identification of crosses having a high probability of producing desirable genotypes.

The repeatability of the visual preference scores for individual clones over generations was shown to be very low. It was concluded that selection of individual clones should not be practised in the early generations.

Mostly candidates are evaluated together with one or more proven varieties serving as a standard by growing them in contiguous plots arranged in strips. The standard plots may be distributed within the strips in a regular or in an irregular pattern. In the regular pattern each third, or each fifth, or each seventh, or ...., *etc.* plot is a standard plot. In the irregular pattern a certain number of plots are assigned at random to the standard variety.

The adjustment may consist of calculating

$$\underline{p}'_i = \underline{p}_i - \bar{\underline{p}}_{i\mathrm{S}} \tag{14.15}$$

where

$\underline{p}_i$:= the phenotypic value of the candidate occurring at plot i; and
$\bar{\underline{p}}_{i\mathrm{S}}$:= the weighted mean phenotypic value of standard plots near to plot i.

The adjustment is followed by truncation selection with regard to $\underline{p}'$.

The symbol $\hat{p}_i$ is introduced for the phenotypic value predicted for the standard variety when it would occur at plot i. Its value is calculated on the basis of phenotypic values obtained from standard plots near to plot i. This quantity may be used in a more general equation for adjustment of the phenotypic value of the candidate actually occurring at plot i, namely

$$\underline{p}'_i = \underline{p}_i - \hat{p}_i \tag{14.16}$$

In the case of a regular pattern of the positions of standard plots there occur n candidate plots in between two standard plots. One may calculate $\hat{p}_i$ then on the basis of some assumption with regard to the trend in the quality of the growing conditions. This is illustrated for the following three assumptions:

1. A constant quality of the growing conditions in the neighbourhood of each standard plot
2. A linear trend in the quality of the growing conditions between two standard plots

Plot	Entry	p
1	C_1	9
2	S	10
3	C_2	7
4	C_3	9
5	S	11
6	C_4	10
7	C_5	10
8	S	12
9	C_6	12
10	C_7	13
11	S	12
12	C_8	13
13	C_9	12
14	S	10
15	C_{10}	13

Fig. 14.3 The phenotypic values (p) obtained from a trial field consisting of a strip of 15 plots. Entry S is a standard variety grown at each third plot; entries $C_1, \ldots, C_{10}$ are 10 candidates

3. A smooth, curvilinear trend in the quality of the growing conditions across the whole field

   N. B. With appropriate modifications the adjustments elaborated in the following may also be applied in case of an irregular distribution of the positions of the standard plots.

A constant quality of the growing conditions in the neighbourhood of each standard plot

In the case of a constant quality of the growing conditions in the neighbourhood of each standard plot, each candidate's phenotypic value is adjusted according to Equation (14.16) with taking for $\hat{p}_i$ simply the phenotypic value of the nearest standard plot. The procedure is applied to the data provided by Figure 14.3 and illustrated by Example 14.9.

Example 14.9 Fig. 14.3 illustrates a trial field consisting of a strip of 15 contiguous plots. The evaluation involves 10 plots used to evaluate candidates $C_1, \ldots, C_{10}$ and five plots grown with standard variety S. The 10 candidates may be randomized across the 10 plots reserved for them, but this is not required. Because S is grown in each third plot, each standard plot is

surrounded by two candidate plots. When taking for $\hat{p}_i$ the phenotypic value obtained from the nearest standard plot, the adjusted phenotypic values are:

	Candidate									
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
p'	−1	−3	−2	−1	−2	0	1	1	1	2

According to this adjustment procedure, the best candidate is C_{10}.

A linear trend in the quality of the growing conditions between two standard plots

In the case of a linear trend in the quality of the growing conditions between two standard plots the phenotypic value predicted for plot $i(\hat{p}_i)$ is calculated as the weighted mean of the phenotypic values of its two nearest standard plots. The procedure is applied to the data provided by Figure 14.3 and illustrated by Example 14.10.

Example 14.10 When assuming a linear trend in the quality of the growing conditions between two standard plots, one may predict the phenotypic value for a standard variety grown on plot $i(\hat{p}_i)$ in the following way. For the data of the trial described by Figure 14.3 the phenotypic value of the standard variety as predicted for plot 3, for instance, is

$$\hat{p}_3 = \tfrac{2}{3}(10) + \tfrac{1}{3}(11) = 10.33$$

Equation (14.16) yields then

$$p_3{}' = 7 - 10.33 = -3.33$$

For plot 4 the adjusted phenotypic value p'_4 is calculated to be

$$9 - [\tfrac{1}{3}(10) + \tfrac{2}{3}(11)] = -1.67$$

The adjusted phenotypic values are thus:

	Candidate									
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
p'	−0.67	−3.33	−1.67	−1.33	−1.67	0	1	1.33	0.67	2.33

As in Example 14.9, candidate C_{10} appears to have the highest adjusted phenotypic value also with the present adjustment procedure.
N. B. The adjusted phenotypic values for C_1 and C_{10} were calculated by extending the trend between plots 2 and 5 and the trend between plots 11 and 14, respectively.

A smooth, curvilinear trend in the quality of the growing conditions across the whole field

When assuming a smooth, curvilinear trend in the quality of the growing conditions one may plot the phenotypic values of the standard plots against their plot numbers and sketch a smooth line indicating for each plot the phenotypic value predicted for the standard variety ($\hat{p}_i$). The predicted phenotypic values may also be obtained by fitting some polynomial function to the phenotypic values obtained for the standard plots. The adjustment procedure is applied to the data provided by Figure 14.3 and illustrated by Example 14.11.

Example 14.11 Let i represent the plot number. The function

$$\hat{p}_i = 8.57 + 0.73i - 0.04i^2$$

can then be calculated to predict, on the basis of a quadratic function in i, the phenotypic value of standard variety S at plot i. The function was obtained by regressing the actual phenotypic values of the standard variety on their plot numbers (see Figure 14.3). The adjusted value for plot 3 amounts then to

$$p'_3 = 7 - [8.57 + 0.73(3) - 0.04(3^2)] = -3.4$$

The complete set of adjusted phenotypic values of the candidates is:

	Candidate									
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
p'	−0.26	−3.4	−1.85	−1.51	−1.72	0.1	1.13	1.43	0.7	2.48

The actual phenotypic values of candidates C_7, C_8 and C_{10} amounted to 13. However, according to its adjusted phenotypic values candidate C_{10} is consistently identified as the one with the highest adjusted phenotypic value in Examples 14.9 to 11.

Federer (1956) defined the **fertility index** $\underline{f}_i$ of plot i (which contains the standard variety), namely

$$\underline{f}_i := \underline{p}_{iS} - \overline{\underline{p}}_S \tag{14.17}$$

where

$\underline{p}_{iS}$:= the actual phenotypic value of the standard variety at plot i
$\overline{\underline{p}}_S$:= the mean phenotypic value across all plots containing the standard variety

The quantity f_i indicates for each plot containing the standard variety the quality of the growing conditions it provides.

When assuming a linear trend in the quality of the growing conditions between two standard plots one may calculate the fertility index of each candidate plot. Regression of the phenotypic values obtained from the candidate

plots on their fertility indices may be used to predict a phenotypic value for candidate plot i:

$$\hat{p}_i = a + bf_i$$

The adjusted phenotypic value of the candidate occurring at plot i is then calculated as

$$p_i{}' = p_i - \hat{p}_i = p_i - (a + bf_i)$$

or, when neglecting the intercept a, as

$$p'_i = p_i - bf_i \tag{14.18}$$

This equation is similar to Equation (14.10).
It is once more emphasized that adjustments aim at improvement of the evaluation of the genotypic values of candidates, *i.e.* at $\rho_{p',g}$ being larger than $\rho_{p,g}$. Such an improvement may appear from the heritability of p' being higher than the heritability of p. In plant breeding practice it is, however, often taken for granted that this goal is attained. Example 14.12 illustrates an unreliable way of establishing whether or not the desired goal is attained.

Example 14.12 Shebeski (1970) expressed the grain yield of individual plots sown with spring wheat lines as a percentage of the grain yield obtained from nearby standard plots. (Thus an adjustment similar to Expression (14.3) was applied.) This yielded a coefficient of correlation between F_3 lines and their F_5 progenies as high as $r = 0.84$. Pembina was the standard variety for the F_3 lines, and Manitou was the standard for the F_5 progenies. Briggs and Shebeski (1971) advised the selection of F_3 lines with a high grain yield *per se* as well as a high relative yield compared to a nearby standard.

One may consider for a certain site a **uniformity trial**, *i.e.* a 'trial' where all plots contain the same genotype. A high coefficient of correlation between the data obtained from adjacent plots in this trial does not warrant that adjustment of candidate data – on the basis of plots containing a standard variety – is efficient at this site: when different candidates are tested at contiguous single-row plots, interplot competition may disturb the quality of the adjustment. (In Section 14.2.3 it was remarked that the change of the environmental variance due to adjustment, namely an increase or a decrease, depends on the combined effect of a trend in the quality of the growing conditions and the effect of intergenotypic competition.) Example 14.13 provides some data concerning the correlation between data obtained from adjacent standard plots.

Example 14.13 Briggs and Shebeski (1968) obtained a highly significant coefficient of correlation, *i.e.* $r = 0.64$, between the yields of standard plots separated 2.7 m. The coefficient of correlation clearly declined at larger distances between the standard plots. At a distance of 19.2 m (in two experiments) or 35.7 m (in a third experiment) the correlation was insignificant; *viz.* $r = 0.08$ in the former case.

Hadjichristodoulou and Della (1976) studied protein content of durum wheat. They estimated coefficients of correlation between standard plots. These were high and significant (0.31 − 0.74) for plots 0.6 m apart, but decreased rapidly with increasing distances between the standard plots. Significant coefficients of correlation were not obtained at distances of 6 m or larger.

High values for the soil heterogeneity index (Section 16.2.3) imply a weak correlation between data of contiguous plots: they imply a weak trend in the quality of the growing conditions. When defining over-adjustment as the situation where the residual variance in the analysis of variance is higher in the presence of the adjustment than in its absence, Baker and McKenzie (1967) deduced from a theoretical study that Equation (14.15) leads to over-adjustment if the soil-heterogeneity index is larger than 0.5. They also considered adjustment in a way similar to the adjustment given by Equation (14.10), namely

$$\underline{p}'_i = \underline{p}_i - b\overline{\underline{p}}_{i\mathrm{S}}, \tag{14.19}$$

This adjustment yielded in all studied cases a reduction of the residual variance. The study confirmed the validity of the positive opinion of Yates (1936) about Equation (14.10) relative to Equation (14.8). The results of the theoretical study were more-or-less confirmed by an experimental verification (Example 14.14).

Example 14.14 Baker and McKenzie (1967) compared oat lines to a standard variety grown in every second plot. Single-row plots were used.

In 1964 adjustment resulted in all cases in an increased residual variance, but especially when Equation (14.15) was applied. In 1965 adjustment gave rise to a decrease of the residual variance, especially when Equation (14.19) was used. The reduction was at most only 14%. The additional costs, due to growing the standard variety at every second plot, could thus not be justified. Baker and McKenzie concluded that adjustment on the basis of regularly inserted standard plots is risky.

Early in the twentieth century the use of standard plots was quite popular. Later, the established opinion was that the accuracy of the evaluation of candidates could be increased sufficiently by using smaller, but more frequently repeated plots. Thus in 1921 a Committee on Standardization of Field Experiments, set up by the American Society of Agronomy, recommended not to make use of standard plots anymore (Kempton, 1984). The present authors support the opinion that adjustment on the basis of standard plots must be critically considered. The reasons for this opinion are the following:

1. The standard plots require an additional part of the trial field and additional attention.
2. The observations on a standard variety include contributions of genotype × location and genotype × season interactions which are specific for the standard variety. The adjusted values of the candidates apply, consequently, only when using the involved standard variety and under the conditions provided by the location and the season. A change of standard variety may affect the ranking of the adjusted values of the candidates. Generalization of the ranking of the candidates to other macro-environmental conditions, *i.e.* to other locations and/or years, is thus risky. The standard variety actually used may thus be inappropriate for regional or international tests. Furthermore standard varieties tend to have a short life span.
3. Dominating random plot-to-plot variation in the quality of the growing conditions may obliterate the course of the trend. The predicted value $\hat{p}$ (Equation 14.16) may then be based on an incorrect assumption about the course of the trend. This gives rise to the application of an inappropriate procedure for data adjustment on the basis of plots containing a standard variety.
4. Sometimes the quality of evaluation of candidates on the basis of adjusted data is worse than the quality of evaluation on the basis of unadjusted data. Then the goal of the adjustment, namely

$$\rho_{p',g} > \rho_{p,g},$$

is not attained. This may appear from comparison of the F value for candidates, calculated in an analysis of variance of the unadjusted phenotypic values, to the F value calculated in an analysis of variance of adjusted phenotypic values.

Notwithstanding the above objections, breeders generally agree that standard plots should be included in an evaluation of candidates in order to be able to perform – throughout the growing season – a visual assessment.

14.3.3 Use of Moving Means

When the number of candidates is to be counted in hundreds instead of in tens, it is unavoidable that the trial field provides heterogeneous growing conditions. Mostly the variation in the quality of the growing conditions occurs partly at random and partly with a trend.

The observation obtained from plot i, *i.e.* $\underline{p}_i$, is thus likely to contain a contribution due to a trend in the quality of the growing conditions. One may

eliminate this contribution by adjustment according to

$$\underline{p}'_i = \underline{p}_i - \overline{\underline{p}}_i \tag{14.20}$$

or

$$\underline{p}'_i = \underline{p}_i - b\overline{\underline{p}}_i \tag{14.21}$$

where

$\overline{\underline{p}}_i$:= the mean of the observations obtained from the k plots surrounding plot i, the considered (central) plot

b := the estimate of the coefficient of regression of $\underline{p}_i$ on $\overline{\underline{p}}_i$

N. B. 1. One may consider regression through the origin.

N. B. 2. Equations (14.20) and (14.21) are identical to Equations (14.8) and (14.10), respectively, but the meaning of $\overline{\underline{p}}_i$ depends on the context.

The idea underlying these adjustment procedures is that the phenotypic value expected for plot i can reliably be predicted by $\overline{\underline{p}}_i$ or by $b\overline{\underline{p}}_i$. Indeed, the moving mean should provide an adequate impression of the quality of the growing conditions offered by plot i. There should be a high correlation ($r > 0.5$) between $\overline{\underline{p}}_i$ (or $\overline{\underline{p}}_{i\mathrm{S}}$) and $\underline{p}_i$ (Mak, Harvey and Berdahl, 1978). Yates (1936) suggested application of analysis of covariance in which $\overline{\underline{p}}_i$ or $\overline{\underline{p}}_{i\mathrm{S}}$ is used as a covariate: linear regression of $\underline{p}_i$ on the covariate, after elimination of replication and candidate effect, yields the regression coefficient b. Over-adjustment is avoided if b differs significantly from zero. This should warrant that the adjusted value ($\underline{p}'_i$), *i.e.* the difference between $\underline{p}_i$ and its predicted value ($\hat{p}_i$), is a better indicator of the genotypic value of the candidate occurring at plot i than the unadjusted phenotypic value.

Example 14.15 illustrates how a moving mean may show a trend in the quality of the growing conditions.

Example 14.15 In a spring wheat breeding programme at Swift Current, Saskatchewan, F_2-derived F_4 lines were evaluated. The plots consisted of four rows. They were separated by two rows sown with winter wheat, staying in the vegetative phase. The plot size was 2.7 m^2. The plots were arranged in strips containing 42 plots. Grain yield per plot was measured in grams. Fig. 14.4 depicts, for one of the strips, the grain yields as well as the moving mean grain yields calculated across $k = 6$ nearby plots (Bos and De Pauw, 1984).

Especially the moving means suggest the presence of a trend in the quality of the growing conditions. It was speculated that this was due to unevenness of the field surface, which gives rise to variation in soil moisture content. In the semi-arid conditions of Swift Current such variation is reflected by variation in grain yield.

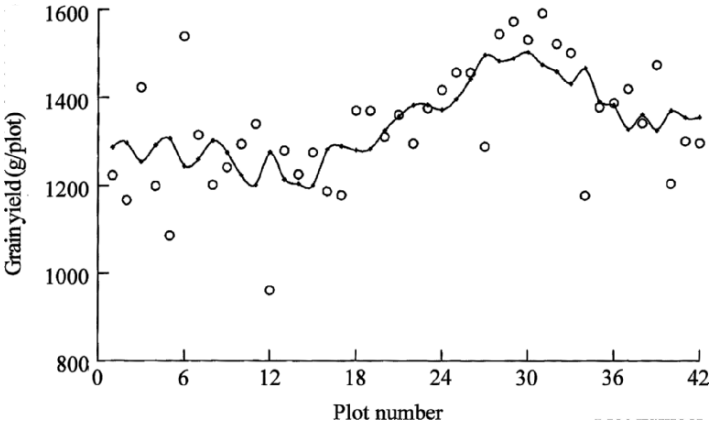


Fig. 14.4 The grain yield (in g/ plot) obtained in a spring wheat trial consisting of 42 plots alongside a strip. The moving mean, calculated across six neighbour plots, is depicted as a smooth line

Examples 14.16 and 14.17 report about the efficiency of adjustment based on moving mean data as compared to adjustment based on standard plot data.

In Example 14.16 the moving mean adjustment was clearly superior, as compared to the standard plot adjustment, with regard to reduction of the error variance. This supports the validity of Baker and McKenzie's conclusion (Example 14.14).

Example 14.16 Townley-Smith and Hurd (1973) tested a series of experimental spring wheat lines: five series with frequently repeated standard plots at Swift Current in either 1969 and 1970 (with $r = 2$ or 3) and eight series without frequently repeated standard plots at four locations in 1970. They applied the following adjustments of the phenotypic value p_i of plot i:

- $\underline{p}'_{1i} = \underline{p}_i - \overline{\underline{p}}_i$, the moving mean adjustment (Equation (14.20)), involving $k = 2(2)20$ neighbour plots
- $\underline{p}'_{2i} = \underline{p}_i - \overline{\underline{p}}_{i\mathrm{S}}$, the standard plot adjustment (Equation (14.15))

The relative efficiency of the adjustments was measured by dividing the error variances obtained from standard analyses of variance of the data for p_i (say: s^2), p'_{1i} (say: s_1^2) and p'_{2i} (say: s_2^2). For series A, for example, it was found that $s_1^2/s^2 = 2038/3728 = 0.547$ (at $k = 8$).

In contrast to adjustment p'_{2i}, adjustment p'_{1i} consistently reduced the error variance: its relative efficiency ranged from 0.55 for $k = 8$ (in series A) to 0.95 for $k = 10$.

In comparison to the adjustments according to Equations (14.20) and (14.15), those according to Equations (14.21) and (14.19) did not yield an additional reduction of the error variance, except when the adjustment was made on the basis of only two standard plots.

Example 14.17 Mak, Harvey and Berdahl (1978) tested, in 1974, 143 homozygous barley lines by means of three randomized complete blocks. Each single-row plot consisted of 20 plant positions. The inter-row distance amounted to 30.5 cm, the intra-row distance to 24 cm. Each third plot contained the standard variety Bonanza. Grain yield and protein content of the 10 central plants in each plot were determined.

The test was repeated in 1975 with 142 lines and two standard varieties (Bonanza and Hector), which were alternately grown at each third plot. Within each single-row plot 150 kernels were sown. The plot length was 366 cm, the plot width, *i.e.* the inter-row distance, 30.5 cm. A 12×12 partially balanced lattice with two replicates (see Chapter 19) was used as the experimental design. For each plot grain yield and protein content were determined.

The data obtained from plot i were adjusted on the basis of:

1. The moving mean of the observations obtained from $k = 2(2)12$ nearby plots ($\overline{\underline{p}}_i$), namely equal numbers of plots on each side of plot i, except at the end of the strip of plots where the appropriate number of plot nearest to plot i was used. Thus

$$\underline{p}'_{11i} = \frac{\underline{p}_i}{\overline{\underline{p}}_i} \text{(percentage adjustment) and } \underline{p}'_{12i} = \underline{p}_i - b\overline{\underline{p}}_i \text{(covariance adjustment)}$$

2. The mean of the observations obtained from two nearby standard plots ($\overline{\underline{p}}_{is}$). Thus

$$\underline{p}'_{21i} = \frac{\underline{p}_i}{\overline{\underline{p}}_{is}} \text{(percentage adjustment) and } \underline{p}'_{22i} = \underline{p}_i - b\overline{\underline{p}}_{is} \text{(covariance adjustment)}$$

The coefficient of residual variation (cv_e) was used to compare the evaluation procedures.
Only cv_e values for grain yield, as obtained with $k = 6$, are given:

	cv_e	cv_{e11}	cv_{e12}	cv_{e21}	cv_{e22}
1974	0.24	0.32	0.23	0.28	0.23
1975	0.18	0.17	0.15	0.17	0.15

In 1974 the percentage adjustments (adjustments 11 and 21) failed. This was avoided by the covariance adjustments (adjustments 12 and 22). In 1975 the covariance adjustments gave, in comparison to the percentage adjustments, a relatively large reduction of cv_e.

It is concluded that percentage adjustments perform badly to poor. (This was already remarked by Yates, 1936). In contrast, covariance adjustments perform poor to good: cv_e was reduced by nearly 20%.

In Example 14.17 adjustment using a moving mean was superior, as compared to adjustment on the basis of standard plot data, with regard to reduction of the coefficient of residual variation. Additionally moving mean adjustments have the advantages that the number of candidates is not limited, that replication is not necessary and that one does not have to grow standard plots.

The optimum value for k, say: k_{opt}, *i.e.* the value giving rise to a minimum error variance of the adjusted data, depends on the nature of the variation in the quality of the growing conditions. This variation may be fine- or coarse-grained. It varies from trial to trial and from location to location. In the case of a fine-grained pattern, a high value for k may imply that $\overline{p}_{i}$ is a biased estimator of the quality of the growing conditions at plot i. However, a smaller value for k implies a higher mean squared error of $\overline{p}_{i}$. Indeed, the accuracy of $\overline{p}_{i}$ as an estimator of the quality of the growing conditions at plot i depends on the genetic variation among the (fortuitous) set of involved candidate genotypes.

The value of k_{opt} depends thus both on the pattern of variation in the quality of the growing conditions and the genetic diversity among the k involved candidates. Guidelines to derive k_{opt} are not (yet) available. It is taken for granted that $k \approx 8$, a commonly applied value, is a reasonable choice. This value can be used in a moving mean adjustment, but to run an analysis of variance for several other values for k (if indeed J, the number of plots per candidate, is at least 2) is so little work, that any improvement with regard to a (further) reduction of the error variance is worth the effort.

Canadian wheat breeders have pioneered work on data adjustment on the basis of standard plots and moving means. Knott (1972) found that both of these adjustment procedures yielded, in comparison to absence of adjustment, reduced residual variances and higher F values in the statistical test. Townley-Smith and Hurd (1973) reported consistently reduced residual variances only when adjusting on the basis of moving mean data (Example 14.16).

Provided that replicated testing is applied, the quality of an adjustment procedure may also be measured in another way. With replicated testing one can estimate, for adjusted as well as for unadjusted data, the coefficient of correlation, across the candidates, between the data from different replications. An adjustment improves the quality of an evaluation if the coefficient of correlation obtained for adjusted data is higher than the coefficient of correlation obtained for unadjusted data. Results reporting with regard to this yardstick are given by Example 14.18.

Example 14.18 Mitchell, Baker and Knott (1982) estimated the coefficient of correlation, across a number of wheat lines, between grain yield data obtained from different replications. For two experiments they obtained for the unadjusted yield data: $r = 0.30$ and $r = 0.35$. After moving mean adjustment (with $k = 6$) they obtained for both experiments $r = 0.47$. This implies that the adjusted phenotypic values give a more consistent, but still unsatisfying, impression of the genotypic values of the candidates than the unadjusted phenotypic values.

Bos and De Pauw (1984) studied yield data for seven series of lines of *Triticale* or tetraploid or hexaploid wheat. They applied moving mean adjustment with $k = 8$ (but for end plots k amounted to 4 to 8). This tended to give rise to an increase of the coefficient of correlation. The increase ranged from -0.01 (at an 'increase' from $r = 0.62$ to $r = 0.61$) to 0.28 (at an increase from $r = 0.12$ to $r = 0.40$). The unweighted average increase of the coefficient of correlation amounted to 0.10.

It is concluded that moving mean adjustment is more efficient than standard plot adjustment. An additional advantage of moving mean adjustment, compared to standard plot adjustment, is the (much) smaller number of standard plots to be inserted in the tests. Finally, the second disadvantage of standard plot adjustment (see the end of Section 14.3.2), hardly plays a role in moving mean adjustment: the interactions of the genotypes involved in the moving mean with the location and the season will largely cancel out among these genotypes.

A minor drawback of the moving mean adjustment is the necessity to obtain observations from each plot, even from apparently inferior plant material.

In recent years the merits of more complicated adjustment procedures have been studied, both theoretically and experimentally. Because they require replicated testing ($J \geq 2$) these procedures are less relevant in the context of Section 14.3. For historical reasons only the proposal of Papadakis (1937) is introduced.

The quality of the growing conditions at plot i, say e_i, may be estimated by subtracting from $\underline{p}_i$, *i.e.* the phenotypic value of the candidate tested at plot i, the estimate of the genotypic value of this candidate ($\hat{\mathcal{G}}$). The latter value is estimated by calculating the mean phenotypic value across all plots containing the considered genotype. Thus:

$$\hat{\underline{e}}_i = \underline{p}_i - \hat{\underline{\mathcal{G}}} \tag{14.22}$$

The quantity $\hat{\underline{e}}$ may be calculated for each plot. Papadakis adjusted $\underline{p}_j$, the phenotypic value of the candidate occurring at plot j, on the basis of information about the two adjacent plots i and k. The sum $\hat{\underline{e}}_i + \hat{\underline{e}}_k$ was thus used as covariate when adjusting $\underline{p}_j$ by means of an analysis of covariance.

This adjustment did not raise attention until about 1970. From the nineteen eighties it became suddenly the subject of study of many statisticians (see, for example, Wilkinson *et al.*, 1983). Example 14.19 reports about an application of estimating the environmental deviation according to Equation (14.22).

Example 14.19 Shorter and Butler (1985) adjusted yield data of F_2-derived F_4 and F_5 lines of peanuts (*Arachis hypogaea* L.) by means of analysis of covariance.

1. The phenotypic values were adjusted according to Equation 14.21, by using as covariate $\overline{\underline{p}}_i$, *i.e.* the moving mean calculated across the observations obtained from k nearby plots. These values are designated by $\underline{p}'_{1i}$.

2. They also calculated $\underline{\overline{e}}_i$, *i.e.* the moving mean of the estimates of the environmental values, calculated according to Equation (14.22), across k nearby plots. From these values they obtained adjusted phenotypic values, say $\underline{p}'_{2i}$, in the following way:

$$\underline{p}'_{2i} = \underline{p}_i - b\underline{\overline{e}}_i$$

The values studied for k were 2, 4, 6,. . . , 14. In most cases adjustment yielded a reduction of the error variance. In 4 of the 8 tests adjustment $\underline{p}'_{1i}$ surpassed adjustment $\underline{p}'_{2i}$, in two tests the reverse occurred. The optimal value for k ranged from 4 to 14.

Comparison of the group of lines selected on the basis of $\underline{p}'_{1i}$ with the group of lines selected on the basis of $\underline{p}'_{2i}$ showed that these two groups coincided largely for $k > 6$. (The test involved only $J = 2$ replications. The mean squared errors of the estimators for g and e_i are then rather high. This low value for J may explain why $\underline{p}'_{2i}$ did not give rise to a lower error variance than $\underline{p}'_{1i}$).

It may be questioned whether a moving mean adjustment requires complete randomization. This should carefully be considered. If the breeder wishes to distinguish different classes of plant material, *e.g.* classes differing in earliness, plant height, *etc.*, he or she will avoid complete randomization and test related or similar entries as groups together. The candidates belonging to the same group will tend to have similar genotypes. The entries selected, either or not on the basis of a moving mean adjustment, may then represent entries tested at plots affording favourable growing conditions.

It is, however, concluded that phenotypic values adjusted on the basis of a moving mean adjustment tend to yield better indications of the genotypic values of the candidates than phenotypic values adjusted on the basis of data from plots containing a standard variety.

The appendix illustrates how a few moving mean adjustment procedures perform with regard to disclosing genotypic values of candidates.

APPENDIX: A study to the relative merits of adjustment procedures involving moving means

In this appendix some procedures for the adjustment of phenotypic values on the basis of moving means are illustrated by means of an artificial example. The relative merits of the procedures are presented simply as ‘food for thought’, not as well established results.

We consider the following model for the phenotypic value of the candidate occurring at plot i:

$$\underline{p}_i = \underline{\mathcal{G}}_i + \underline{E}_i + \underline{e}_i \tag{14.23}$$

where

$\underline{E}_i :=$ the contribution due to the quality of the growing conditions varying with a trend
$\underline{e}_i :=$ the contribution due to the quality of the growing conditions varying at random.

The variables $\underline{p}_i$ and $\underline{\mathcal{G}}_i$ have their usual meaning

Table 14.3 presents for plot i, where $i = 1, \ldots, 36$, the phenotypic value $(\underline{p}_i)$ as well as its components $\underline{\mathcal{G}}_i, \underline{E}_i$ and $\underline{e}_i$. The values for $\underline{\mathcal{G}}_i$ and for $\underline{e}_i$. were drawn from the normal distributions $N(10,1)$ and $N(0,1)$, respectively. The values for E_i were simply calculated as $10sin(10i)$. For plot 6, for example, E_6 was calculated to amount to $4sin(60) = 3.46$. The 18 candidates were coded 1 to 18. They were tested in each of two blocks: block I comprised plots, 1, ..., 18 and block II comprised plots 19, ..., 36. The 36 plots occurred alongside one strip.

The table shows that candidates 8, 5 and 9 have the highest genotypic values. According to the phenotypic values candidates 8, 9 and 10, occurring in neighbouring plots, are the most attractive.

Table 14.4 presents, for each candidate, the values for $\underline{p}_i$ and $\overline{\underline{p}}_i$. The latter values are averages calculated across $k = 6$ nearby plots. These nearby plots are usually plots $i-3, i-2, i-1, i+1, i+2$ and $i+3$, but for the end plots, *i.e.* plot 1, 2, 3, 4, 5, or 6, they are the six plots nearest to the considered plot. (For end plots one may, as an alternative, calculate $\overline{\underline{p}}_i$ also from only the three, four or five plots within three plots distance.)

The candidates with the lowest and the highest mean phenotypic values are candidate 17 and 5, respectively. The candidates with the lowest and the highest genotypic values are candidates 18 and 8, respectively. The coefficient of correlation of $\underline{p}_i$ and $\overline{\underline{p}}_i$, where $i = 1, \ldots, 36$, can be estimated to amount to 0.81^{**}; amply larger than 0.50, the minimum required for successful application of a moving mean adjustment (Mak, Harvey and Berdahl, 1978). The coefficient of correlation of $\underline{\mathcal{G}}$ and $\underline{p}$ amounts to 0.78 in block 1 and to 0.60 in block 2.

We now consider adjustment on the basis of the moving mean $(\overline{\underline{p}}_i)$, namely adjustment on the basis of Equation (14.20):

$$\underline{p}_{1i}{}' = \underline{p}_i - \overline{\underline{p}}_i$$

The adjusted phenotypic values are presented by Table 14.4. The coefficients of correlation of $\underline{\mathcal{G}}$ and $\underline{p_1}'$ amount to 0.65 both in block 1 and block 2. Adjustment of individual plot data was thus not clearly advantageous for the considered data.

Table 14.3 A trial field consisting of a strip of 36 plots ($i = 1, \ldots, 36$), *i.e.* two blocks of 18 plots each. Indicated are the phenotypic values (p_i) of the 18 tested candidates and their components $\mathcal{G}_i, E_i$ and e_i

Plot (i)	Candidate	$\mathcal{G}_i$	E_i	e_i	p_i
1	1	9.70	0.69	0.03	10.42
2	2	9.24	1.37	0.87	11.48
3	3	9.48	2.00	0.13	11.61
4	4	8.44	2.57	1.97	12.98
5	5	11.39	3.06	0.49	14.94
6	6	9.66	3.46	−1.42	11.70
7	7	10.88	3.76	−0.52	14.12
8	8	12.38	3.94	0.24	16.56
9	9	11.25	4.00	0.24	15.49
10	10	9.73	3.94	1.51	15.18
11	11	8.41	3.76	−0.99	11.18
12	12	10.24	3.46	−0.40	13.30
13	13	10.65	3.06	−0.81	12.90
14	14	9.64	2.57	0.33	12.54
15	15	10.05	2.00	−0.87	11.18
16	16	9.10	1.37	−1.81	8.66
17	17	8.49	0.69	−1.78	7.40
18	18	8.16	0.00	1.19	9.35
19	9	11.25	−0.69	−0.89	9.67
20	6	9.66	−1.37	0.03	8.32
21	5	11.39	−2.00	1.55	10.94
22	16	9.10	−2.57	−0.37	6.16
23	3	9.48	−3.06	−1.03	5.39
24	13	10.65	−3.46	2.09	9.28
25	12	10.24	−3.76	1.23	7.71
26	11	8.41	−3.94	−0.36	4.11
27	2	9.24	−4.00	−0.69	4.55
28	14	9.64	−3.94	−1.75	3.95
29	8	12.38	−3.76	−0.14	8.48
30	4	8.44	−3.46	1.42	6.40
31	7	10.88	−3.06	−2.02	5.80
32	17	8.49	−2.57	0.62	6.54
33	18	8.16	−2.00	−0.55	5.61
34	1	9.70	−1.37	1.82	10.15
35	10	9.73	−0.69	−0.27	8.77
36	15	10.05	0.00	−1.65	8.40

When analysing the averages across the two blocks, the coefficient of correlation of $\underline{\mathcal{G}}$ and $\underline{p}$ amounts to 0.83 and the coefficient of correlation of $\underline{\mathcal{G}}$ and $\underline{p}_1{}'$ to 0.76. These figures illustrate for the present data that replicated testing is advantageous, but data adjustment did not give further improvement of the quality of the evaluation.

Table 14.4 A summary per candidate of the data presented in Table 14.3. The symbol $\overline{p}_i$ represents the moving mean across six nearby plots; $p_1{}'$ represents the adjusted phenotypic value calculated according to Equation (14.20)

Candidate		Block 1			Block 2			Averages across the two blocks		
	$\mathcal{G}$	p	$\overline{p}_i$	$p_1{}'$	p	$\overline{p}_i$	$p_1{}'$	p	$\overline{p}_i$	$p_1{}'$
1	9.70	10.42	12.81	−2.39	10.15	6.92	3.23	10.29	9.87	0.42
2	9.24	11.48	12.63	−1.15	4.55	6.65	−2.10	8.02	9.64	−1.63
3	9.48	11.61	12.61	−1.00	5.39	7.75	−2.36	8.50	10.18	−1.68
4	8.44	12.98	12.38	0.60	6.40	5.82	0.58	9.69	9.10	0.59
5	11.39	14.94	13.08	1.86	10.94	8.03	2.91	12.94	10.56	2.39
6	9.66	11.70	14.28	−2.58	8.32	8.15	0.17	10.01	11.22	−1.21
7	10.88	14.12	14.48	−0.36	5.80	6.85	−1.05	9.96	10.67	−0.70
8	12.38	16.56	13.77	2.79	8.48	5.22	3.26	12.52	9.50	3.03
9	11.25	15.49	13.67	1.82	9.67	8.47	1.20	12.58	11.07	1.51
10	9.73	15.18	13.93	1.25	8.77	7.15	1.62	11.98	10.54	1.44
11	8.41	11.18	14.33	−3.15	4.11	6.56	−2.45	7.65	10.45	−2.80
12	10.24	13.30	13.08	0.22	7.71	5.57	2.14	10.51	9.33	1.18
13	10.65	12.90	12.01	0.89	9.28	6.48	2.80	11.09	9.25	1.85
14	9.64	12.54	10.77	1.77	3.95	6.17	−2.22	8.25	8.47	−0.23
15	10.05	11.18	10.69	0.49	8.40	6.00	2.40	9.79	8.35	1.45
16	9.10	8.66	10.51	−1.85	6.16	8.55	−2.39	7.41	9.53	−2.12
17	8.49	7.40	9.95	−2.55	6.54	7.53	−0.99	6.97	8.74	−1.77
18	8.16	9.35	9.36	−0.01	5.61	7.67	−2.06	7.48	8.52	−1.04
Mean			12.46			6.97			9.72	

The candidates with the lowest and the highest mean adjusted phenotypic values are candidates 11 and 8, respectively.
Now adjustment according to Equation (14.21) is considered. Linear regression of $\underline{p}_i$ on $\underline{\overline{p}}_i (i = 1, \ldots, 36)$ yields

$$\hat{p}_i = 0.968 + 0.904\overline{p}_i$$

with $r = 0.81$. The covariance adjusted phenotypic values are calculated as

$$\underline{p}_{2i}{}' = \underline{p}_i - 0.904\underline{\overline{p}}_i$$

These values are presented in Table 14.5.

The coefficient of correlation of $\underline{\mathcal{G}}$ and $\underline{p}_{2i}{}'$ amounts to 0.69 in block 1 ($i = 1, \ldots, 18$), and to 0.65 in block 2 ($i = 19, \ldots, 36$). These values are only marginally higher than the corresponding coefficients of correlation of $\underline{\mathcal{G}}$ and $\underline{p}_{1i}{}'$. Across all plots the coefficient of correlation is 0.66. The coefficient of correlation of $\underline{\mathcal{G}}$ and $\overline{p}_2{}'$ amounts to 0.77. This is only marginally higher than the coefficient of correlation of $\underline{\mathcal{G}}$ and $\overline{p}_1{}'$.

The candidates with the lowest and the highest mean adjusted phenotypic values are again candidates 11 and 8, respectively.

Table 14.5 A summary per candidate of the data presented in Table 14.3. The symbol $p_2{}'$ represents the adjusted phenotypic value calculated according to Equation (14.21)

		$p_2{}'$		
Candidate	$\mathcal{G}$	Block 1	Block 2	$\overline{p}_2{}'$
1	9.70	−1.16	3.89	1.37
2	9.24	0.06	−1.46	−0.70
3	9.48	0.21	−1.62	−0.70
4	8.44	1.79	1.14	1.46
5	11.39	3.12	3.68	3.40
6	9.66	−1.21	0.95	−0.13
7	10.88	1.03	−0.39	0.32
8	12.38	4.11	3.76	3.94
9	11.25	3.13	2.01	2.57
10	9.73	2.59	2.31	2.45
11	8.41	−1.77	−1.82	−1.80
12	10.24	1.48	2.67	2.08
13	10.65	2.04	3.42	2.73
14	9.64	2.80	−1.63	0.59
15	10.05	1.52	2.98	2.25
16	9.10	−0.84	−1.57	−1.21
17	8.49	−1.59	−0.27	−0.93
18	8.16	0.89	−1.32	−0.22

According to Mak, Harvey and Berdahl (1978) contributions to the moving mean which are due to the block and to the candidate should be eliminated prior to the calculation of the adjusted phenotypic values. This requires calculation of $\underline{\overline{p}}'(i,j)$ according to

$$\underline{\overline{p}}'(i,j) = \underline{\overline{p}}(i,j) - \underline{\overline{p}}(\cdot,j) - \underline{\overline{p}}(i,\cdot) + \underline{\overline{p}}(\cdot,\cdot)$$

where
$\underline{\overline{p}}(i,j)$:= the moving mean value calculated for candidate i as occurring in block j

$\cdot$ implies calculation of the mean across the involved block and/or candidate.

For candidate 1 in block 1 it can, for example (see Table 14.4), be derived that $\underline{\overline{p}}'(1,1)$ amounts to

$$\underline{\overline{p}}'(1,1) = 12.81 - 12.46 - 9.87 + 9.72 = 0.20$$

Altogether this yields the data presented by Table 14.6.

Linear regression across the two blocks of $\underline{p}_i$ on $\underline{\overline{p}}_i{}'$ yields

$$\hat{p}_i = 9.75 + 0.736\underline{\overline{p}}_i{}'$$

Table 14.6 A summary per candidate of the data presented in Table 14.3. The symbol $\overline{p}'$ represents the moving mean adjusted for block and genotype effect; $p_3{}'$ represents the adjusted phenotypic value according to Equation (14.24)

	Block 1				Block 2			
Candidate	$\mathcal{G}$	p	$\overline{p}'$	$p_3{}'$	p	$\overline{p}'$	$p_3{}'$	$\overline{p}_3{}'$
1	9.70	10.42	0.20	10.27	10.15	−0.20	10.30	10.29
2	9.24	11.48	0.25	11.30	4.55	−0.24	4.73	8.02
3	9.48	11.61	−0.31	11.84	5.39	0.32	5.15	8.50
4	8.44	12.98	0.54	12.58	6.40	−0.53	6.79	9.69
5	11.39	14.94	−0.22	15.10	10.94	0.22	10.78	12.94
6	9.66	11.70	0.32	11.46	8.32	−0.32	8.56	10.01
7	10.88	14.12	1.07	13.33	5.80	−1.07	6.59	9.96
8	12.38	16.56	1.53	15.43	8.48	−1.53	9.61	12.52
9	11.25	15.49	−0.14	15.59	9.67	0.15	9.56	12.58
10	9.73	15.18	0.65	14.70	8.77	−0.64	9.24	11.97
11	8.41	11.18	1.14	10.34	4.11	−1.14	4.95	7.65
12	10.24	13.30	1.01	12.56	7.71	−1.01	8.45	10.51
13	10.65	12.90	0.02	12.89	9.28	−0.02	9.29	11.09
14	9.64	12.54	−0.44	12.86	3.95	0.45	3.62	8.24
15	10.05	11.18	−0.40	11.47	8.40	0.40	8.11	9.79
16	9.10	8.66	−1.76	9.96	6.16	1.77	4.86	7.41
17	8.49	7.40	−1.53	8.53	6.54	1.54	5.41	6.97
18	8.16	9.35	−1.90	10.75	5.61	1.90	4.21	7.48

with $r = 0.21^{\text{ns}}$. The symbol $\hat{p}_i$ represents the phenotypic value, predicted on the basis of the trend in the quality of the growing conditions, for the candidate tested at plot i, irrespective of its genotype. This trend is indicated by the moving mean $\underline{\overline{p}}_i{}'$.

To evaluate the genotypic value of the candidate occurring at that plot i ($\underline{\mathcal{G}}_i = \underline{p}_i - \underline{E}_i - \underline{e}_i$; see Equation (14.23)) as well as possible one may consider the difference between $\underline{p}_i$ and $\hat{p}_i$ (like in Equations (14.10a) and (14.10b)):

$$\underline{p}_{3i}{}' = \underline{p}_i - \underline{\hat{p}}_i = \underline{p}_i - 0.736\underline{\overline{p}}_i{}' \tag{14.24}$$

Table 14.6 presents the phenotypic values adjusted in this way. For these adjusted phenotypic values the coefficient of correlation with the genotypic values amounts to $r = 0.83$, *i.e.* higher than the correlations of $\underline{p}_i$, or $\underline{p}_{1i}{}'$, or $\underline{p}_{2i}{}'$ and $\underline{\mathcal{G}}$. The candidates with the lowest and the highest mean adjusted phenotypic values are now candidates 17 and 5, respectively (like for $\underline{p}_{1i}{}'$).

The replicated testing permits the estimation of the phenotypic values presented in Table 14.3 as well as, for corresponding adjusted phenotypic values, analyses of variance. These give rise to the following values for the estimates of the error variance (s^2) and for the F value (F):

	s^2	F
p	3.123	2.318^{*}
$\underline{p}_1{}'$	2.204	2.699^{*}
$\underline{p}_2{}'$	2.126	2.799^{*}
$\underline{p}_3{}'$	2.074	3.587^{**}

These results apply to the data presented by Table 14.3. They show that adjustment may give rise to reduction of the error variance and – consequently – higher F values. The power of the test is promoted by the data adjustment. It is, however, emphasized that these results are been obtained for a specific set of data.

Chapter 15
Reduction of the Detrimental Effect of Allocompetition on the Efficiency of Selection

In the early generations of a plant breeding programme selection is mostly based on observations of individual plants or on small plots. This is due to lack of seed, to large numbers of candidates to be tested, and/or to limited resources. The efficiency of selection for yield tends then to be very low. Low or even negative correlations between yields of single plants and their progenies have been reported.

Variation in the quality of the growing conditions is one of the factors responsible for such low efficiency of selection. Chapter 14 deals with procedures to improve the efficiency of selection under such conditions. Intergenotypic competition is another factor responsible for the low efficiency of selection among candidates represented by individual plants or by small plots. The best competing candidates tend to have the highest phenotypic values for yield. This implies that preferentially candidates with a high competitive ability are selected instead of candidates with a high potential yield in the absence of intergenotypic competition. Indeed, competitive ability, as expressed in a genetically heterogeneous population grown at a high plant density, tends to be negatively correlated with yielding capacity at high plant density in the absence of genetic variation.

For this reason selection in the absence of competition, i.e. selection at a very low plant density, has been advocated. In contrast it has been stated that at low plant density the bias due to competitive ability, which occurs when selecting at high plant density, is replaced by a bias due to genotype × density interaction: selection at low plant density would imply selection of candidates only performing in a superior way when grown at low plant density.

This chapter gives attention to the negative effect of intergenotypic competition on the efficiency of selection. It is shown how the breeder might reduce the detrimental effect.

15.1 Introduction

The environmental conditions of a plant comprise both physical growth factors and the growth habits of neighbouring plants. Interference between plants consists usually of competition for the same growth requisites, like water, light and mineral nutrients. These are present in a limited supply. The competition between plants for the limited resources result in an uneven sharing of these

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 381–403.

resources (Spitters, 1979, p. 1). The competition is stronger as the **plant density**, *i.e.* the number of plants per unit area, is higher or as, at a given plant density, the canopy of the plants increase or the amount of resources is reduced.

Commercial cultivation of crops generally consists of the growing of monocultures at high plant density. These monocultures consist of genetically homogeneous plant material, *viz.* clones or pure lines, in the case of vegetatively reproducing plant material or self-fertilizing crops. In the case of many cross-fertilizing crops, however, they also consist of genetically homogeneous plant material, *viz.* single-cross hybrid varieties, or of fairly homogeneous plant material, *viz.* (in decreasing order):

- Three-way cross hybrids
- Double-cross hybrids
- Synthetic varieties
- Open pollinating varieties

The cultivation at high plant density implies the presence of strong interplant **competition**. Because of the absence of genetic variation (or nearly so) the competition is called **intragenotypic competition**, or **isocompetition**. Isocompetition is thus competition among plants with the same genotype. It concerns competition within clones, pure lines and single-cross hybrids. It is the competition commonly occurring within cultivars.

When developing new varieties, breeders often apply in the selection field, the physical growing conditions prevailing in commercial cultivation. Selection occurs, consequently, at high plant density, *i.e.* in the presence of interplant competition among genetically diverse candidates. (With regard to the latter conditions, the conditions in the selection field differ from the conditions occurring at commercial cultivation.) The individual plants, clones, lines, families or hybrids are thus evaluated when being subjected to **intergenotypic competition** (or **allocompetition**). Allocompetition is thus competition among candidates with different genotypes. It is the competition commonly occurring in breeding nurseries.

Caligari (1980) pointed out that plants of more related genotypes tend to compete more strongly than plants of unlike genotype: the closer the genetic relationship between plants, the more similar, at a given date, their requirements from the environment. This implies that, at the same plant density and at the same amount of nutrients, isocompetition may have more severe effects with regard to fitness, *i.e.* with regard to the proportion of surviving plants (vitality) and/or the number of viable seeds produced (fertility), than allocompetition.

The strength of isocompetition or allocompetition may be measured in the following way. When plant material representing the same single genotype is grown in a series of densities and the studied character, for instance the portion of surviving plants or the average single-plant biomass, shows a linear

regression on the density the regression coefficient b_m provides a measure of the strength of isocompetition (Mather and Caligari, 1981).

Effects of allocompetition have been studied by growing mixtures of different genotypes, or even different species, which can be identified in the mixtures. If a variable number (x) of plants with genotype B is added to a reference number of plants with genotype A, one may calculate b_d, *i.e.* the regression on x of the expression of the character observed on the plants representing genotype A. The difference $b_d - b_m$ provides then a means of measuring the relative strength of allocompetition as it affects A.

In this chapter attention is focussed on allocompetition as a factor interfering with artificial selection. Numerous studies of the effect of allocompetition on the efficiency of selection have been conducted with commercial varieties (reviewed by Spitters, 1979). These cultivars represent genotypes that survived during their development the full selection process. They have been selected and were, consequently, successful under allocompetition. They also have been tested and selected under isocompetition. Studies involving mixtures of commercial varieties are thus not representative of sets of random inbred lines (as obtained by single-seed descent or by doubling the chromosome number of haploid plants). This should be kept in mind when appreciating results of competition experiments (Powell *et al.*, 1985a).

Another common feature of competition studies is the use of a restricted number of genotypes. This usually arises from the practical difficulty of handling a large number of different genotypes but does give rise to concern about the representiveness of the sample of genotypes to more general situations.

In groups of plants which are in competition with one another, each plant plays a dual part: it exerts competitive pressure on its fellows and at the same time it responds to competitive pressure from them. It is possible to quantify the competitive pressure exerted (aggression) and the response to the competitive pressure experienced (response) (Mather and Caligari, 1983). However, for a particular genotype it is the balance between these forces that determines whether it will survive. The most successful competitor is likely to be one that combines strong aggression with a low response (Hill et al., 1987). It has been shown that aggression and response can vary independently. This implies that each of these two items is subject to its own genetic control. They should, therefore, be separately adjustable by selection (Powell *et al.*, 1985a).

As competition concerns both below- and above-ground factors, competitive ability concerns both below-ground and above-ground aspects of plant growth (Example 15.1).

Example 15.1 Satorre and Snaydon (1992) used boxes to separate the effects of above-ground (shoot) and below-ground (root) competition between a barley, wheat or oats variety and the common wild oat (*Avena fatua*). They found that the severity of root competition was greater than the severity of shoot competition. Thus competition for soil resources was

greater than that for aerial resources. Indeed, soil resources are often limiting. The most marked effect of root competition from *A. fatua* was on the number of grains per ear. All three cereal species were more competitive than *A. fatua*, which has a slow early root development.

The differences in shoot competitive ability between the cereals were only partially related to plant height.

In this chapter the balance between aggression and response is indicated as **competitive ability**. Ideas about how to deal with competition are exposed without elaborating procedures for estimating aggression, response or competitive ability.

Following Spitters (1979), in this chapter the terms **monoculture** and **mixture** are used with a more restricted meaning: the term monoculture will refer to genetically homogeneous plant material (mostly a pure line variety) and the term mixture to genetically heterogeneous plant material. Thus within a monoculture isocompetition will occur exclusively; whereas allocompetition will occur within a mixture.

The competitive ability of a certain candidate (grown as a single plant or as a plot) is expressed if the candidate interferes with surrounding candidates. With a regular pattern of plant positions each plant has the same number of neighbours within a certain distance. Then the competitive ability of a plant is not confounded with the number of neighbours.

In the literature it is often reported that traits like biomass, ear weight and grain yield per plant are strongly affected by competition. Number of ears per plant is influenced to a somewhat lower degree. In the experiments described by Spitters (1979), traits like number of grains per ear or harvest index were hardly affected by competition. The competitive ability was not clearly related to any of these traits. Sakai (1961) concluded from his experiments that 'competitive ability was not associated with morphological traits which might be supposed to favour competition'. Spitters remarks that in his experiments 'differences in competitive ability between barley genotypes could be mainly ascribed to differences in **juvenile growth**.' And: 'Relating competitive ability to characters that express themselves late in the development is doomed to fail' (Spitters, 1979, p. 186). Also in this book it is assumed that, at a regular pattern of plant positions, the competitive ability is mainly determined by juvenile growth.

At a high plant density and at a regular pattern of plant positions, the success of the juvenile growth is determined by the following factors:

1. Pre-seedling conditions like
   - Seed size
   - Depth of sowing
   - Orientation of the seed after being sown, *e.g.* upside down or not

These conditions are mainly of an environmental nature. They affect the date of emergence. This is a very important factor with regard to the success of juvenile growth. Harper (1977, p. 165) stated: 'The advantage which an early emerging seedling gains is far greater than can be accounted for merely by the greater time that it has been allowed to grow'.

2. Juvenile plant traits like
   - Date of emergence
   - Growth rate of the seedling
   - (Early) plant height

These conditions are mainly genetically controlled. The first two aspects concern juvenile plant growth. Successful juvenile growth may be due to an early germination of the seeds. The seedlings grow quickly and develop into leafy, early shooting plants. Early plant height may determine the competitive ability, but it may reflect its effect as well. It is a matter of speculation whether it is a cause of or a consequence from variation in competitive ability.

Variation with regard to the factors mentioned above induces variation in the success of juvenile growth. It induces variation in competitive ability. In cereals, and many other crops, a positive correlation between plant height and yield is commonly observed, both in monocultures and in mixtures. It may be due to variation in competitive ability.

The status for each of the six factors mentioned above may be due to environmental conditions or to a combination of environmental conditions and genetic disposition. In monocultures the variation in competitive ability is entirely due to environmental variation for these factors. In mixtures it is due to environmental and genetic variation with regard to these factors. Example 15.2 presents some data concerning the correlation between juvenile and/or adult plant traits.

Example 15.2 Evans and Bhatt (1977) found for wheat a high positive coefficient of correlation ($r = 0.73$) between the weights of the kernels and the vigour of the seedlings emerging from these.

Soetono and Donald (1980) studied, in barley, the relationship between date of emergence and plant performance. They found that the date of emergence significantly affected plant weight and number of grains. The earlier plants to emerge were larger at day 70 and at day 90 than those that emerged later. This applied to each of three plant densities.

Bos and Kleikamp (1985) studied, in spring rye, how variation in preseedling factors gave rise to variation for some adult plant traits, *viz.* number of tillers, plant height and grain yield. At a plant density of 205.3 plants/m^2 the coefficient of correlation between initial seed weight and plant yield amounted to 0.44; at a density of 51.3 it was 0.49.

Brown and Caligari (1986) established that the coefficient of correlation between the weight of a potato tuber, produced by a seedling, and the total

weight produced by the first-year-clonal plant obtained from it, ranged from 0.15 (site S, block 2) to 0.49 (site S, block 1) (see also Example 14.8). However, a number of genotypes performing well in the first clonal year had only a few very small tubers produced by the 'parental' seedling.

As explained, the competitive ability in a regular pattern of plant positions, is due to several factors. If, additionally, an irregular pattern of plant positions applies, the competitive ability is determined in an even more complicated way (Example 15.3).

Example 15.3 Knight (1983) observed plants belonging to a wheat cultivar broadcasted at a density of 180 kernels/m^2. The effect of date of emergence and the area available for each plant, including the shape and the position in that area of the considered plant, on plant weight and grain yield was studied. The relationship between date of emergence and plant performance was clearly negative. For a fixed date of emergence, however, the variation in plant performance was enormous.

The area per plant ranged from 2 to 249 cm^2. The average area was 61.0 cm^2, implying a density of 164 plants/m^2. The coefficients of correlation between plant area and plant weight ($r = 0.36$) or plant area and grain yield ($r = 0.35$) were highly significant. Plants occupying the same area varied greatly in performance.

The strength of the competition experienced by each individual plant was quantified in the following way. Across all neighbours within a distance of either 10 or 20 cm around the considered plant the sum:

$$z = \sum_i \left(\frac{y_i}{{d_i}^{\theta}} \right)$$

was calculated for $\theta = 0,\ \frac{1}{2}$, 1 and 2, where

$y_i :=$ the performance of neighbour plant i; and
$d_i :=$ the distance between neighbour plant i and the considered plant,

The coefficient of correlation between the performance of a plant and its z value was negative. At a radius of 10 cm the distance hardly played a role because for $\theta = 0, \frac{1}{2}$ and 1 the coefficient of correlation only ranged from -0.22 to -0.34. At a radius of 20 cm the values $\frac{1}{2}$ and 1 for θ yielded stronger correlations (range -0.08 to -0.40) than the values 0 and 2 (r ranging from -0.04 to -0.32).

Multiple linear regression of grain yield or plant weight on date of emergence, plant area and z showed plant area to be the most important predictor, but the average value for the coefficient of determination amounted to only 0.14. Simultaneously the three predictors explained only 19% of the variation

for plant performance. Date of emergence explained only 3.8%. The shape of the available area and the position of the plant in that area did not appear to be good predictors.

The frequency distribution of the observations obtained for a trait may give evidence of the strength of the interplant competition. Characteristic features of the frequency distribution are: mean, variance and skewness. At increased strength of interplant competition, due to ongoing growth or increased plant density, the frequency distribution for plant weight tends to show a stronger positive skewness (Spitters, 1979, p. 91).

Many breeders tend to provide, in the selection field, the growing conditions prevailing under commercial cultivation. This includes use of a high plant density. Candidates represented either by single plants or by small plots with a high competitive ability will then behave in a superior way. Such entries will be selected. It is questionable, however, whether genotypes with a strong competitive ability will perform in a superior way when grown as monoculture. The superiority shown by a candidate when grown in a mixture of genotypes does not necessarily imply superiority if the candidate is grown as a monoculture, *i.e.* as a variety *per se.* Candidates with a successful juvenile growth will only out-yield other candidates when occurring in a mixture (see Example 8.8, columns 2 and 4, especially variety Goudgerst).

Fasoulas (1981) held the opinion that allocompetition should be avoided when evaluating plant material with the aim to select genotypes that have high yields when grown as monoculture. In his view, selection should occur at a very low plant density. Fasoulas stated that it is impossible for breeders to provide in the selection field the growing conditions prevailing at commercial cultivation, because the required plant density implies allocompetition whereas isocompetition will occur at commercial cultivation.

According to Fasoulas, inefficiency of truncation selection at high plant density is primarily due to the failure to disclose $\mathcal{G}_{\text{mono}}$, *i.e.* the genotypic value under monoculture conditions. The unmasking of $\mathcal{G}_{\text{mono}}$ is said to require absence of interplant competition. Differences in performance are then not due to differences in competitive ability, but to differences in growing conditions and/or genotype.

Selection of superior genotypes, *e.g.* by means of honeycomb selection at a very low plant density, may yield varieties that are superior with regard to $\mathcal{G}_{\text{mono}}$. An important condition is, of course, that the ranking of the genotypes is density independent, *i.e.* that genotype × plant density interaction does not occur. Fasoulas (1981) observed absence of such interaction (Example 15.4).

Example 15.4 Fasoulas (1981, p. 50) studied the grain yield of seven maize hybrids. He found that the ranking at a low density, *viz.* 1.4 plants/m^2, coincided with the ranking at the density of commercial cultivation.

An F_2 population of cotton was grown at a plant density of 1.4. The F3 lines derived from honeycomb selected plants of the F_2 generation were tested at three plant densities: 4, 8 and 16 plants/m^2. At each density the F_3 lines descending from highly productive F_2 plants were also highly productive.

Fasoulas concluded from the results of his experiments that one should not fear that the ranking at low plant density (absence of competition) deviates from the ranking in monoculture at high plant density (presence of isocompetition). Fasoulas and Tsaftaris (1975) did not dare to draw this conclusion for genetically heterogeneous varieties (of crops such as grasses, beet and rye) implying presence of allocompetition.

Spitters (1979, p. 77), reviewing the literature on experiments with self-fertilizing cereals, mentioned the occurrence of genotype × plant density interaction. He himself also observed such interaction, see Example 15.5.

Example 15.5 Table 8.3 presents another illustration of the occurrence of genotype × plant density interaction. Comparison of conditions 1 and 4, *i.e.* cultivation at low and at high plant density respectively, shows a strong effect of the plant density on the ranking of barley variety L98. The coefficient of correlation between the ranks at the two densities amounted to 0.24 (and to 0.55 in the absence of L98).

Kelker and Briggs (1979) concluded that it is impossible to recommend for selection one single plant density, which is optimal with regard to all traits of interest. Crosbie and Mock (1979) found that the performance at a high plant density of genotypes selected at a lower plant density was disappointing. Faris and De Pauw (1981) reported the presence of genotype × plant density interaction in spring wheat. Bussemakers and Bos (1999) concluded on the basis of experiments lasting five generations that mass selection should be applied at the plant density used in commercial cultivation (Example 15.6).

Example 15.6 Bussemakers and Bos (1999) studied the effect of interplant distance on the efficiency of honeycomb selection in spring rye by performing five generations of selection at two interplant distances, *viz.* 100 cm (implying absence of intergenotypic competition) and 15 cm (implying presence of intergenotypic competition). The offspring of plants selected either at low or at high plant density were compared, both at high and at low plant density, with offspring of plant taken at random from the original population.

At high plant density offspring of plants selected at high density performed better than the original population for most of the characters observed on a per plant basis.

At low plant density offspring of plants selected either at high or at low plant density performed better than the original population for the characters recorded on a per plant basis. The selections differed, however, significantly from each other: the offspring of plants selected at low plant density

performed better. As the latter did not occur athigh density genotype × plant density interaction was indicated.

The authors concluded that mass selection should be applied at the plant density used in commercial cultivation.

Spitters (1979, p. 248) described the questions to be considered when studying the detrimental effects of environmental variation and allocompetition on the efficiency of selection: "Selection occurs in a heterogeneous population; one tries to choose the genotypes that perform best in monoculture. Therefore, the central question is: to what extent are the genotypes with the highest yield in monoculture in generation $t + 1$ chosen when selection is for the phenotypes yielding highest in a mixture in generation t? The central question is split into three:

(1) To what extent are the highest yielding phenotypes in the mixture in generation t also the highest yielding genotypes in that mixture in that generation?
(2) To what extent are the genotypes that give the highest yield in the mixture in generation t also the genotypes yielding highest in monoculture in that generation?
(3) To what extent do the genotypes selected in generation t maintain their expected monoculture yield in generation $t + 1$?

The first question refers to the degree to which the genotypes with the highest yield in the mixture are identified by selection in that mixture. The progress that is made for yielding ability in that mixture is called the **direct response** to selection. The second question defines the effect of intergenotypic competition on the outcome of selection. Selection for yield in the mixture leads to a **correlated response** for monoculture yield. The third question concerns the effect of heterozygosity and mode of reproduction". The possible complicating effect of genotype × environment interaction can, of course, be added.

Section 15.2 deals with procedures that aim to reduce the detrimental effect of interplant competition on the efficiency of individual plant selection. Section 15.3 deals with procedures aimed at reduction of the detrimental effect of interplot competition on the efficiency of selection among candidates evaluated on the basis of plots.

15.2 Single-Plant Evaluation

The detrimental effect of competition on the response to selection can be quantified by applying the theory developed for indirect selection (Section 12.3). In this section it is shown how this can be done.

In Section 15.1 it was questioned to what extent the candidates with the best phenotypic values in the mixture in generation t are also the candidates with the best genotypic values in that mixture in that generation. The response to selection in that mixture in that generation was called the **direct response** to selection. It is

$$R_{\text{mix}} = \text{E}\underline{\mathcal{G}}_{\text{mix,s}} - \text{E}\underline{\mathcal{G}}_{\text{mix}} \tag{15.1}$$

where

$\text{E}\underline{\mathcal{G}}_{\text{mix,s}}$:= the expected value of $\underline{\mathcal{G}}_{\text{mix,s}}$, *i.e.* the expected genotypic value in mixture, calculated across the selected plants

$\text{E}\underline{\mathcal{G}}_{\text{mix}}$:= the expected value of $\underline{\mathcal{G}}_{\text{mix}}$, calculated across all candidate plants

R_{mix} is the average genetic superiority of the selected plants under the environmental conditions, especially the strength of the allocompetition, to which the plants constituting the mixture are subjected.

In Section 15.1 also the relationship between the genotypic values of the candidates in the mixture in generation t and their genotypic values in monoculture in generation t was questioned. This concerns the **correlated response** to selection, *i.e.* the response when cultivating the selected candidates in monoculture, *i.e.* in the absence of allocompetition. It is measured by

$$CR_{\text{mono}} = \text{E}\underline{\mathcal{G}}_{\text{mono,s}} - \text{E}\underline{\mathcal{G}}_{\text{mono}} \tag{15.2}$$

where

$\text{E}\underline{\mathcal{G}}_{\text{mono,s}}$:= the expected value of $\underline{\mathcal{G}}_{\text{mono,s}}$, *i.e.* the expected genotypic value in monoculture, calculated across the selected plants

$\text{E}\underline{\mathcal{G}}_{\text{mono}}$:= the expected value of $\underline{\mathcal{G}}_{\text{mono}}$, calculated across all candidate plants

Example 15.7 illustrates the calculation of R_{mix} and CR_{mono}. The example shows that the plants selected under competition did not represent genotypes yielding higher in the absence of competition. The effect of allocompetition nullified the monoculture response to selection. In Example 8.8 it was illustrated that, due to allocompetition, $\text{var}(\underline{\mathcal{G}}_{\text{mix}})$ tends to be much greater than $\text{var}(\underline{\mathcal{G}}_{\text{mono}})$.

Example 15.7 Spitters (1979, pp. 159–167) applied single-plant selection in a mixture of 12 homozygous barley varieties. The mixture was grown at a pattern of $5 \times 25\,\text{cm}^2$, *i.e.* at a density of 80 plants/m^2. The selection field consisted of five grids. Each variety was represented by eight plants in each grid. The 10 top-yielding plants were selected in each grid. Altogether $5 \times 10 = 50$ out of $12 \times 5 \times 8 = 480$ plants were selected. Because the variety to which each selected plant belonged could be identified, f_i, *i.e.* the relative frequency of variety $i (i = 1, \ldots, 12)$ among the selected plants could

be determined. The quantity $\underline{\overline{p}}_{\text{mix,i}}$, *i.e.* the mean phenotypic value of the 40 plants representing variety i, was used as estimator of $\underline{\mathcal{G}}_{\text{mix},i}$. Table 15.1 presents the data required to calculate various quantities.

Table 15.1 Grain yield (in g / plant) and rank (from 1 = lowest to 12 = highest) of 12 spring barley varieties grown in 1977 under two different conditions (see text). The symbol f_i represents the relative frequency of variety $i (i = 1, \ldots, 12)$ among the 50 selected plants (source: Spitters, 1979, Tables 25, 27, Figure 32)

Variety (i)	Condition				
	Monoculture		Mixture		
	yield	rank	yield	rank	f_i
Varunda	5.3	6.5	5.1	5.5	0.04
Tamara	5.7	10	7.8	12	0.26
Belfor 5.3	5.3	6.5	5.4	9.5	0.06
Aramir	6.1	12	5.3	7.5	0.04
Camilla	5.0	5	5.4	9.5	0.06
G. Promise	4.5	1	4.9	4	0.02
Balder	4.8	4	5.1	5.5	0.08
WZ	5.5	8	4.8	3	0.04
Goudgerst	4.7	3	7.7	11	0.26
L98	6.0	11	3.5	2	0.02
Titan	4.6	2	1.6	1	0.00
Bigo	5.6	9	5.3	7.5	0.12

$$\overline{\mathcal{G}}_{mono} = 5.26 \quad \overline{\mathcal{G}}_{mix} = 5.16$$

Column 2 of the table presents for each variety the estimate of $\mathcal{G}_{\text{mix},i}$ for grain yield. Thus $\text{E}\underline{\mathcal{G}}_{\text{mix}}$ can be estimated to be

$$\frac{\sum_i \overline{p}_{\text{mix},i}}{12} = 5.16\,\text{g/plant},$$

and $\text{E}\underline{\mathcal{G}}_{\text{mix,s}}$ can be estimated to be

$$\sum_i f_i \overline{p}_{\text{mix},i} = 6.498\,\text{g/plant}$$

The direct response to selection (Equation (15.1)) is then

$$R_{\text{mix}} = 6.50 - 5.16 = 1.34\,\text{g/plant}.$$

Because the varieties were simultaneously grown as monocultures at a density of 80 Spitters could use $\underline{\overline{p}}_{\text{mono},i}$, *i.e.* the mean phenotypic value of 200 plants representing variety i, as estimator of $\underline{\mathcal{G}}_{\text{mono},i}$. Column 2 of Table 15.1

presents for each variety the estimate of $\mathcal{G}_{\text{mono},i}$ for grain yield. Thus $\mathrm{E}\underline{\mathcal{G}}_{\text{mono}}$ can be estimated to be

$$\frac{\sum_i \overline{p}_{\text{mono},i}}{12} = 5.26\,\text{g/plant},$$

and $\mathrm{E}\underline{\mathcal{G}}_{\text{mono,s}}$ can be estimated to be

$$\sum_i f_i \overline{p}_{\text{mono},i} = 5.26\,\text{g/plant}.$$

Equation (15.2) implies then that the correlated response to selection amounted to

$$CR_{\text{mono}} = 0.00\,\text{g/plant}$$

The selection under allocompetition was not at all effective with regard to monoculture yield! The main reason for this was that as many as 26% of the selected plants belonged to variety Goudgerst. The monoculture grain yield genotypic value ($\mathcal{G}_{\text{mono}}$) of this variety is, however, very low, *viz.* 4.7 g/plant.

Caligari and Powell (1986), reporting on an experiment with spring barley, stated: 'The results reinforce the general theme that early-generation selection should be avoided in barley breeding programmes. ... if selection is practised in early generations, when genotypes are present in heterogeneous mixtures, its effects will be confounded with the effects of competition in its broadest sense'.

The genotypes with the highest $\mathcal{G}_{\text{mix}}$ values are thus not necessarily also the genotypes with the highest $\mathcal{G}_{\text{mono}}$ values. Selection in a mixture aiming at improvement of monoculture performance is a form of indirect selection (Section 12.3). The predicted response in monoculture to selection in mixture is described by Equation (12.4), *i.e.* in the present context by

$$CR_{\text{mono}} = i_{\text{mix}} \rho_{g_{\text{mix}},g_{\text{mono}}} h_{w_{\text{mix}}} \sigma_{g_{\text{mono}}} \tag{15.3}$$

Example 15.8 presents an application.

Example 15.8 Spitters (1979, p. 164) presents the following estimates for the genetic parameters in Equation (15.3):

$$r_{g_{\text{mix}},g_{\text{mono}}} = 0.11$$

(This estimate is obtained from the data in Table 15.1.)

$$s_{g_{\text{mono}}} = 0.47\,\text{g/plant}$$

and

$$\hat{h}_{w_{\text{mix}}} = 0.54$$

At $i_{\text{mix}} = 1.76$ (this applies when selecting 10 %, Section 11.1) Equation (15.3) yields

$$CR_{\text{mono}} = 1.76 \times 0.11 \times 0.47 \times 0.54 = 0.05\,\text{g/plant}$$

This is very close to the actual correlated response to selection calculated in Example 15.7, *i.e.* 0.00 g/plant.

The direct response to selection, *i.e.* the response in mixture, can be predicted on the basis of Equation (11.20), *viz.*:

$$R_{\text{mix}} = i_{\text{mix}} {h_{\text{mix}}}^2 \sigma_{p_{\text{mix}}}$$

For

$$\begin{aligned} i_{\text{mix}} &= 1.76, \\ \hat{h}^2_{\text{mix}} &= 0.29 \end{aligned}$$

and

$$s_{p_{\text{mix}}} = 2.94\,\text{g/plant}$$

the predicted direct response amounts to

$$R = 1.76 \times 0.29 \times 2.94 = 1.5\,\text{g/plant}$$

This comes very close to the actual direct response, *i.e.* 1.34 g/plant.

15.2.1 The Optimum Plant Density

The topic of the optimum plant density for selection was already briefly touched upon in Section 12.3.3. Different opinions exist with regard to this topic. Spitters (1979, p. 117) concluded that selection can best be carried out at the plant density applied for commercial cultivation. Fasoulas (1981), on the other hand, advocated selection in the absence of allocompetition.

Experimental evidence is scarce and inconsistent. With regard to honeycomb selection in cereals Bos (1981, pp.150), Mitchell, Baker and Knott (1982) and Pasini and Bos (1990a,b) obtained weak indications that selection should be done at a (very) low plant density. Bussemakers and Bos (1999), see Example 15.6, however, concluded that mass selection can best be carried out at the plant density applied by growers.

Kyriakou and Fasoulas (1985) reported unambiguous results. At the 'high' plant density of 51.3 winter rye plants per m^2, the response to honeycomb selection was negative at each of three selection intensities. At the plant density of 1.4 plants/m^2 they got, however, a positive response for each selection

intensity. (The plant density used in commercial cultivation of cereals is about 250 plants/m^2). The relationship between the response to selection and the selection intensity was negative at the high plant density and positive at the low plant density. Example 15.9 present some details concerning a study to the optimum plant density for selection.

Example 15.9 Kramer (1983) simulated selection fields by mixing 12 spring wheat lines. These lines were similar for plant height and date of flowering and maturity, but for each plant in the selection fields the line to which it belonged could be identified visually. The mixtures were grown at plant densities of 400, 100, 44, 25 and 6.25 plants/m^2. The selection fields consisted of 12 grids, and each grid of four plants from each of the 12 lines. (At the 6.25 density each grid contained only one plant from each line). The 12 lines were also evaluated in monoculture. The trait studied was grain yield (in g/m^2).

Table 15.2 presents estimates for $\mathrm{E}\underline{p}$, $\mathrm{var}(\underline{e})$, $\mathrm{var}(\underline{\mathcal{G}})$ and $r_{g_{\mathrm{mix}},g_{\mathrm{mono}}}$.

Table 15.2 Estimates, for grain yield (in g/m^2) of spring wheat, of mean yield ($\mathrm{E}\underline{p}$), $\mathrm{var}(\underline{e})$, $\mathrm{var}(\underline{\mathcal{G}})$ and $r_{g_{\mathrm{mix}},g_{\mathrm{mono}}}$. Data for the derived quantities h_w, vc_e, vc_g and the relative selection efficiency (RSE) are also presented (source: Kramer, 1983)

Estimated	Plant density of the mixture (plants/m^2)					
parameter	400	100	44	25	6.25	Monoculture
$\mathrm{E}\underline{p}$	356	402	356	334	203	379
$\mathrm{var}(\underline{e})$	21,248	17,430	11,945	10,100	25,401	521
$\mathrm{var}(\underline{\mathcal{G}})$	3,018	1,584	1,422	849	413	280
$r_{g_{\mathrm{mix}},g_{\mathrm{mono}}}$	0.10	0.40	0.04	0.64	0.33	
$\hat{h}_w$	0.35	0.29	0.33	0.28	0.37	0.59
vc_e	0.41	0.33	0.31	0.30	0.25	0.06
vc_g	0.15	0.10	0.11	0.09	0.10	0.04
RSE	0.06	0.20	0.02	0.30	0.21	

From the figures for RSE, *i.e.* the ratio of the correlated monoculture response to selection in mixture and the monoculture response to selection in monoculture, it appears that the indirect selection was very inefficient, whatever the plant density. A tendency for a higher efficiency at a lower plant density may, however, be calculated: $r = -0.49$.

15.2.2 Measures to Reduce the Detrimental Effect of Allocompetition

Spitters (1979, pp. 176–192) reviewed the literature on suggestions to reduce or avoid the detrimental effect of allocompetition on the efficiency of mass selection. The considered suggestions are:

1. Application of a very low plant density in the selection field
2. Use of seeds or tubers with a uniform size or weight
3. Application of a pattern of plant positions such that alongside each row plants belonging to the population to be improved are alternated with plants belonging to a standard variety
4. Application of indirect selection for monoculture yield

Application of a very low plant density in the selection field

In the presence of interplant competition there is a tendency for plants with a high competitive ability to be selected. Generally the genotype of such plants does not give rise to a superior yield performance in monoculture. One may try to avoid this detrimental effect on the monoculture yield response of allocompetition by applying a very low plant density. However, the effect of genotype × plant density interaction may then take over as a cause of inefficiency of the mass selection. This could imply that selection at low plant density yields disappointing results, when growing the selected candidates in monoculture at high plant density. Example 15.10 presents results of a study where the genotype with the highest monoculture yield was neither predominantly selected at high plant density nor at low plant density.

Example 15.10 Spitters (1979, pp.167–176) cultivated a mixture of 12 barley varieties both at plant densities of both 80 and 3.2plants/m^2. The mixture contained varieties like Goudgerst, which has a fast emergence and development, and Titan and L98, which have slower juvenile growth.

At the high plant density, large differences in grain yield occurred. The coefficient of genetic variation (vc_g) amounted to $vc_g = \sqrt{2.04}/5.69 = 0.25$. Tamara and Goudgerst plants performed best and were predominantly selected.

At the low plant density the differences in grain yield were much smaller. The coefficient of genetic variation amounted to $vc = \sqrt{33.5}/45.5 = 0.13$, *i.e.* half as high as at the high plant density. Here Belfor and Tamara plants were predominantly selected.

According to Table 8.3, column 9, the monoculture yield rank at high plant density was 11.5 for Tamara, 3 for Goudgerst, and 10 for Belfor (1 = lowest; 12 = highest).

It can be concluded that the detrimental effect of competition at high plant density, due to variation in date of emergence and/or date of anthesis, was a greater nuisance than the disturbing effect of genotype × plant density interaction.

Across the 12 varieties the coefficient of correlation between mixture yield and monoculture yield at high plant density amounted, at high plant density of the mixture, to −0.11 and at low density of the mixture to 0.39. These estimates did not differ significantly from 0, but they suggest the

application of selection at low plant density. However, one should keep in mind that this study involved a mixture of varieties, *i.e.* a mixture of genotypes that survived during their development the full selection process. These genotypes have been selected and were, consequently, successful under allocompetition. They also have been tested and selected under isocompetition (Section 15.1).

Use of seeds or tubers with a uniform size or weight

Especially at a high plant density, plants developing from larger seeds (or tubers) tend to surpass plants emerging from smaller seeds (or tubers). This has been observed both within and across genotypes (Example 15.2). The detrimental effect of allocompetition on the efficiency of single-plant selection, in as far as it is due to differences in size or weight of seeds or tubers, may thus be reduced by grading the seeds or the tubers with regard to their size or weight and using a uniform portion to grow the selection field. However, there is a danger to this. The selected candidates may represent genotypes not deserving to be selected when grown according the commercial practice of planting less strictly graded material.

Alternated growth of plants belonging to the population to be improved and plants belonging to a standard variety

The use of a selection field where plants belonging to the population to be improved are alternated with standard plants might reduce the detrimental effect of allocompetition because it eliminates variation in the strength of the competition experienced by the candidate plants.

Alternating candidate plants with standard plants has two clear disadvantages.

1. It doubles the area of the selection field. This implies testing at less uniform soil conditions field. It also implies doubling of the amount of costs. This must be the reason why this measure is hardly applied in the case of single-plant evaluation.
2. The competitive ability of a candidate plant relative to adjacent standard plants will vary among the candidate plants. It will depend on the standard variety and it is not necessarily positively correlated with monoculture performance.

Example 15.11 reports about the results of an application of the measure.

Example 15.11 Spitters (1979, p. 178, Table 25) studied the effectivity of this measure by means of mixtures of 12 barley varieties. Within separate rows 12 plants, one from each variety, were alternated with plants of the

standard variety Varunda. Besides, the 12 varieties were grown in a normal mixture. Both types of mixture were grown at a density of 80 plants/m^2. For ear weight, the coefficient of correlation across the 12 varieties between their ratio of normal-mixture yield to alternated-mixture yield and their competitive ability amounted to $r = 0.76$. Thus the ear weight of a variety with a strong competitive ability tended to be higher in the normal mixture than in the alternated mixture, and – conversely – the ear weight of a variety with a weak competitive ability tended to be lower in the normal mixture than in the alternated mixture.

This tendency was not confirmed by another experiment. Altogether Spitters concluded that alternating candidate and standard plants is useless to reduce the detrimental effect of allocompetition on the efficiency of selection.

Application of indirect selection for monoculture yield

In Section 15.2 indirect selection for monoculture yield at high plant density, via selection for a trait observed in mixture, was considered. Of course one does not know in advance which trait should be used as auxiliary trait. One may speculate that mixture yield will tend to be the best. Then the detrimental effect of allocompetition on the efficiency of selection cannot further be reduced by indirect selection.

Spitters (1979, pp. 187–192) proposed harvest index (grain yield/biomass, HI) as auxiliary trait for indirect selection aiming at high monoculture yield at high plant density (Y_{mono}). He established for barley that the harvest index of a given genotype is not affected by the strength of (allo)competition experienced. Thus

$$HI_{\text{mix}} = HI_{\text{mono}} \tag{15.4}$$

for the considered genotype. If, additionally, monoculture biomass (B_{mono}) is constant across genotypes, then the relation

$$Y_{\text{mono}} = B_{\text{mono}} \times HI_{\text{mono}} = \text{const.} \times HI_{\text{mix}} \tag{15.5}$$

implies for the considered genotype that Y_{mono} is linearly related to its value for HI_{mix}. This applies whatever the genotype.

On the basis of this – somewhat speculative – reasoning, indirect selection for Y_{mono} via selection for HI_{mix} is promising. The degree to which B_{mono} is really constant across genotypes is, of course, decisive. Example 15.12 supports the reliability of using HI_{mix} as auxiliary trait.

Example 15.12 Nass (1980) pursued a higher grain yield in spring wheat. He applied indirect selection in F_2 populations by selecting for a high harvest index. This was done at normal plant density (59 seeds/m alongside a row) and at low plant density (6 seeds/m). The indirect selection was successful, especially at normal plant density.

15.3 Evaluation of Candidates by Means of Plots

In the early phase of a selection programme the number of candidates (clones, lines or families) is still high. The candidates may then be evaluated by mean of plots, often consisting of a single row of plants. The result of the evaluation is then affected by allocompetition among the candidates. Also in this situation allocompetition has a detrimental effect on the efficiency of the selection. This is illustrated by Example 15.13.

Example 15.13 Spitters (1979, pp.210–219, Table 38, Fig. 44) evaluated 12 barley varieties by means of single-row plots. Each plot had a size of $0.2 \times 2\,\mathrm{m}^2$ and was sown with 100 kernels belonging to a single variety. Each complete block comprised 12 plots. Twelve blocks, located alongside one strip, formed a grid. Four of such grids comprised altogether $4 \times 12 \times 12 = 576$ plots.

When applying grid selection, the varieties Tamara and Bigo were most frequently selected. In monoculture at commercial plant density these varieties ranked 11.5 and 8.5 (when 1 = lowest and 12 = highest) (Table 8.3, column 9). Camilla, also with monoculture rank number 11.5, was not selected. However, L98 and Goudgerst, ranking with regard to monoculture yield only 1 and 3, were selected.

As in the case of mass selection (Example 15.8) a fair similarity between the actual direct response (R_{mix}) and the predicted direct response ($\hat{R}_{\mathrm{mix}}$) appeared: when selecting within each grid the top 10% of the candidates both R_{mix} and $\hat{R}_{\mathrm{mix}}$ amounted to 17%. However, the actual correlated response under monoculture conditions (CR_{mono}) amounted to only 8%.

Spitters (1979, pp. 13–16, 225–232) discussed suggestions in the literature that aim at reducing or avoiding the detrimental effect of allocompetition on the efficiency of selection of candidates evaluated by means of plots. The considered suggestions were

1. Use of multi-row plots
2. Selection on the basis of the observations obtained from the central row(s) of multirow plots
3. Use of a large interplot distance
4. Evaluation of candidates on the basis of single-row plots alternated with single-row plots containing a standard variety
5. Application of indirect selection for monoculture yield
6. Grouping of candidates, which are similar with regard to a trait affecting competitive ability, followed by within-group evaluation
7. Growing of a fixed number of plants per plot

Use of multirow plots

Yield data from multirow plots estimate the monoculture yield performance with a smaller mean squared error than yield data from single-row plots, implying reduction of the residual variance of yield per unit area (Section 16.2.3). The use of multirow plots gives thus rise to a higher heritability, *i.e.* to a higher efficiency of the selection. However, the use of multirow plots may imply reduction of the number of replications.

The size of the plots is further considered in Section 16.2.

Selection on the basis of the observations obtained from the central row(s) of multirow plots

Yield data from the central row(s) of multirow plots yield an unbiased estimate of monoculture yield performance. The involved mean squared error of yield per unit area is, however, larger than the mean squared error of yield per unit area as calculated for the whole plot. Spitters (1979, p. 226) derived, in mathematical terms, the condition for selection based on observations representing the whole plot to have a higher response than selection based on observations obtained from the central row(s). He concluded that selection on the basis of whole plot yield data tends be give rise to a higher response than selection based on central row(s) yield data. For his own experiments Spitters established that selection based on three-row plot data was to be preferred. Apparently the advantage of a smaller residual variance of yield per unit area applying to whole-plot yield data was larger than the disadvantage of the bias due to allocompetition. Example 15.14 describes an experiment dedicated to the present issue.

Example 15.14 Bradshaw (1986) tested 29 fodder kale (*Brassica oleracea* var. *acephala* (D.C) Alef.) varieties by means of single-row plots as well as by means of five-row plots where only the central row was harvested. The coefficient of correlation across the varieties between the yield data obtained for the two methods of evaluation amounted only to 0.20. When considering only single-row plots, the coefficient of correlation between yield and plant height amounted to 0.89, for the other evaluation procedure it was 0.19^{ns}.

According to this experiment selection for yield on the basis of observations obtained from single-row plots is inefficient. When using such yield data as a basis for selection tall types may be favoured. This could imply an increased risk of lodging.

Use of a large interplot distance

Allocompetition does not occur if a large interplot distance, including the area used as an alley, is applied. A bias of the evaluation of the monoculture

performance, occurring in the presence of allocompetition, is then eliminated. However, the use of a large interplot distance may be accompanied by introduction of a bias due to genotype × interplot distance interaction.

According to Spitters (1979, pp. 227–228) the bias replacement is nearly complete. He argued that evaluation by means of single-row plots should occur at the row spacing applied in commercial cultivation. Example 15.15 describes a situation where a large interplot distance is applied as a routine.

Example 15.15 In the semi-arid conditions of Central Canada the soil moisture in the alleys around the plots provides advantageous growing conditions for the plants at the periphery of the plots. To prevent effects of a differential response of spring wheat candidates to the space provided by the alleys around the plots, at the Swift Current Research Station these alleys are drilled (in the spring!) with winter wheat, which stays – because vernalization does not occur – in the vegetative growing stage. An improved accessibility of the trial is an additional advantage of this measure.

Evaluation of candidates on the basis of single-row plots alternated with single-row plots containing a standard variety

Spitters (1979, p. 228) considered the alternating of single-row candidate plots with single-row plots containing a standard variety with an intermediate competitive ability. The only effect of the common 'genetic' environment of all candidates he observed was a reduction of the environmental variance to the level occurring in monoculture. This advantage is small and it is questioned whether it cancels out the increase of the environmental variance due to the doubling of the area of the trial field (which implies also a doubling of the costs of the evaluation). Example 15.16 describes experiments where genotypes were evaluated in a common genetic environment.

Example 15.16 Bradshaw (1986) evaluated the yield potential of 16 fodder kale varieties by means of five-row plots. The row length was 6 m, the interrow distance 0.5 m. Graded seed was used for a 6.25 cm seed spacing within the rows. Three evaluation procedures were compared in an experiment with three replication:

1. Within each plot rows 1, 2, 4 and 5 contained the short variety Maris Kestril, whereas row 3 was planted with the variety to be evaluated.
2. Within each plot rows 1, 2, 4 and 5 contained the tall variety Vulcan, whereas row 3 was planted with the variety to be evaluated.
3. Within each plot all five rows are planted with the variety to be evaluated.

The coefficient of correlation, across the varieties, between the yield data obtained from evaluation procedures 1 and 3 amounted to only 0.56; for

procedures 2 and 3 it amounted to 0.64. There was a statistically significant cultivar × guard variety interaction. The evaluation of the varieties by means of single-row plots cultivated in between four-row standard plots did thus not reflect the ranking at monoculture conditions as appearing from evaluation procedure 3. It was concluded that the use of single distinct cultivar to guard single-row plots was not a satisfactory solution to the problem of interplot competition. The competition for water during the dry weather in the summer was probably more important than shading by a tall variety in the autumn.

Spitters (1979, p. 243) compared in his barley experiments two evaluation procedures:

1. Single-row plots
2. Single-row plots alternated with single-row standard

With procedure 1 the environmental variance was somewhat higher than with procedure 2. The genetic variance was somewhat lower. This gave rise to a lower heritability. Procedure 1 yielded altogether a slightly lower correlated monoculture response than procedure 2.

Evaluation of candidates such that these are subjected to the same competitive stress is pursued by an experimental design called nearest neighbour balance (Dyke and Shelley, 1976). In this design each candidate has each of the other candidates equally often as neighbour to the left and to the right. For eight entries, for example, this design consists of

$$7(72385614)(47512683)(35467821)(17324865)(58741362)(28431576)$$
$$(64253718)(81634527)7$$

The eight complete blocks are represented within brackets. The very first and the very last plots, both containing candidate 7, are required for the balance. Such a balance may be required when the candidates differ in plant height. The degree of exposure of a candidate to the sun is then fully due to the plant height of the considered candidate and not to the plant height of neighbours.

Due to the required costs and/or seed, this design will only rarely be used by a plant breeder.

Indirect selection for monoculture yield

If the mixture harvest index of a candidate is equal to its monoculture harvest index (Spitters, 1979, p. 229), *i.e.* $HI_{\text{mix}} = HI_{\text{mono}}$, then the detrimental effect of allocompetition on the efficiency of selection may be avoided by indirect selection for monoculture yield (Y_{mono}) via selection for HI_{mix}. This is based on Equation (15.5)

$$Y_{\text{mono}} = B_{\text{mono}} \times HI_{\text{mono}} = B_{\text{mono}} \times HI_{\text{mix}}$$

where B_{mono} represents biomass in monoculture. The relative efficiency of such indirect selection depends on the coefficient of genetic correlation between Y_{mono} and $HI_{\text{mix}}(\rho_g(Y_{\text{mono}}, HI_{\text{mix}}))$ and on the ratio of the heritability of HI_{mix} to the heritability of Y_{mono} (Section 12.3). If B_{mono} is constant across the candidates, *i.e.* $\rho_g(Y_{\text{mono}}, HI_{\text{mix}}) = 1$, then

$$Y_{\text{mono}} = B_{\text{mono}} \times HI_{\text{mono}} = \text{const.} \times HI_{\text{mix}}$$

Spitters (1979, p. 191) admitted that the assumption of constancy of biomass across genotypes was not always valid, but harvest index was found not to be influenced by allocompetition, nor by isocompetition. The relative efficiency of the indirect selection is high if $h^2(HI_{\text{mix}})$ is (much) larger then $h^2(Y_{\text{mono}})$.

Grouping of entries, which are similar with regard to a trait affecting competitive ability, followed by within-group evaluation

Grouping of candidates having a similar competitive ability can be pursued by grouping them according to traits such as seed size, seed quality, date of emergence, growth habit (*e.g.* prostrate or erect), plant height or date of maturity. Plant height is often considered as a trait indicating competitive ability. Thus lines of cereals, for example *Triticale*, may be grouped according to their genotype for height reducing loci (Kempton et al., 1986). When selecting the most attractive candidates within each group, instead of selecting among ungrouped candidates, the breeder avoids selection of candidates which are attractive because of their competitive ability relative to their neighbours.

Growing of a fixed number of plants per plot

Within a certain range of plant densities the competitive ability of a candidate being tested in a certain plot is higher as the number of plants in that plot is higher. Variation among plots in the number of plants they contain contributes, consequently, to a competitional bias of the evaluation. One may try to eliminate variation in competitive ability, in as far as determined by variation in the number of germinating seeds per plot, by sowing in each plot the same number of germinating seeds.

The number of plants per plot will, nevertheless, vary for several reasons. At an insufficient depth of sowing germinating seeds may be consumed (by birds, mice, *etc.*) or seedlings may wither. Too great a depth of sowing may lead to unsuccessful emergence. After emergence other accidents may happen. Sometimes plants are damaged or (partly) consumed by snails, mice, birds, moles, rabbits, hares, *etc.*

Plants take advantage of any nearby unoccupied plant position. This occurs especially within rows, but also between rows. This compensation hampers correct adjustment for variation in number of plants per plot. A positive correlation is to be expected between plot yield and number of plants in the plot.

Evaluation of candidates on the basis of plot yield is then biased. Simultaneously, the correlation between mean single-plant yield, calculated across the plants in a plot, and the number of plants in the plot may be negative. Evaluation of candidates on the basis of mean single plant yield per plot is then biased as well. Example 15.17 presents some estimates of coefficients of such correlations.

Example 15.17 Bos (1981, Table 42) reported positive coefficients of correlation across plots between the number of winter rye plants per plot and total grain yield per plot (r ranging from 0.22 to 0.38) as well as negative coefficients of correlation between number of plants per plot and yield per plant (r ranging from -0.12 to -0.35).

Adjustment of the yield of a plot by an analysis of covariance, using the number of plants in the plot as covariate, may be incorrect if the number of plants in the plot reflects the intrinsic genotypic value of the candidate occurring at the considered plot, *e.g.* for seed quality. Thus one should first study whether the candidates differ significantly with regard to the trait used as covariate.

Chapter 16
Optimizing the Evaluation of Candidates by means of Plots

At the end of Section 8.1 it was indicated why selection tends to be an inefficient process. Section 12.1 presented additional causes for this phenomenon. Breeders should therefore make efforts to promote the efficiency of selection in a situation where opportunities to be successful are apparently unfavourable.

In the two preceding chapters the topic of disclosure of the genotypic values of candidates in situations where their phenotypic values strongly depend on the quality of the growing conditions, including the strength of the competition exerted by nearby fellow candidates, has been thoroughly considered. However, the evaluation of the genotypic values of candidates needs also to be optimized with regard to other points of view. A few of these are considered in this chapter.

The optimum number of plots per candidate, i.e. the optimum number of replications, is considered in Section 16.1. Section 16.2 gives attention to the size, the shape and the positioning of the test plots. The optimum plot size from an economic point of view is also considered (Section 16.2.3).

16.1 The Optimum Number of Replications

In the initial phases of a breeding programme the number of candidates to be evaluated is often very large. Selection is then mostly based on a single-site, single-year evaluation of the candidates. For such an evaluation a choice with regard to J, the number of plots per candidates, *i.e.* the number of replications, should be based on consideration of the following aspects:

1. The total number of plots allowed to evaluate candidates representing the considered crop
2. The optimum size and shape of the plots
3. The amount of seed available per candidate

These aspects should preferably be considered simultaneously. The number of replications that can actually be applied, whatever the optimum number of replications, is, for instance, limited by $J_{\max}$, *i.e.* the ratio of the amount of seed available per candidate to the amount of seed required for a single plot. The latter amount of seed depends, of course, on the plot size. (The latter subject is considered in Section 16.2.)

To avoid the complications arising when considering these three aspects with regard to the choice of J simultaneously, the second aspect is considered in a separate section. The present section considers therefore a way to determine

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 405–420.

the optimum value for J, say J_{opt}, in the situation where total number of plots allowed to evaluate candidates representing the considered crop (N) is given. Mostly N is limited. It is often determined by policy makers higher in the hierarchy of the organization than the breeder.

The question facing the breeder is then: How to allocate these N plots? Should (s)he use them for a non-replicated test of N candidates, or for a test in duplicate of only $\frac{1}{2}N$ candidates? When evaluating each candidate by means of J plots, the number of candidates that can be evaluated (C_J) amounts to

$$C_J = \frac{N}{J} \tag{16.1}$$

The quantity J_{opt} is defined as the value for J such, that the ratio of the response to selection expected when selecting among C_J candidates each evaluated by means of J plots (R_J), to the response to selection expected when selecting among N candidates, each evaluated at a single plot (R) is maximal. The present section gives thus attention to optimizing the number of replications from a genetic point of view. Section 16.2.3 considers the determination of J_{opt} from an economic point of view.

N.B. In this section symbols without a subscript refer to non-replicated testing ($J = 1$) and symbols supplied with the subscript J to replicated testing ($J \geq 2$).

When applying Equation (11.20) the value for J yielding the maximum value for the ratio

$$\frac{R_J}{R} = \frac{{h_J}^2 S_J}{h^2 S} = \frac{i_J {h_J}^2 \sigma_J}{i h^2 \sigma} \tag{16.2}$$

is taken to be J_{opt}. In the present section J_{opt} is derived for the situation where each of the candidates is evaluated by means of $J = 1$, or 2, or 3 or 4 plots. A larger number of plots per candidate is considered to be unrealistic.

J_{opt} is derived under the following conditions:

1. The number of plots used in the evaluation (N) is fixed. It does not depend on the number of replications. Thus

$$N = C_J \times J$$

2. The amount of seed allocated to a plot does thus not depend on the number of replications. In other words: the plot size does not depend on J.
3. The breeder decides beforehand to select the n most attractive candidates, whatever the number of candidates being evaluated. Thus the portion

$$v_J = \frac{n}{C_J} \tag{16.3}$$

is selected. Equation (16.3) implies that this portion is equal to J times the portion (v) that would be selected when evaluating N candidates:

$$v_J = J\left(\frac{n}{N}\right) = Jv \tag{16.4}$$

N.B. It is useless to have derived a value for J_{opt} larger than $J_{\max}$, where $J_{\max}$ is equal to $J_{\max} = \frac{\text{Available amount of seed}}{\text{Amount of seed required per plot}}$

In Section 11.2.1 it was indicated that the residual variance of a mean across J plots is

$$\frac{\sigma^2}{J}$$

The largest marginal decrease of the residual variance of the mean occurs when applying $J = 2$ instead of $J = 1$.

N.B. It is, by the way, impossible to estimate the residual variance applying to a single-plot observation (σ^2) at $J = 1$.

As J increases the power of statistical tests, for instance tests of hypotheses with regard to equivalence of candidates, increases.

The ratio given by Equation (16.2) is now considered with regard to the elements h^2, σ and i, respectively.

1. Equation (11.34) presents the ratio of the heritability when evaluating each candidate on the basis of J plots to the heritability applying to non-replicated testing, namely

$$\frac{{h_J}^2}{h^2} = \frac{J}{1 + h^2(J-1)} \tag{16.5}$$

2. Equation (11.33) presents the ratio of the phenotypic variance of candidate means across J plots to their phenotypic variance at non-replicated testing, namely

$$\frac{{\sigma_J}^2}{\sigma^2} = \frac{1 + h^2(J-1)}{J} \tag{16.6}$$

where

$${\sigma_J}^2 = {\sigma_g}^2 + \frac{\sigma^2}{J}$$

3. According to Equation (16.1), a higher number of replicates, *i.e.* a higher value for J, implies – at a fixed number of plots (N) – a lower number of candidates. At a fixed number of selected candidates (n) the latter implies (Equation (16.3)) an increase of the portion of selected candidates, *i.e.* a reduction of the selection intensity (i).

When selecting the portion v from a population with a normal distribution for the considered trait, the appropriate standardized minimum phenotypic value ($z_{\min}$, Section 11.1) can be read from tables in statistical handbooks (for example Pearson and Hartley, 1970, Table 16.1). According to Expression (11.8) this selection implies a selection intensity equal to:

$$i = \frac{f(z_{\min})}{v} = \frac{1}{v\sqrt{2\pi}} e^{-\frac{1}{2}{z_{\min}}^2} \tag{16.7}$$

Table 16.1 The ratio of the expected response to selection of n candidates when testing each of $C_J(= N/J)$ candidates at J plots to the expected response to selection of n candidates when testing each of N candidates at a single plot. The proportion of selected candidates amounts to $v_J = n/C_J = Jv$ and $v = n/N$, respectively (source: Bos, 1983a)

			h^2				
v	v_J	J	0.1	0.2	0.3	0.4	0.5
0.005	0.01	2	1.244	1.192	1.145	1.103	1.066
0.005	0.015	3	1.380	1.278	1.195	1.127	1.069
0.005	0.02	4	1.468	1.323	1.214	1.128	1.059
0.01	0.02	2	1.223	1.171	1.125	1.084	1.048
0.01	0.03	3	1.344	1.244	1.164	1.097	1.041
0.01	0.04	4	1.415	1.276	1.171	1.088	1.021
0.02	0.04	2	1.200	1.149	1.104	1.064	1.028
0.02	0.06	3	1.296	1.200	1.123	1.059	1.005
0.02	0.08	4	1.347	1.214	1.114	1.035	0.972
0.03	0.06	2	1.180	1.130	1.085	1.046	1.011
0.03	0.09	3	1.258	1.165	1.090	1.028	0.975
0.03	0.12	4	1.289	1.162	1.066	0.991	0.930
0.04	0.08	2	1.163	1.114	1.070	1.031	0.997
0.04	0.12	3	1.224	1.133	1.066	0.999	0.948
0.04	0.16	4	1.240	1.118	1.026	0.953	0.894
0.05	0.10	2	1.146	1.097	1.054	1.016	0.982
0.05	0.15	3	1.192	1.104	1.032	0.973	0.942
0.05	0.20	4	1.191	1.073	0.985	0.915	0.859
0.06	0.12	2	1.132	1.084	1.042	1.004	0.970
0.06	0.18	3	1.162	1.076	1.006	0.949	0.900
0.06	0.24	4	1.145	1.032	0.948	0.880	0.826

Thus

$$\frac{i_J}{i} = \frac{\frac{f(z_{\min,J})}{v_J}}{\frac{f(z_{\min})}{v}} = \frac{f(z_{\min,J})}{Jf(z_{\min})} = \frac{e^{-\frac{1}{2}({z_{\min,J}}^2 - {z_{\min}}^2)}}{J} \tag{16.8}$$

Altogether, the ratio given by Equation (16.2) is equal to:

$$\frac{R_J}{R} = \frac{e^{-\frac{1}{2}({z_{\min,J}}^2 - {z_{\min}}^2)}}{J} \cdot \sqrt{\frac{J}{1 + h^2(J-1)}} \tag{16.9}$$

With regard to the part of the expression under the square root one can easily see that this part is larger, at $h^2 < 1$, as J increases. It is smaller at higher values for h^2.

The parameters $z_{\min}$ and $z_{\min,J}$, and consequently the ratio R_J/R, can be obtained for any value for J at any value for v. Example 16.1 illustrates the calculation of R_J/R. Table 16.1 presents the ratio for a number of values for v, J and h^2.

Example 16.1 For $v = 0.005$ the standardized minimum phenotypic value amounts to $z_{\min} = 2.576$ (Falconer, 1989, Appendix, Table A). For $J = 2$, *i.e.* $v_2 = 0.01$, it amounts to $z_{\min,2} = 2.326$. Thus

$$\frac{i_2}{i} = \frac{e^{-1/2(2.326^2 - 2.576^2)}}{2} = 0.923$$

At $h^2 = 0.10$ and $J = 2$ Equation (16.5) yields

$$\sqrt{\frac{2}{1 + 0.1}} = 1.348$$

Then

$$\frac{R_2}{R} = 0.923 \times 1.348 = 1.244$$

(see also Table 16.1). Apparently duplicated testing yields at $h^2 = 0.10$ a 24.4% higher expected response to selection than selection, on the basis of non-replicated testing, of the same number of entries from twice the number of candidates.

The optimum value for J is found by looking, in Table 16.1, for the highest value for the ratio R_J/R. This is illustrated by Example 16.2.

Example 16.2 In a fictitious breeding programme the heritability of some trait of some population of candidates amounts under non-replicated testing to $h^2 = 0.30$. The number of plots available for testing amounts to $N = 450$. Selection of $n = 18$ candidates implies $v = 0.04$. Table 16.1 presents then relevant data with regard to the table below

N	n	J	C_J	v_J	R_J/R
450	18	1	450	0.04	1.000
450	18	2	225	0.08	1.070
450	18	3	150	0.12	1.066
450	18	4	112	0.16	1.026

The table shows that a higher response to selection may be expected if less than 450 candidates are tested: selection of 18 candidates when evaluating each of 225 candidates by means of two plots is expected to yield a 7% higher response. When testing each of 150 candidates in three plots, or each of only 112 candidates in four plots, selection of the best 18 candidates is expected to yield a 6.6% and a 2.6% higher response, respectively. Thus at $N = 450, h^2 = 0.3$ and $n = 18$ the optimal value for J is 2. For $h^2 = 0.2$ and $h^2 = 0.1$ one may derive from Table 16.1 that J_{opt} amounts to 3 and 4 (or perhaps even more!), respectively. At $h^2 = 0.5$ the table presents $J_{\text{opt}} = 1$.

The part of the potential response which is not realized when testing with a lower or with a higher value for J than J_{opt} can also be derived from Table 16.1. At $N = 450, n = 18$ and $h^2 = 0.2$ one can determine J_{opt} to be 3. The non-exploited part of the potential selection response due to testing with $J = 2$ amounts then to

$$100\left(1 - \frac{1.114}{1.133}\right) = 1.7\%$$

Comparison of R_4, due to selecting the portion $v_4 = 0.02$, and R_2, due to selecting the same portion ($v_2 = 0.02$) from twice the number of candidates, shows that $J = 4$ is expected to yield a higher response to selection than $J = 2$ (at twice the number of candidates) for each value considered for h^2; see lines 3 and 4 in Table 16.1.

When selecting 8% of the candidates, testing with four replications is to be preferred over testing with two replications for each considered value for h^2 up to 0.4. For $h^2 = 0.5$ the optimum value for J is 1.

16.2 The Shape, Positioning and Size of the Test Plots

16.2.1 General considerations

If, for each candidate, the amount of seed available suffices to grow N rows of a fixed length, one may allocate the seed in different ways. The amount of seed allows to grow J plots each consisting of K rows, on the condition that $JK = N$. The one extreme is testing each candidate by means of a single, *i.e.* a non-replicated, N-row plot. The other extreme consists of testing each candidate by means of N single-row plots. The determination of the optimum plot size in an economic sense (K_{opt}) is considered in Section 16.2.4. It requires a yardstick for measuring the trend in soil fertility, *i.e.* a part of the soil heterogeneity. This subject is considered in Section 16.2.3.

Ideal trial fields, in the open or in glasshouses, provide uniform growing conditions across their whole area. Such trial fields do not exist if their area is 'somewhat large'. In the latter situation the trial field will contain better and poorer sections. These sections may change in time; their contours may depend on the (previous) crop. When evaluating candidates the breeder should try to make allowance for this source of variation. The latter is only possible in as far as the quality of the growing conditions varies from plant-to-plant or from plot-to-plot according to a known pattern (Chapter 14). Random plant-to-plant or plot-to-plot variation in the quality of the growing conditions causes the estimates of the genotypic values of the candidates to be biased.

Additionally, evaluation of candidates by means of **small plots** tends to yield a biased assessment of their monoculture performances. Example 16.3, however, illustrates that this is not always the case.

Example 16.3 Caligari, Brown and Manhood (1985) studied the plot size and the number of replications to be used in potato breeding. Plot size was determined by the number of rows of a given length, *i.e.* six plant positions alongside 270 cm, per plot. The intra-row distance was 75 cm. The basic plot size was a single row. Six clones and four plot sizes, *viz.* one, two, four and eight rows, were examined. Each clone was represented by two single-row plots, two two-row plots, one four-row plot and one eight-row plot in each of two blocks. No evidence of any effect of plot size (or its interaction with genotype) was found when using mean yield per row as the observation characterizing a plot. There was no evidence of any difference between the yields obtained from outer or inner rows; nor did competition from adjacent rows appear to have any effect.

With regard to the single-row and the two-row plots the effect of increasing the number of replicates, while the number of rows is fixed, was also studied. The residual standard deviation decreased with increasing replication according to what should be expected. It was concluded that the most efficient procedure for potato yield trials consists of using single-row plots with as many replicates grown as can be handled.

These results cannot be taken as providing a definitive answer for all such trials. They do, however, show that the generally accepted opinion of 'the larger the plot the better', because 'bigger plots reflect more accurately agricultural conditions', is not necessarily correct.

Example 16.3 refers to the mathematical fact that the environmental variance of a candidate, *i.e.* σ^2/J, is at its smallest when evaluating single plant plots with a very high value for J. However, this ignores possible agricultural or biological differences that are introduced by growing single plants as opposed to plots which contain more than one plant (Caligari, Brown and Manhood, 1985).

Notwithstanding Example 16.3, there is often a bias in the estimate of the monoculture performance of a candidate when estimating this performance by means of small plots. It may be caused by effects of genotype × density interaction due to alleys and to effects of interplot competition (Chapter 15). It is smaller as larger test plots are used: a candidate shows its monoculture performance more precisely under a larger test plot than in a smaller test plot. Section 16.2 considers shape, positioning and size of the plots. This topic is also covered by LeClerg et al. (1962, pp. 111–126) and by Gomez and Gomez (1976, pp. 203–222).

Example 16.4 shows how one may determine empirically the quality of the evaluation of the monoculture yields of candidates by means of small plots.

Example 16.4 Kramer, van Ooijen and Spitters (1982) studied the effect of plot size on the quality of the evaluation of the monoculture performance of 16 homozygous spring wheat lines. In order to determine the latter, the lines were cultivated in six-row plots. The length of the rows was 6 m, the inter-row distance was 0.25 m. Yield data obtained from four complete blocks was taken to represent monoculture performance, *i.e.* the performance at commercial cultivation. They were used to measure the quality of the evaluation of the 16 lines by means of small plots.

Four complete blocks were used to study four types of small plots, namely

		Number of rows	Inter-row distance (m)	Yield evaluated on the basis of*
Plot type:	1	1	0.208	C, D
	2	1	0.416	C, D
	3	3	0.208	A, B, C, D
	4	6	0.208	A, B, C, D

*A: all rows; B: the central row(s), *i.e.* one or four rows; C: 1.5 m of the row length; D: the whole length of the row(s)

The length of the rows was 2 m. The number of kernels sown per m^2 was 250, except for plot type 2, where it amounted to 125.

The quality of the evaluation of the monoculture performance of the lines, by means of a particular type of small plot, was measured by the coefficient of phenotypic correlation (r), estimated across the 16 lines, between small plot yield and monoculture yield at 'commercial cultivation'. The highest coefficient of correlation for each of a number of different evaluation procedures amounted to:

Plot type	Yield of area	Yield adjusted for	r
1	D	number of plants	0.5
2	C	number of plants	0.65
3	B		0.75
3	A	number of plants	0.78
4	A		0.88
4	B	interplot competition	0.92

For plot types 3 and 4 it hardly mattered whether all rows were harvested or only the central row(s). Certainly the rows should be harvested across their whole length.

In some situations breeders decide to use large (or wide) plots. This decision may be made because of

1. The available equipment,
2. The wish to reduce the effect of interplot competition, and/or
3. Interest in traits that are best expressed in large plots, *e.g.* lodging resistance.

The decision may imply application of a low number of replications. This is undesirable in the presence of soil heterogeneity (Section 16.2.3).

In practice breeders often use small plots. They may do so because of the following reasons:

1. Per candidate only a small amount of seed is available
2. They wish to evaluate a large number of candidates when having available limited resources

A small amount of seed per candidate

In the case of selection among candidates obtained from single plants the amount of tubers or seed per candidate tends to be limited. For crops such as peas (*Pisum sativum* L.) or field beans (*Vicia faba* L.) individual plants produce a small number of seeds. F_3 lines will then consist of a small number of plants and evaluation of F_3 lines by means of large plots is then prohibited. The F_3 lines are then to be tested in small plots or one may decide to evaluate F_2-derived F_4 lines by means of large, possibly replicated plots.

Evaluation of a large number of candidates

Breeders tend to prefer evaluation of a large number of candidates by means of small plots over evaluation of a smaller number of candidates by means of larger plots or by means of replicated testing. The theory developed in Section 16.1 allows a check of whether evaluation of as large a number of candidates as possible (namely N), considering the number of plots the breeder is allowed to plant (N), is to be recommended.

16.2.2 Shape and Positioning of the Plots

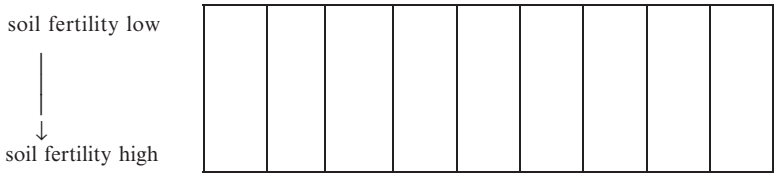
When dividing the trial field into blocks, a part of the residual sum of squares showing up when applying a completely randomized experiment can be shown to be due to differences between the blocks. This may endow statistical tests involving the randomized block experiment with a larger power: more pairs of candidates can be shown to consist of non-equivalent candidates.

The partitioning of the trial field into blocks should be done such that as much as possible of the residual sum of squares, which would show up when applying a completely randomized experiment instead of randomized block design, is assigned to the blocks.

If one does not know the trend in the quality of the growing conditions, it may be best to use square blocks containing square plots. An advantage of square plots is their minimum circumference at a given size. This minimizes

effects of allocompetition. (It is, regrettably, admitted that breeders commonly decide to apply oblong plots because of the equipment available for drilling, maintenance and harvesting).

If the trend in the quality of the growing conditions is known, the blocks should be oblong, with their longitudinal axis parallel to the soil fertility contour lines. This ensures a minimum residual (within blocks) sum of squares and a maximum between blocks sum of squares. The power of statistical tests is further promoted by adapting the positioning of the plots to the direction of the trend in the quality of the growing conditions. Oblong plots should thus be positioned with their longitudinal axis perpendicular to the 'soil fertility contour line':



Beside statistical arguments, economic considerations with regard to size, shape and orientation of plots also play a role. The plot size is equal to the product of its length and its width. For economic reasons the width of the plots is often determined by the available equipment for drilling, maintenance and harvesting. Then the length of the plots is decisive for its size and its shape. Size and shape of the test plots are then directly related.

The size (and shape) to be recommended for test plots have not only been studied generally (Hatheway, 1958), but also for

- Special crops, *e.g.* for tobacco (Crews, Jones and Mason, 1963) or sugar beet (Jaggard, 1975)
- Groups of crops, *e.g.* horticultural crops (Ferguson, 1962).

16.2.3 Yardsticks to Measure Soil Heterogeneity

Soil heterogeneity, *i.e.* variation in the quality of the growing conditions, is often studied by means of a so-called **uniformity trial**. Then all plots contain the same genetically uniform plant material. Such studies have shown that the growing conditions provided by a particular field may appear homogeneous when observed in some season and for some trait of some crop, but they may appear heterogeneous when observed in a different season or for some trait of a different crop. For a given crop, different traits may differ with regard to their capacity to bring soil heterogeneity to light. Flower colour of tulip

plants respond less to the quality of the growing conditions than the size of the tulip plants. The reader is thus reminded to the fact that the environmental variance depends on the genotype considered and, within a given genotype, on the trait considered (Sections 8.1, 8.2). This implies that measurement of the soil heterogeneity is not a straightforward activity.

The subject of measuring soil heterogeneity has been considered by LeClerg, Leonard and Clark (1962, pp. 105–107). Here the theory developed by Fairfield-Smith (1938) for a uniformity trial is presented. The basic area unit plays a role in this approach. This basic area unit may be:

- the area occupied by a single plant
- the area occupied by a single row or
- simply $1\,\mathrm{m}^2$.

The plot size can then be expressed as the number of basic area units. A plot size amounting to K means that the plot consists of K subplots, each with an area of one basic area unit.

Let x_i represent the yield of subplot $i (i = 1, \ldots, K)$ within a plot, and let

$$\underline{X} = \sum_{i=1}^{K} \underline{x}_i$$

represent the total yield of the plot. Because a uniformity trial is studied one may assume that the same error variance, $\mathrm{var}(\underline{x})$, applies to different subplots.

It is quite common to express the yield of a plot in $\mathrm{g/m}^2$, or even in t/ha. Thus one may calculate for each plot the quantity

$$\underline{y}_K = \frac{\underline{X}}{K}$$

where $\underline{y}_K$ represents the plot yield per unit area when dealing with plots of size K. Then

$$var(\underline{y}_K) = var\left(\frac{\underline{X}}{K}\right) = \frac{var(\underline{X})}{K^2} \tag{16.10}$$

The relation between $\mathrm{var}(\underline{y}_K)$ and K, both in the absence and in the presence of a trend, is now considered.

1. In the absence of a trend, the yields of subplots belonging to the same plot are stochastically independent, *i.e.* $\mathrm{cov}(\underline{x}_i, \underline{x}_{i'}) = 0$. This implies that

$$var(\underline{y}_K) = \frac{var(\underline{x}_1 + \ldots + \underline{x}_K)}{K^2} = \frac{K var(\underline{x})}{K^2} = \frac{var(\underline{x})}{K} = \frac{var(\underline{y}_1)}{K} \tag{16.11}$$

2. In the presence of a trend, *i.e.* $\mathrm{cov}(\underline{x}_i, \underline{x}_{i'}) > 0$, implies

$$var(\underline{y}_K) = \frac{K' var(\underline{X})}{K^2} \tag{16.12}$$

where, due to the positive covariances of the observations obtained from subplots belonging to the same plot,

$$K' > K$$

A general expression for the relation between $\text{var}(\underline{y}_K)$ and K, is thus

$$var(\underline{y}_K) = \frac{var(\underline{x})}{K\left(\frac{K}{K'}\right)} = \frac{var(\underline{y}_1)}{K^b} \tag{16.13}$$

where $b = 1$ (absence of a trend) or $b < 1$ (presence of a trend). The parameter b is the so-called **soil heterogeneity index**. Example 16.5 assists in understanding parameter b.

Example 16.5 At $K = 4$ and $K' = 4.44$, *i.e.* at $\frac{K}{K'} = 0.9$, the denominator $K\left(\frac{K}{K'}\right)$ amounts to $4 \times 0.9 = 3.6036$. The appropriate value parameter b is then calculated from $b\log(4) = \log(3.6036)$. This yields $b = 0.9247$.

Equation (16.13) implies

$$\log(var(\underline{y}_K)) = \log(var(\underline{y}_1)) - b\log(K) \tag{16.14}$$

If $\text{var}(\underline{y}_K)$ is estimated for different plot sizes, *i.e.* for different values for K, one may estimate b by means of linear regression of $\log(\text{var}(\underline{y}_K))$ on $\log(K)$. This is illustrated by Example 16.6.

Example 16.6 Gomez and Gomez (1976; p. 208) analysed a uniformity trial with rice. Grain yield (in g/m^2) was determined for basic units of 1×1 m^2. The following scheme presents a part of the data:

842	844	808	822	979	954
803	841	870	970	943	914
773	782	860	822	932	971
912	887	815	937	844	661
874	792	803	793	818	799
908	875	899	788	867	790

Different vertical or horizontal combinations of the $N = 36$ basic area units yield different plot sizes (K; in m^2) and accordingly different numbers of plots ($J = N/K$). For each possible value for K first the grain yield $\underline{y}_K$ (in g/m^2) was calculated for each of the J plots and next $\text{var}(\underline{y}_K)$ was estimated. This yielded the following values:

Plot size (K)	Number of plots (J)	$\hat{\text{var}}(\underline{y}_K)$
$1 \times 1 = 1$	36	4,818
$2 \times 1 = 2$	18	2,028
$1 \times 2 = 2$	18	3,611
$3 \times 1 = 3$	12	3,179
$1 \times 3 = 3$	12	2,418
$2 \times 2 = 4$	9	1,509
$6 \times 1 = 6$	6	464
$1 \times 6 = 6$	6	707
$2 \times 3 = 6$	6	1,704
$3 \times 2 = 6$	6	2,909
$3 \times 3 = 9$	4	2,518

For the four plots of $9\,\text{m}^2$ one gets, for instance:

824.78	923
862.78	810.78

Then one can calculate: $\hat{\text{var}}(\underline{y}_9) = 2518$.

Linear regression of ${}^{10}\log(\hat{\text{var}}(\underline{y}_K))$ on ${}^{10}\log(K)$ yielded

$$ {}^{10}\hat{\text{log}}(\hat{\text{var}}(\underline{y}_K)) = 3.634 - 0.6025 \times {}^{10}\log(K) $$

The estimate of the soil heterogeneity index amounted thus to 0.6025.

Uniformity trials have rarely been carried out over several years at the same site. Data on the stability of the soil heterogeneity index are thus scarce. Koch and Rigney (1951) reported for three years: 0.65, 0.42 and 0.67 for one experiment; and 0.56, 0.68 and 0.76 for another experiment. Apparently the parameter b varied both across sites and across years within site. The crop and the observed traits will also have played a role.

For a trial field with a very strong trend in the quality of the growing conditions the covariance of the performances observed for adjacent basic area units, *i.e.* $\text{cov}(\underline{x}_i, \underline{x}_{i'})$, will be high. The value of b will then be much smaller than 1, it might possibly be a value close to 0. An increase of the plot size, *i.e.* an increase of $\log(K)$, yields then hardly a decrease of $\text{var}(\underline{y}_K)$ (Equation (16.14)). Then $\text{var}(\underline{y}_K) \approx \text{var}(\underline{y}_1)$. In this situation it is advisable to apply small plots and a high number of replications, or to use a different trial field with a weaker trend in the quality of the growing conditions. In such a trial field an increase of K leads to a decrease of $\text{var}(\underline{y}_K)$. At $b = 1$, *i.e.* in the absence of a trend in the quality of the growing conditions, Equation (16.11) applies. In that case $\text{var}(\underline{y}_K)$ is equal to the variance of the average across K separate plots each of size 1.

Occurrence of a soil heterogeneity index larger than 1 may seems improbable. It requires the covariance between the performances observed for adjacent plots in a uniformity trial to be negative. This indicates isocompetition between the plots. Example 16.7 describes a situation where this occurred.

Example 16.7 Spitters (1979, p. 235) found that in his experiments with barley that the grain yield of subplots (in this case single-row plots) varied with a trend. Rows 1 and 6 of a sowing pass yielded systematically better than the intermediate rows obtained from the 6 six-row sowing drill. It was guessed that this phenomenon was caused by a wider distance between rows due to different sowing passes than between rows due to the same sowing pass. It might also be explained from better growing conditions in rows 1 and 6 due to soil compaction in the wheel track of the drill.

Absence of a trend in the quality of the growing conditions does not necessarily imply absence of heterogeneity. With irregular plot-to-plot variation in the quality of the growing conditions the covariance of the performances observed for adjacent basic area units might be near to zero. This implies a soil heterogeneity index of near to one. One should thus not rush to the conclusion that the trial field provides uniform growing conditions. This may be one of the reasons why b is seldom determined in plant breeding practice. But undoubtedly the effort required to perform and to analyse a uniformity trial is the main reason.

Estimation of the soil heterogeneity index on the basis of an experimental design dedicated to the evaluation of candidates is therefore an attractive alternative. Gomez and Gomez (1976, pp. 214–218) illustrated the estimation of b from a split-plot design. Lin and Binns (1984) estimated b from a randomized complete block design. They showed how the accuracy of the evaluation can be improved by increasing K, when keeping J constant, or by the use of incomplete blocks.

Notwithstanding the ambiguity in the interpretation of the soil heterogeneity index, the index in introduced because of its role in the determination of the optimum plot size from an economic point of view (Section 16.2.4).

Sometimes the measurement of a trend in the quality of the growing conditions has been attempted by calculating, for a uniformity trial, the coefficient of correlation between the yields obtained from pairs of adjacent plots (Example 16.8).

Example 16.8 Pijper (1981) estimated for a uniformity trial using the barley variety Varunda the coefficient of correlation between the biomass data of pairs of adjacent rows. For 1976 the mean of six estimates of the coefficient of correlation, each based on 90 pairs of adjacent rows, amounted to 0.2. For 1977 the mean of four estimates, each based on 72 pairs of adjacent rows, amounted to -0.09.

Example 16.7 showed that isocompetition between single-row plots may occur. To eliminate this effect, the coefficient of correlation between pairs of rows separated by a single row was also estimated. This yielded, for 1976, $r = 0.22$ and for 1977 $r = 0.01$. Thus in 1977 the coefficient of correlation between pairs of adjacent rows was reduced by isocompetition.

Pijper also considered two-row plots. Then the mean coefficient of correlation between pairs of adjacent plots amounted, for 1976, to 0.51 and for 1977 to 0.28. This shows that the biomass of two-row plots is less affected by isocompetition and random variation than the biomass of single-row plots.

In all cases the 1976 estimate was higher than the 1977 estimate.

However, the interpretation of this yardstick is not unambiguous:

- A high coefficient of correlation implies that the soil fertilities of adjacent plots are similar. Simultaneously the range, across a large distance, in the quality of the growing conditions may be wide.
- A low coefficient of correlation implies that the qualities of the growing conditions of adjacent plots tend to be different, but simultaneously the range in the quality of the growing conditions across a large distance may be narrow.

It is concluded that the coefficient of correlation is a poor yardstick to measure soil heterogeneity.

16.2.4 The Optimum Plot Size from an Economic Point of View

From an economic point of view, the optimum plot size (K_{opt}) at a given soil heterogeneity index b, is the plot size such, that the total of the costs of the evaluation of a candidate is minimal per unit information. The determination of K_{opt}, at a given value for b, is elaborated in the present section.

In this book the residual variance of single-plot yield data, say $\underline{X}$ (in g or in kg), was designated by σ^2. The residual variance of the mean value across J plots of a candidate amounts then to σ^2/J (Section 11.2.1). With these representations the area of the plot can be anything (Section 16.2.3). When expressing yield in g/m^2 or in t/ha, the yield was represented by $\underline{y}_K$, with residual single-plot variance represented by $\text{var}(\underline{y}_K)$ and with residual variance of the mean across J plots equal to $\text{var}(\underline{y}_K)/J$. All these variances are scale-dependent. For this reason the scale independent quantity $vc = \sigma/\mu$ is a better indicator of the accuracy of the observations.

The reciprocal of the residual variance of the mean across J plots, *i.e.* $J/\text{var}(\underline{y}_K)$, is called the **precision** or the **information** (Steel and Torrie, 1980, p. 123) of the evaluation. It is a meaningful yardstick for the accuracy of

the observations: the larger this quantity, *i.e.* the smaller the residual variance of the mean, the larger the accuracy of the evaluation. An increase of the information is realized by:

1. Increasing J, *i.e.* the number of plots per candidate. An increase of J implies an increase of the cost of the evaluation proportional to the relative increase of the number of replications. The parameter C_1 is introduced to designate the fixed cost of an additional plot of a given candidate.
2. Increasing K, *i.e.* the size, in standard area units, of a plot. An increase of K may tend to reduce). This type of increase involves costs proportional to the plot size. The parameter C_2 designates for a given plot the cost of increasing the size of the plot with an additional standard area unit.

The total costs of the evaluation of a candidate by means of J plots, each of size K, is then

$$J(C_1 + KC_2) \tag{16.15}$$

They amount per unit of information to

$$\frac{J(C_1 + KC_2)}{\left(\frac{J}{var(\underline{y}_K)}\right)} = var(\underline{y}_K) \cdot (C_1 + KC_2) \tag{16.16}$$

The minimum value of the costs of evaluation of a candidate per unit of information occurs if the derivative of this function to K is zero.

Substitution of Equation (16.13) transforms Equation (16.16) into

$$\frac{var(\underline{y}_1) \cdot (C_1 + KC_2)}{K^b} \tag{16.17}$$

Differentiation to K yields:

$$var(\underline{y}_1)\left[-bK^{-b-1}C_1 + (1-b)K^{-b}C_2\right] = \frac{var(\underline{y}_1)}{K^b}\left[\frac{-bC_1}{K} + (1-b)C_2\right] \tag{16.18}$$

- At $b = 1$, *i.e.* in the absence of a trend in the quality of the growing conditions, the plot size is optimal if

$$\frac{C_1}{K} = 0$$

  This occurs approximately at $K \to \infty$.
- At $b < 1$ the derivative is zero for

$$K_{\text{opt}} = \frac{b}{1-b} \cdot \frac{C_1}{C_2} \tag{16.19}$$

At C_1 low and C_2 high, the optimum plot size is smaller than for a high ratio of C_1 to C_2. Swallow and Wehner (1986), dealing with the testing of cucumber, present an application of the present theory.

Chapter 17
Causes of the Low Efficiency of Selection

The effect of improvements in agricultural techniques and of the introduction of higher-yielding varieties has often been studied. The results may seem unimpressive compared to the progress due to scientific advances with other fields of application. (However one should not overlook the fact that such, indeed, low annual progresses have been maintained for decades and decades).

This chapter summarizes causes for the low efficiency of selection in a looking back manner.

Sneep and Hendriksen (1979, p. 421) summarized some data concerning increases of yield attained in agriculture. They wrote: 'In the USA no improvement whatsoever has been obtained for beans and only a moderate improvement for soya beans but the results for maize and groundnuts are very good. In The Netherlands the increase over the high yield levels already attained in 1930 is partly due to improvement in cultivation techniques and partly to the introduction of higher yielding varieties. By considering the two factors separately it is possible to show that the yield increase due to the introduction of better varieties since 1980 was 0.8% per year for potatoes and 0.4% per year for sugar yield of sugar beets. For spring wheat the yield increase due to breeding has been shown to be 1% per year over the same period.... Similar calculations were made in England.... The increase in yield due to breeding was 1.8% per year over a period of 30 years in wheat, accounting for 67% of the total yield increase. For barley the corresponding figures are 1% per year and 50% of the total increase'.

In Section 8.1 it was already indicated that selection tends to be an inefficient process. This was attributed to

1. Non-identical reproduction
2. Variation in the quality of the growing conditions
3. Competition

Non-identical reproduction as a cause for inefficient selection

In the case of identical reproduction the genotypic composition of a population does not change from generation to generation. This occurs at asexual reproduction of clones, at selfing of pure lines, and at re-production (by making the underlying crosses again) of single-cross hybrids.

Non-identical reproduction of the selected candidates is a genetic cause for a disappointing response to selection. The genotypes of the selected candidates are not identically reproduced and do, consequently, not reoccur unaltered in

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 421–428.

the next generation. This occurs when selecting among heterozygous candidates representing a self-fertilizing crop and when selecting among candidates representing a cross-fertilizing crop.

This cause of inefficiency of selection does not occur when selecting among clones, among completely homozygous candidates of a self-fertilizing crop or among test hybrids when developing a single cross hybrid.

Variation in the quality of the growing conditions as a cause for inefficient selection

In order to rank candidates according to their genetic quality, the breeder should make effort to compare the candidates in identical growing conditions. To pursue this, the breeder might compare the candidates within blocks. Additional tools are randomization and replication.

Uniformity of the growing conditions within blocks is an idealization. This implies that ranking of candidates according to their genotypic value is impossible. It is, therefore, unavoidable that candidates with a superior genotypic value are overlooked, whereas candidates with a genotypic value not justifying selection are selected. This applies even more strongly to evaluation procedures employing incomplete block designs and/or non-replicated evaluation procedures using standard plots (Section 14.3.2) or moving means (Section 14.3.3).

Competition as a cause for inefficient selection

Variation among candidates in allocompetitive ability reduces the efficiency of selection of candidates with a genetically superior monoculture performance. Candidates with a strong competitive ability are apt to be selected. They may perform disappointingly when grown in monoculture, *i.e.* in the absence of variation in competitive ability (Chapter 15).

Rasmusson (1987) mentioned four additional causes for an impeded improvement of plant material with regard to traits with quantitative variation. They apply because of interrelationships of the traits.

1. The requirement of a harmonically tuned sizes of different organs, *e.g.* ear size and leaf area. Selection efforts not in consonance with this requirement are expected to be less successful than efforts respecting such harmony.
2. Due to a physiological limit there is mutual compensation between or within organs with regard to their number or size. Thus an increase of the number of ears per rye plant is associated with
   - a decrease of single kernel weight and/or
   - a decrease of the ear size, *i.e.* the number of grains per ear.
3. Pleiotropy (Section 12.2): some genotype with regard to a pleiotropic locus may be favourable with regard to one trait but unfavourable with regard to another trait.

4. The genotype giving rise to a favourable expression of the studied trait may fail to yield an attractive expression in a different genetic background.

The response to selection is mostly low, certainly if each candidate is evaluated by means of only one small plot. Looking back to topics elaborated in preceding chapters the following general causes for a low efficiency of selection, additional to those presented above, can be summarized:

1. In the early phases of a breeding programme breeders usually want to test large numbers of candidates. Because of this, and possibly also because at that stage there is usually only a small amount of plant material per candidate available, the evaluation is done in a single macro-environment. In the presence of genotype × environment interaction one may then tend to select candidates that do not perform in a superior way when grown in another macro-environment. A higher response to selection may be expected if the selection is based on tests covering at least two growing seasons at two or more locations. This is illustrated in Example 18.3.
2. Evaluation of candidates by means of small plots implies not only an inaccurate but, due to allocompetition, also a biased estimation of the monoculture performances of the candidates. Selection based on mixture performance tends thus to yield a disappointingly low correlated response for monoculture performance. Examples 15.7 and 15.8 show that this reason for inefficient selection can be quit important.
3. Usually selection occurs under growing conditions deviating from the conditions provided in commercial farming. Certainly the candidates are subjected to allocompetition, whereas the breeders pursues selection of candidates with a good performance in the absence of such competition. This means that it is unavoidable that the breeder applies indirect selection, which tends to be less efficient than direct selection (see Sections 12.3 and 12.3.1). It implies reduction of the potential response to selection (see Example 15.7). Section 12.3.3 considered the relative loss in potential selection response due to indirect, instead of direct selection.

For cross-fertilizing crops one may add the following cause. When selecting after pollen distribution, the next generation (G_{t+1}) will contain, compared to later generations (G_{t+k}, where $k \geq 2$), an excess of heterozygous plants (Section 2.2.1). If dominance plays a role in the inheritance of the considered trait the lasting response to selection cannot be measured by comparison of the performances of generations t and $t+1$. Such a comparison would yield a biased estimation of the realized response to selection because generation $t+1$ will contain an excess of heterozygous plants. The lasting response to selection should, therefore, be measured as

$$R = E\underline{\mathcal{G}}_{t+2} - E\underline{\mathcal{G}}_{t},$$

i.e. from comparison of the performances of generations $t+2$ (if obtained in the absence of selection in generation $t+1$) and t.

For self-fertilizing crops one may also consider a specific cause for inefficiency of selection. Selection, among heterogeneous and heterozygous plant material, of candidates producing – in later generations – pure lines with superior monoculture performance, may yield disappointing results. This could (partly) be due to the following. In the case of dominance (and when selecting in an 'early' segregating generation), especially plants or lines with a higher than average level of heterozygosity may tend to be selected. The plants descending from the selected candidates will be more homozygous than the selected candidates. There performance may, consequently, differ from the performance of their parents. This is a special drawback of selecting in early generations. Selection in an advanced generation does not suffer from this drawback and may, therefore, be expected to be more efficient. This is illustrated in the next chapter (Example 18.1).

The next section focuses on the fact that the probability of coincidence of the candidate with the highest phenotypic value, *i.e.* the candidate favoured by the breeder, with the candidate with the highest genotypic value, may be (considerably) less than 1. The selection performed by the breeder is then incorrect. The section shows once more that the candidate(s) favoured by a breeder are not necessarily the candidates with the most attractive genotypic values.

17.1 Correct Selection

When comparing K candidates usually the null hypothesis H_0: 'the genotypic values of the K candidates are equal' is tested. Example 17.1 illustrates the situation facing a breeder in the case of equivalence of the candidates.

Example 17.1 The expected value of a single draw from a standard normal distribution is 0. If one takes two draws the expected value of the largest of the two is 0.564. The expected value of the largest draw in a sample of 10 is 1.539; in a sample of 100 it is 2.51 (Pearson and Hartley, 1972, Table 9).

Consider now a yield trial involving 100 candidates. Assume that for each candidate the genotypic value for yield amounts to 5000 kg/ha and that the standard deviation of the mean yield across the B plots, which are distributed over B blocks, amounts for each candidate to 200 kg/ha. Then the highest phenotypic value expected for yield will amount to

$$5000 + (1.539 \times 200) = 5502 \text{ kg/ha}$$

and the lowest expected phenotypic value to

$$5000 - (1.539 \times 200) = 4498 \text{ kg/ha}.$$

The above null hypothesis is usually tested against the alternative hypothesis H_a: 'the genotypic values of the K candidates are not equal'. Such a test hardly makes sense.

If the candidates are truly different (and for what other reason would one want to compare them), it is reasonable to expect differences, however small these may be. Acceptance of H_0 indicates that the power of the test was apparently insufficient; *i.e.* the numbers of locations (L), of years (Y), and/or of blocks per test (B) were too small to succeed in rejection of H_0.

What to do in case of rejection of H_0? The main interest of a breeder is not in acceptance or rejection of H_0 but in the identification of the candidate with the most attractive (expected) genotypic value. The statistical approach to this problem is indicated by the term **statistical selection**. If the main goal of an evaluation trial is indeed identification of the candidate with the most attractive genotypic value, then that goal should be pursued explicitly. Statistical selection procedures aim at goals such as selection of the best candidate, or selection of a subset of candidates, containing the best candidate. These procedures are hardly applied in present plant breeding practice. Practitioners are used to applying analysis of variance techniques. However, statistical selection procedures deserve recognition as useful tools.

The present section introduces the subject of ranking and selection procedures. The concept of correct selection and its probability are introduced by Example 17.2.

Example 17.2 Assume that 11 candidates are tested. Ten of these are equivalent. They have the same (expected) genotypic value, -1, and their phenotypic variance is 1. Candidate 11 has a genotypic value equal to 0 and its phenotypic variance is 1.

Correct selection (CS) is defined as selection of the candidate with the most attractive genotypic value. In the present case $P(\text{CS})$, *i.e.* the probability of correct selection, is the probability that the phenotypic value observed for candidate 11 is larger than the largest phenotypic value observed for the ten other candidates.

It can be derived that $P(\text{CS})$ is only 0.324 at $B = 1$ (Gauch and Zobel, 1989). At random selection, *i.e.* selection in the absence of testing, $P(\text{CS}) = \frac{1}{11} = 0.091$. Gauch and Zobel commented: 'Frequently selection tasks are considerably more difficult than may be recognized'.

Two procedures will be described: the indifference zone approach of selection and the subset selection procedure.

Indifference zone approach of selection

The **indifference zone procedure** (Bechhofer, 1954) proceeds as follows. The genotypic values of the K candidates are $\mathcal{G}_1, \ldots, \mathcal{G}_K$. These candidates are assumed to have the same phenotypic variances, viz. ${\sigma_p}^2$. Correct selection,

i.e. selection of the candidate with the highest genotypic value, is pursued by selecting the candidate with the highest observed mean phenotypic value across the B plots.

The probability of **correct selection**, $P(\text{CS})$, is considered. It depends on ${\sigma_p}^2$, B, K and $\delta = \mathcal{G}_{[K]} - \mathcal{G}_{[K-1]}$, *i.e.* the difference between the highest genotypic value, $\mathcal{G}_{[K]}$, and the highest but one genotypic value, $\mathcal{G}_{[K\text{-}1]}$. One may require that:

$$P(\text{CS}) \geq P^*$$

where

$$\frac{1}{K} < P^* < 1$$

if δ is at least δ^*.

The probability of selecting the candidate with the highest genotypic value is at least P^*, whenever the genotypic value of this candidate is at least δ^* units higher than the genotypic value of the second best candidate.

The minimum for $P(\text{CS})$ is attained for the least favourable situation, *i.e.* for the situation where K-1 candidates have the same genotypic value, whereas one candidate has a genotypic value at least δ^* units higher. (This situation was considered in Example 17.2). Example 17.3 illustrates how one may calculate δ^*.

Example 17.3 Van der Laan and Verdooren (1990) illustrated an application of the indifference zone approach. They used magnesium content of leaf 17 (LMG, in %) data of *tenera* oil palm families. This trait was shown to be clearly related to fresh bunch yield of the families (Example 12.5). The standard deviation of this trait was known to be 0.0186. The data were obtained from an experiment involving $K = 10$ families evaluated by means of $B = 4$ randomized complete blocks containing 16 palms per plot. The average LMG values ($\overline{p}$) of the ten families are summarized below:

Family	$\overline{p}$	Rank number
1	0.212	[5]
2	0.222	[7]
3	0.242	[8]
4	0.204	[3]
5	0.210	[4]
6	0.186	[2]
7	0.218	[6]
8	0.244	[9]
9	0.162	[1]
10	0.248	[10]

The least favourable situation is given by $\mathcal{G}_{[1]} = \ldots = \mathcal{G}_{[9]} = \mathcal{G}_{[10]} - \delta^*$. When choosing the minimum probability of correct selection, P^*, to be equal to 0.90, the quantity δ^* can be calculated to be

$$\delta^* = \frac{\tau \times 0.0186}{\sqrt{B}} = \frac{2.9829 \times 0.0186}{2} = 0.0277$$

where $\tau = 2.9829$ is obtained from Dickinson-Gibbons, Olkin and Sobel, (1977).

Otherwise one can determine the value for B needed for $\delta^* = 0.020$. This value amounts to

$$B = \left(\frac{\tau \times \sigma}{\delta^*}\right)^2 = \left(\frac{2.9829 \times 0.0186}{0.020}\right)^2 = 7.7$$

It appears that this is a value not commonly applied in plant breeding practice.

The minimal probability P^* of correct selection can only be guaranteed if B is high enough. The value for B such that the conditions posed by the choices for P^* and δ^* are met at given values for K and ${\sigma_p}^2$ can be derived (Van der Laan and Verdooren, 1990). It is the number of blocks to be used when applying the indifference zone approach of selection. If ${\sigma_p}^2$ is unknown an initial experiment is required to get the estimate s^2 which is used to determine B.

Subset selection

The **subset selection procedure** (Gupta, 1956) aims to select a subset from the K candidates, including (with a certain probability) the candidate with the most attractive genotypic value. Candidate $k(k = 1, \ldots, K)$ is included in this subset if its mean phenotypic value across the B plots ($\overline{p}_k$) exceeds

$$\max_{1 \le j \le K} (\overline{p}_j) - \frac{\tau \times \sigma}{\sqrt{B}} \tag{17.1}$$

where the numeric value of τ must be determined such that

$$P(\text{CS}) \ge P^*$$

for all possible genotypic values of the K candidates.

For this case it can be shown that the least favourable situation is the situation where all candidates have the same genotypic value. Appropriate values for τ are given by Dickinson-Gibbons, Olkin and Sobel, (1977).

The number of candidates in the subset is random. It depends on the (common) residual variance (σ^2) and on B. Its expected value should, of course, be as small as possible. The size of the subset reflects the confidence in identification of the most attractive candidate. A large subset means that the candidates have similar genotypic values and/or that B is small.

If ${\sigma_p}^2$ is unknown, one may include in the subset candidate k if its mean phenotypic value exceeds

$$\max_{1 \le j \le K} (\overline{p}_j) - \frac{\tau \times s_p}{\sqrt{B}}$$

where ${s_p}^2$ is an unbiased estimate of the residual variance ${\sigma_p}^2$ based on ν degrees of freedom. The parameter τ has been tabulated for different values for K, ν and P^* (Dickinson-Gibbons, Olkin and Sobel, 1977)

Example 17.4 illustrates how the subset selection procedure proceeds.

Example 17.4 When applying the subset selection procedure to the data given by Example 17.3, the breeder will select the families for which $\overline{p}_k$ is larger than

$$0.248 - \frac{\tau \times 0.0186}{\sqrt{4}}$$

As for $P^* = 0.90$ the appropriate value for τ amounts to 2.9829, the breeder will select all candidates with a mean phenotypic value for leaf Magnesium content (in %) larger than

$$0.248 - \frac{2.9829 \times 0.0186}{2} = 0.220$$

The subset of selected candidates will thus consist of candidates 2, 3, 8 and 10.

Chapter 18
The Optimum Generation to Start Selection for Yield of a Self-Fertilizing Crop

There are theoretical reasons as well as reasons of a practical nature to start, in a self-fertilizing crop, selection for yield in an early segregating generation.

However, in advanced generations the coefficient of correlation between related lines belonging to successive generations is, due to the greater number of homozygous loci, stronger. This implies that the selection is more efficient in later generations than in earlier generations. From a genetic point of view it is, therefore, efficient to postpone the starting point of the selection to a later generation.

This chapter considers the dilemma concerning the optimal generation to start selection in a self-fertilizing crop.

18.1 Introduction

For self-fertilizing crops it is difficult to arrive at a decision with regard to the generation in which selection for a trait showing quantitative variation should be started.

Section 18.2 presents several reasons to start the selection in an early segregating generation. These are either of a theoretical or a practical nature. However, the coefficient of correlation with regard to the considered trait between related lines belonging to successive generations is, due to the greater number of homozygous loci, stronger in later generations. This suggests that selection in later generations is more efficient. From a genetic point of view it is, therefore, efficient to postpone the starting point of the selection to a later generation. Section 18.3 considers the pros-and-cons of starting selection for yield in later generations.

Example 18.1 describes an experiment allowing verification of opinions with regard to the question in which generation selection for a trait with quantitative variation should start.

Example 18.1 Whan, Knight and Rathjen (1982) observed grain yield of wheat lines representing the F_3, F_4, F_5 and F_6 populations obtained from two initial crosses. The lines were developed in the absence of selection, *i.e.* by 'random selection'. This allows calculation cf the response to selection that would have been realized when starting the selection in a certain generation. When measuring the response to the selection, the level of heterozygosity did not play a role because the yield of the plant material obtained after

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 429–436.

selection (population P'_{t+1}, Section 11.1) was compared to the yield of the lines retained by random selection (population P_{t+1}). By using remnant seed, all lines of all populations were tested both in the same and in successive seasons and both at the same and at different locations.

The result of the simulated selection was established:

1. In the same growing season and at the same location, *i.e.* at the conditions prevailing at selection;
2. In the same season but at a different location and
3. At the same location but in a different season.

Response to selection in the same season and at the same location

The response to selection among lines derived from plants of the F_2, the F_3 or the F_4 population was favourable. Offspring of F_3 lines in generations F_4, F_5 and F_6 showed progressive inbreeding depression. The response to selection among F_3 lines, when selecting on the basis of the average grain yield of the descending F_4 lines (progeny testing), was higher than the response to selection among F_3 lines, when selecting on the basis of the yield of the F_3-lines *per se*. The latter selection was not very efficient.

Response to selection in the same season but at a different location

When evaluated at a different location, the response to selection was satisfying for only one of the two initial crosses. The response to selection of F_3 lines, when selecting on the basis of progeny testing, did not surpass anymore the response to selection with regard to the yield of the lines *per se*. Line × location interaction was clearly present.

Response to selection at the same location but in a different season

When evaluated in a different year the response to selection was very low. Apparently large effects of line × year interaction applied. The response to line selection in generations F_3, F_4 or F_5 did not show a clear trend.

A conclusion with regard to the optimum generation to start selection could not be drawn. The important effects of line × year and line × location interaction showed, however, that the selection should be based on tests across different macro-environments.

18.2 Reasons to Start Selection for Yield in an Early Generation

Negative effects of line × year interaction on the response to selection with regard to a trait with quantitative variation (Example 18.1) can be reduced by evaluating the candidates for several successive seasons. As the latter tends to

cause a delay in the release of a new variety one may start the selection in an earlier generation. This reason for an early start of selection is of a practical nature. It was suggested by Whan, Knight and Rathjen (1982). (Alternatively one may apply breeding techniques reducing the time required to attain complete homozygosity. Such techniques, *e.g.* single-seed descent (Section 6.1) or doubling the chromosome number of haploid plants (Section 3.1), tend to be applied more and more often).

Theoretical considerations as well, may induce breeders to start selection in an early generation. Here the views developed by Shebeski (1970) are presented. Shebeski tried to explain why a spectacular breakthrough with regard to yield is attained so rarely. In his view it is caused by either too small a population size or by the inability of the breeder to identify plants or lines with a superior genotype. He elaborated the following reasoning.

Yield is a trait affected by many loci. Each chromosome arm may be assumed to contain at least one relevant locus. Then grain yield of bread wheat is controlled by at least 42 more-or-less independently segregating loci (Chapter 1). If two wheat varieties, which have a different homozygous genotype with regard to 25 of such loci, are crossed, the probability that a plant of the F_2 generation has genotype BB or Bb for each of the 25 loci amounts to

$$(\tfrac{3}{4})^{25} = 0.00075$$

Thus it is expected that 1 out of 1329 plants of the F_2 has the complex genotype $B_1 \cdot B_2 \cdot \ldots B_{25}\cdot$, where $\cdot$ indicates the presence – at the considered locus – of either allele b or B. The probability that a plant of the F_2 generation has a different genotype is 0.99925.

The probability that a plant has genotype BB or Bb for each of the 25 loci amounts in the F_3 to

$$(\tfrac{5}{8})^{25} = 0.0000079$$

Then 1 out of 126,765 plants of the F_3 is expected to have genotype $B_1 \cdot B_2 \cdot \ldots B_{25}\cdot$. In the F_4 this expected relative frequency amounts to 1 out of 1,765,781 plants!

Apparently the probability that a plant has a genotype capable of producing in a later generation a plant with the best possible complex genotype $B_1B_1B_2B_2 \ldots B_{25}B_{25}$, is highest in the F_2 generation.

This theoretical consideration suggests to start selection of plants or lines capable of generating the best possible genotype as early as possible, *i.e.* in generation F_2 and F_3, respectively. Example 18.2 illustrates that it is difficult to realize this goal in a practical situation. It shows the inefficiency of mass selection for yield considered in Section 14.2.

Example 18.2 Shebeski (1970) reported that each of four breeders selected, apparently visually, in each of 11 F_2 populations of wheat, each of which consisted of about 10,000 plants, the 10 best plants. The 440 F_3

lines descending from the $4 \times 11 \times 10 = 440$ selected F_2 plants were separately tested adjacent to standard plots containing the offspring of plants selected at random from the corresponding F_2 population. About 50% of the F_3 lines yielded better than the corresponding standard.

Apparently the highest-yielding plants in the F_2 populations were rarely plants with a superior genotypic value. Selection of individual, possibly highly heterozygous, plants of an F_2 population is then in vain.

Shebeski (1970) continued his reasoning as follows. Assume that the breeder succeeds in retaining, possibly together with other plants of the F_2 generation, a plant of the F_2 with the complex genotype $B_1 \cdot B_2 \cdot \ldots B_{25}\cdot$. This plant is expected have genotype BB at $25 \times 1/3 \approx 8$ of the 25 considered loci and genotype Bb at the other 17 loci. The probability that a plant of population F_2 has such a complex genotype is, in fact, equal to

$$\frac{25!}{0!17!8!}\left(\frac{1}{4}\right)^0\left(\frac{2}{4}\right)^{17}\left(\frac{1}{4}\right)^8 = \binom{25}{17}\left(\frac{1}{2}\right)^{17}\left(\frac{1}{4}\right)^8 = 0.0001259$$

This means that one out of 7943 plants may be expected to have this complex genotype. The genotypic composition of the F_3 line descending from it, is equal to the genotypic composition of an F_2 population segregating for 17 unlinked loci. The probability that a plant, belonging to this F_3 line, has genotype $B\cdot$ for each of the 17 loci for which the F_2 parent plant had a heterozygous genotype, amounts to

$$(\tfrac{3}{4})^{17} = 0.0075$$

Thus 1 out of 133 plants of the considered F_3 line is expected to have a genotype capable of producing in a later generation a plant with complex genotype $B_1B_1B_2B_2 \ldots B_{25}B_{25}$.

In the absence of selection among the plants in the F_2 population, it is expected that one out 7,943 F_3 lines will descend from an F_2 plant with the considered complex genotype. When each F_3 line would consist of 133 plants it is, in the absence of selection, expected that one F_4 line will descend from a plant of the F_3 generation with genotype BB for eight loci and genotype $B\cdot$ for the other 17 loci. As the efficiency of mass selection for yield within the F_3 must be expected to be very low, each of the best F_3 lines will give rise to many F_4 lines to be tested.

This reasoning suggests extensive testing, associated with selection, in the early segregating generations. Sneep (1977) supported Shebeski's plea. It is in contrast to the conventional approach in cereal breeding.

A main weakness of conventional breeding is that it does not rectify the error arising from not selecting in the F_2, F_3 or F_4 population at least one plant with genotype $B\cdot$ for all relevant loci. This shortcoming can be adjusted by intercrossing in later generations (Section 9.3).

18.3 Reasons to Start Selection for Yield in an Advanced Generation

Plants belonging to an early segregating generation of a self-fertilizing crop vary with regard to the number of loci, affecting yield, with a heterozygous genotype. If a high number of loci with a heterozygous genotype is associated with a high genotypic value for the complex genotype, and consequently with a high phenotypic value, selection in an early generation implies preferential selection of highly heterozygous plants or lines. The genotypes of the offspring obtained from the selected candidates will be more homozygous than the selected candidates themselves. The performance of the offspring may then be disappointing. The coefficient of correlation, across lines, between the performance of a line and the performance of its offspring is, consequently, expected to be lower in an early segregating generation than in an advanced generation. The efficiency of selection in an advanced generations tends thus to be higher. (The higher efficiency of selection in the case of identical was already emphasized in Section 6.1.)

Additionally, the quantity of seed available per candidate may be larger in the more advanced generations. The amount of seed representing an F_3 line depends on the amount of seed produced by the parental F_2 plant. It is less than the amount of seed representing the corresponding F_2-derived F_4 line. In its turn the latter amount is less then the amount of seed available when evaluating a pure line. Monoculture performance (Section 15.1) is thus better evaluated in a more advanced generation. The genotypic values of the candidates may then be estimated more accurately and with a lower bias than in earlier generations.

The above reasoning suggests postponing extensive testing, associated with selection, to advanced generations. This is supported by Example 18.3, Table 3(a), but the described experiment also teaches, once more, how disturbing effects of genotype × macro-environment interaction may be.

Example 18.3 Whan, Rathjen and Knight (1981) observed grain yield of wheat lines representing the F_3, F_4, F_5 and F_6 populations of two initial crosses. The lines were developed in the absence of selection (see also Example 18.1).

In 1975 F_3 lines and their corresponding F_2-derived F_4 lines were tested in Roseworthy by means of two-row plots, 2 m long. Moving mean adjustment of the data, involving 14 contiguous plots, was applied. In 1976 the same lines as well as F_5 and F_6 lines were tested, both in Roseworthy and in Mortlock, by means of four-row plots, 2.5 m long. The data were adjusted on the basis of the mean across two contiguous standard plots.

Table 18.1(a) shows that the coefficient of correlation, across lines, between the grain yield of a line in generation t (F_t line) and the mean grain yield of the corresponding F_{t-1}-derived F_{t+1} lines was higher as t was higher.

The coefficient of correlation between the grain yield of a line in generation t and the mean grain yield of the corresponding F_{t-1}-derived F_{t+k} lines k generations later was for $k = 1$ (Table 18.1(a)) higher than for $k = 2$ or 3 (Table 18.1(b)). The coefficient of correlation tended to be higher for higher values for t. At $k = 3$ some correlations were not significant.

Table 18.1 Coefficients of correlation, across lines, concerning grain yield of related wheat lines evaluated in different generations. Data obtained in 1976, in Roseworthy and Mortlock (source: Whan, Rathjen and Knight, 1981)

	Generations	Roseworthy	Mortlock	Pooled
(a)	$F_3 - \overline{F}_4$	0.59	0.44	0.51
	$F_4 - \overline{F}_5$	0.62	0.57	0.68
	$F_5 - \overline{F}_6$	0.69	0.75	0.78
(b)	$F_3 - \overline{F}_5$	0.38	0.28	0.39
	$F_4 - \overline{F}_6$	0.34	0.49	0.42
	$F_3 - \overline{F}_6$	0.25	0.28	0.29
(c)	$\overline{F}_4 - \overline{F}_5$	0.61	0.56	0.62
	$\overline{F}_5 - \overline{F}_6$	0.46	0.76	0.63
	$\overline{F}_4 - \overline{F}_6$	0.28	0.53	0.42

(a) The coefficient of correlation between a line and its offspring one generation later.
(b) The coefficient of correlation between a line and its offspring two or three generations later.
(c) The coefficient of correlation between an F_2-derived F_4 or F_5 line and its offspring one or two generations later.

The separate lines were tested by means of non-replicated plots. This was expected to give rise to relatively low coefficients of correlation. Thus the coefficient of correlation between the mean, per F_2 plant, across the F_2-derived F_4 or F_5 lines and the mean across the corresponding F_2-derived F_5 or F_6 lines was also estimated. Table 18.1(c) shows that in this way higher coefficients of correlation were obtained; compare *e.g.* $r_{\overline{F}4,\overline{F}5} = 0.62$ to $r_{F3,\overline{F}5} = 0.39$.

The coefficient of correlation, across lines, between 1976 grain yield in Roseworthy and in Mortlock was rather low (Table 18.2). When considering the same plant material, the correlation was relatively high for F_3-derived F_4 lines and for F_4-derived F_5 lines.

Table 18.2 Coefficients of correlation concerning grain yield of related wheat lines in Roseworthy and Mortlock. Data obtained in 1976 (source: Whan, Rathjen and Knight, 1981)

		Mortlock			
		Same material	$\overline{F}_4$	$\overline{F}_5$	$\overline{F}_6$
Roseworthy	F_3	0.19	0.26	0.21	0.24
	F_4	0.42	0.24	0.34	0.17
	F_5	0.54	0.34	0.45	
	$\overline{F}_4$	0.36	0.29	0.21	
	$\overline{F}_5$	0.28	0.19		

Table 18.3 Coefficients of correlation concerning grain yield, in different years and/or locations, of related wheat lines and their offspring. Data obtained in 1975 in Roseworthy and in 1976 in Roseworthy and Mortlock (source: Whan, Rathjen and Knight, 1981)

		1976					
		Roseworthy			Mortlock		
		$\overline{F}_4$	$\overline{F}_5$	$\overline{F}_6$	$\overline{F}_4$	$\overline{F}_5$	$\overline{F}_6$
Roseworthy, 1975	$\overline{F}_3$	0.12	0.00	0.16	0.18	0.00	0.03
	$\overline{F}_4$	0.12	−0.13	0.23	0.21	0.05	0.11

The coefficient of correlation between the 1975 yield and the 1976 grain yield in Roseworthy or in Mortlock was also estimated. Table 18.3 shows that these correlations were very low. Apparently the effects of line × year interaction were very large. This was already noted in Example 18.1, sub 3.

The coefficient of correlation, across the lines, between the grain yield of an F_3 line, or the mean grain yield of F_2-derived F_4 lines, and the mean grain yield of the corresponding F_2-derived F_5-lines (Table 18.3: $r_{1975R,1976R}$ = 0.00 and −0.13, respectively). This may be due to using in 1975 and 1976 different test and adjustment procedures.

In this experiment the coefficient of correlation between different generations was more strongly reduced by line × year interaction than by line × location interaction (compare Table 18.2 to Table 18.3). It is tentatively concluded that the coefficient of correlation was more strongly decreased by line × year or line × location interaction than by a change in the heterozygosity (compare Tables 18.2 and 18.3 to Table 18.1).

Shebeski (1970) evaluated grain yield of F_3 lines of wheat by means of 750 plants per line. In order to reduce the effect of allocompetition, a large interplot distance (60 cm) was applied and a standard variety was grown adjacent to each line. Shebeski estimated the coefficient of correlation, across the lines, between the grain yield of an F_3 line and the grain yield of the corresponding F_2-derived F_5 line to be as high as 0.85.

Early in this section it was explained why it is difficult in an early segregating generation to identify genotypes capable of producing, in later generations, homozygous lines with superior monoculture performance. Thus it is often decided to delay intense selection for yield until pure lines are available. Pure lines can be developed

1. by conventional inbreeding, mostly continued selfing;
2. by application of the single seed descent method (SSD-method; Section 7.1) or
3. by doubling the number of chromosomes of haploid plants (DH-method; Section 3.1). The DH-method yields pure lines which are the products of recombination during one generation, whereas selfing and the SSD-method yield pure lines which are the product of recombination during several

generations of sexual reproduction. Example 18.4 presents a comparison, on the basis of the performances of the obtained lines, between the three procedures for developing (pure) lines.

Example 18.4 Powell, Caligari and Thomas (1986) compared 92 F_4 lines tracing back to 20 randomly chosen F_2 plants, 54 random F_4-derived F_7 lines (obtained by application of the SSD-method) and 18 F_1-derived DH-lines. The initial cross involved the spring barley varieties Universe and Mazurka. The 164 lines were tested in each of two randomized complete blocks. Each single-row plot consisted of up to 10 seeds, sown at 5 cm spacings, with a wheat guard plant at each end. The rows were spaced 22.5 cm apart. From each plot five randomly chosen plants were observed.

The mean phenotypic values of the three types of lines differed significantly for

- number of grains per ear on the main stem,
- final plant height,
- yield of grain on the main stem and
- thousand grain weight.

For these traits the F_4 lines scored higher than the SSD- or the DH-lines. This could be due to dominance effects in the still heterozygous F_4 lines (such effects would reduce the response to early generation selection). The DH-lines had a lower value for thousand grain weight than the SSD-lines. This could be due to epistasis and, if epistasis is present, to linkage.

The three types of lines did not differ significantly for the estimates of $\text{var}(\underline{\mathcal{G}}_{F\infty})$ for the studied traits.

It was concluded that the choice between the use of F_4 and DH- or SSD-lines should be made on non-genetic considerations, such as the cost of each method. An increase in the frequency of plants with a desirable recombinant genotype, resulting from early generation selection, is the only condition for to favour pedigree selection.

Chapter 19
Experimental Designs for the Evaluation of Candidate Varieties

Selection for a trait with quantitative variation is often based on a comparison of the candidates with regard to the trait. The probability of correct selection is higher as the environmental conditions under which the candidates are compared are more similar. For this reason breeders always seek trial fields providing growing conditions as uniform as possible. When knowing the plant-to-plant or the plot-to-plot trend in the quality of the growing conditions, the breeder may adjust the phenotypic values observed for the candidates according to a procedure dedicated to the elimination of the contribution of the trend to these values (Chapter 14).

Sometimes there is not a gradual trend in the quality of the growing conditions, but a rather sudden, sometimes even clear-cut change. Such a change may be due to physical properties of the field, e.g. the moisture content in the presence of a slope, or it may be due to effects of the cultivation regime in the previous growing season, e.g. cultivation of different crops in different parts of the trial field, or application of a different crop husbandry in different parts of the field, even when growing the same crop, e.g. different dates of harvest. In the situation of a sudden change in the quality of the growing conditions, the breeder may partition the trial field in different parts, usually called blocks (Section 8.1), which are assumed to provide uniform growing conditions.

How should a breeder or how should testing authorities compare candidates, which are evaluated in different blocks, in an unbiased way with regard to their genotypic values? How should one compare candidates without being disturbed by the fact that not all of them have been tested under the same conditions? This problem plays especially a role in the stage of variety testing, when a correct comparison is of utmost importance.

This chapter introduces experimental designs designed to provide unbiased comparisons of candidates, possibly candidate varieties, in the case of replicated testing of the candidates by means of incomplete blocks, i.e. by means of blocks not accommodating all candidates, whereas different blocks provide different growing conditions. Problems concerning the estimation of differences between candidates and testing their significance are not considered. These topics belong to a special branch of statistics, i.e. design and analysis of experiments.

Partitioning of the trial field into blocks, allows partitioning of the residual sum of squares occurring in a **completely randomized experiment** into the between-block sum of squares and the residual within-block sum of squares. If each block contains a plot for each of the t candidates, a so-called **randomized**

I. Bos and P. Caligari, Selection Methods in Plant Breeding – 2nd Edition, 437–443.

complete block design is used. In that case the classification of the data according to the blocks and the classification according to the candidates are **orthogonal**. This allows partitioning of the total sum of squares according to Pythagoras.

When using a randomized complete block design the residual variance of the difference between candidate mean values is the same for all pairs of candidates. This is a property of so-called **balanced** designs. Because the candidates are all tested in the same set of r blocks the comparisons of the candidates are not biased by block effects.

Indeed, comparison of candidates by means of blocks such that each block accommodates all candidates is very attractive. Blocks providing uniform growing conditions, including the harvest conditions, allow then unbiased and accurate comparison of the genotypic values of the candidates.

Often, however, the number of candidates is so large that one may not anymore assume that a block accommodating all candidates provides uniform growing conditions. Larger blocks tend to provide less uniform growing conditions than smaller blocks. The residual (within-block) variance (σ^2) applying to individual plots of such blocks tends to be larger than the residual variance applying to smaller blocks. This means that the residual variance of the difference of the mean phenotypic values of candidates i and j, both evaluated in the same r complete blocks, *viz.*

$$var(\overline{p}_{i\cdot} - \overline{p}_{j\cdot}) = \frac{2\sigma^2}{r} \tag{19.1}$$

tends to be large. The power of the test of H_0: 'The genotypic values $\mathcal{G}_i$ and $\mathcal{G}_j$ of candidates i and j are equivalent' is then small.

In order to have a test with a reasonable power, the number of candidates tested by means of a randomized complete block design should not be too high. This number depends, of course, on the plot size: the larger the plot size, the larger the block at a given number of plots per bock. If the actual value of t, *i.e.* the number of candidates, is 'high', *e.g.* larger than 15, the use of **incomplete** blocks should be considered. In this case each incomplete block accommodates only k of the t candidates ($k < t$). The complete set of all t candidates is then tested by means of a number of incomplete blocks. The total number of these incomplete blocks, say b, exceeds, of course, r, the number of replicates.

The use of incomplete blocks may imply that the contrast of the phenotypic values of candidates i and j, *i.e.*

$$\overline{p}_{i\cdot} - \overline{p}_{j\cdot}$$

is a biased estimator of the difference of the genotypic values of these candidates.

This is illustrated by the following example. Assume that $t = 4$ candidates are tested by means of two incomplete blocks each accommodating $b = 3$ candidates:

		Block	
		1	2
Candidate	1:	x	
	2:	x	x
	3:	x	x
	4:		x

The expected value of the difference of the mean phenotypic values of candidates 2 and 3, *i.e.* is equal to

$$E(\overline{p}_{2\cdot} - \overline{p}_{3\cdot}) = \tfrac{1}{2}E(\underline{p}_{21} + \underline{p}_{22} - \underline{p}_{31} - \underline{p}_{32})$$
$$= \tfrac{1}{2}([\mu + \alpha_2 + \beta_1] + [\mu + \alpha_2 + \beta_2] - [\mu + \alpha_3 + \beta_1] - [\mu + \alpha_3 + \beta_2])$$
$$= \alpha_2 - \alpha_3 = \mathcal{G}_2 - \mathcal{G}_3$$

The difference of the mean phenotypic values is thus an unbiased estimate of the difference of the genotypic values.

The expected value of the difference of the phenotypic values of candidate 1, in block 1, and candidate 4, in block 2, amounts to

$$[\mu + \alpha_1 + \beta_1] - [\mu + \alpha_4 + \beta_2] = [\alpha_1 + \beta_1] - [\alpha_4 + \beta_2]$$

The difference of the phenotypic values of candidates 1 and 4 is thus not an unbiased estimate of the difference of their genotypic values.

An unbiased estimate of the difference $\mathcal{G}_1 - \mathcal{G}_4$ consists of the contrast

$$(\underline{p}_{11} - \underline{p}_{21}) - (\underline{p}_{42} - \underline{p}_{22})$$

This appears as follows

$$E[(\underline{p}_{11} - \underline{p}_{21}) - (\underline{p}_{42} - \underline{p}_{22})]$$
$$= ([\mu + \alpha_1 + \beta_1] - [\mu + \alpha_2 + \beta_1]) - ([\mu + \alpha_4 + \beta_2] - [\mu + \alpha_2 + \beta_2])$$
$$= (\alpha_1 - \alpha_2) - (\alpha_4 - \alpha_2) = \mathcal{G}_1 - \mathcal{G}_4$$

The use of incomplete blocks does, consequently, not exclude unbiased estimation of contrasts of genotypic values, but the accuracy of the unbiased estimates will vary across the estimators. In this case the experimental design is called **unbalanced**. The residual variance of the difference of the phenotypic values of candidates tested within the same block(s) will be lower than the residual variance of the difference of the phenotypic values of candidates tested in different blocks.

It is self-evident that the residual variances of differences of phenotypic values of candidates should be as low and as uniform as possible. Therefore, experimental designs employing incomplete blocks have been developed having the property that the residual variance of the difference of the phenotypic

values is equal to a constant value or, depending on the considered pair of candidates, to one out of a very small number of different values.

The category of experimental designs with one constant value for the residual variance of the difference of the phenotypic values of candidates for all pairs of candidates is indicated as **balanced incomplete block** (BIB) designs. Experimental designs where the residual variance of the difference between the phenotypic values of candidates may adopt one out of only two different values are indicated as **partially balanced incomplete blocks** (PBIB). Depending on the considered pair of candidates, the two candidates occur together in either λ_1 blocks or in λ_2 blocks. Example 19.1 illustrates an incomplete block design where the values $\lambda_1 = 0$ and $\lambda_2 = 1$ apply.

Example 19.1 Cochran and Cox (1957, p. 453) describe the following experimental design:

		Block								
		1	2	3	4	5	6	7	8	9
Candidate	1:	x			x			x		
	2:	x				x			x	
	3:	x					x			x
	4:		x		x					x
	5:		x			x		x		
	6:		x				x		x	
	7:			x	x				x	
	8:			x		x				x
	9:			x			x	x		

The design is specified by the parameters $t = 9, k = 3, r = 3$ and $b = 9$. Candidates 1 and 2 occur only together in block 1 ($\lambda = 1$). Candidates 1 and 3 also occur only once together (also in block 1) and so do candidates 2 and 3. Candidates 1 and 4 occur once together (in block 4). Candidates 6 and 8 do not occur together with candidate 1 in the same block. For these pairs of candidates the value of λ amounts to 0.

The residual variance of the difference between the phenotypic values of two candidates adopts the lower value if λ_2 applies to the pair of candidates and the higher value if λ_1 applies.

The manuals by Cochran and Cox (1957) and Kuehl (2000) are important sources of information about incomplete block designs. They present references to balanced and partially balanced experimental designs for combinations of values for t and k. In order to accommodate a given number of candidates, which are to be evaluated by means of incomplete blocks consisting of a more-or-less predetermined size, adaptation of t and/or k may be required. If several

designs can be used, designs where combinations of incomplete blocks coincide with replicates are to be preferred. The grouping into complete replications is handy in the management of an experiment belonging to this category of so-called **resolvable designs**. It allows management of large trials on a replication-by-replication basis.

A special group of BIBs consists of the so-called **balanced lattice** designs. These are characterized by special values for t, k and r, *viz.*

$$t = 9, 16, 25, 49, 64 \quad \text{or} \quad 81$$

whereas

$$k = \sqrt{t} \quad \text{and} \quad r = k + 1$$

Balanced lattice designs belong to the category of resolvable designs.

One may try to modify the number of candidates into one of the values for t mentioned before by adding or eliminating one or a few candidates. Another way-out is the use of a so-called **rectangular lattice**, where

$$t = k(k + 1) \quad \text{for} \quad k = 3(1)9$$

These designs are not balanced but may be considered as PBIBs for practical purposes. Still another solution is provided by the **cubic lattices**, where

$$t = k^3 \quad \text{for} \quad k = 3(1)10 \quad \text{and} \quad r = 3 \text{ or a multiple of } 3$$

The requirement

$$r = k + 1$$

for a balanced lattice may be too demanding. Since the lattice designs are resolvable, one or more of the replicate groups may be eliminated to get a partially balanced lattice design. Thus, in practice, smaller values for r are applied. Designs with $r = 2$ are called **simple lattice**; those with $r = 3$, **triple lattice** (Example 19.1 provides a triple lattice design); those with $r = 4$, **quadruple lattice**. Cochran and Cox (1957, pp. 428–38) present designs for the lattices.

Incomplete block designs intend to reduce the residual variance of the difference between the mean phenotypic values of two candidates as compared to its value when using a randomized complete block design. The efficiency of an incomplete block design relative to using a randomized complete block design appears from the ratio of the residual variances of the difference between the mean phenotypic values of two candidates. Example 19.2 deals with a comparison of analysis of data according to a lattice design with the analysis according to a randomized complete block design.

Example 19.2 Mak, Harvey and Berdahl (1978) compared a statistical analysis of data according to a lattice design with the analysis according to a randomized complete block design (Example 14.17). The efficiency of data adjustment was the main subject of the study.

Analysis of the data according to a lattice design yielded for grain yield a relative efficiency (RE) of 1.16 and for protein content RE = 1.27. Analysis of covariance, with the moving mean as covariate, yielded at the complete block approach for yield, when involving eight neighbours, RE = 1.25 and for protein content, when involving 10 neighbours, RE = 1.23.

The results suggest that an analysis as a randomized block design, combined with an analysis of covariance using a moving mean as covariate, is a good substitute if – for the actual value of t – a partially balanced lattice design cannot be applied.

Kuehl (2000) provides plans for small experiments ($t \leq 11$) and gives references to plans for other values for t, as well as to computer programmes developing incomplete block design plans, including so-called **alpha designs** (α designs). The latter category of designs were developed by Patterson, Williams and Hunter (1978). A feature that these designs share with lattice designs is that combinations of incomplete blocks coincide with replicates. However, this does not imply that t is a multiple of k.

For $k = 4(1)8$ and $t = 26$, for instance, there is no α-design, such that each block contains the same number of candidates. One should then apply two block sizes, *e.g.* two blocks with $k_1 = 5$ and four blocks with $k_2 = k_1 - 1 = 4$. The use of two block sizes, such that $k_2 = k_1 - 1$, is a typical but not necessary feature of certain α-designs.

Patterson, Williams and Hunter (1978) described the construction of α-designs starting from s (the number of blocks within a complete replicate), k and r. The restrictions are

$$k \leq s, t \leq sk \quad \text{and} \quad t < 100$$

The number of replicates is $r = 2$, 3 or 4. An α-design is called an $\alpha(0,1)$-design if, depending on the considered pair of candidates, the two candidates occur together in either 0 blocks or in one block. Likewise there are $\alpha(0,1,2)$-designs.

Alpha designs are most effective if $k < \sqrt{t}$, *i.e.* if $k^2 < t$. This condition means for situations with $t = ks$, that they are most effective if $k^2 < ks$, *i.e.* if $k < s$. When $k > s$ some pairs of candidates occur together in a block in more than one replicate.

A computer programme generating optimal α-designs for $t = 2(1)500$ has been made available by Williams and Talbot (1993).

Note 19.1 describes how plans for incomplete block designs may be used when designing an incomplete crossing scheme. This is illustrated by Example 19.3.

Note 19.1 If it is not feasible to make a complete diallel or factorial cross, one may consider intercrossing the lines to be studied according to an experimental design using incomplete blocks. Thus, if t maternal genotypes and b paternal genotypes are to be test crossed, each maternal genotype may be pollinated by r paternal genotypes and each paternal genotype should pollinate k maternal genotypes.

Example 19.3 Melchinger (1984) made an incomplete factorial set of crosses involving $t = 11$ maternal dent maize lines and $b = 11$ paternal flint maize lines. The lines were crossed according to the balanced incomplete block design with $k = r = 6$ given by Cochran and Cox (1957, plan 11.20) as well as Kuehl (2000, Plan 9A.16). Thus paternal line 1 pollinated the six maternal lines 4, 6, 7, 9, 10 and 11.

When using (in)complete blocks the selected experimental design requires randomization at several stages:

1. The code numbers 1,2,..., t are assigned at random to the t candidates.
2. The k entries that, according to the design, are to be evaluated in a certain (in)complete block are assigned at random to the k plots in the block.
   For a resolvable design, where combinations of blocks coincide with a complete replicate, this is followed by:
3. Within each replicate, the s incomplete blocks are assigned to random positions.

References

Allard, R.W. (1960). *Principles of Plant Breeding*. Wiley, New York.

Allard, R.W. and Bradshaw, A.D. (1964). Implications of genotype-environmental interactions in applied plant breeding. *Crop Sci.*, **4**, 503–508.

Allard, R.W., Jain, S.K. and Workman, P. (1968). The genetics of inbreeding populations. *Adv. Genet.*, **14**, 55–131.

Arboleda-Rivera, F. and Compton, W.A. (1974). Differential response of maize to mass selection in diverse selection environments. *Theor. Appl. Genet.*, **44**, 77–81.

Baker, R.J. (1986). *Selection Indices in Plant Breeding*. CRC Press, Boca Raton, Florida.

Baker, R.J. and McKenzie, R.I.H. (1967). Use of control plots in yield trials. *Crop Sci.*, **7**, 335–337.

Baltjes, H.J. (1975). Natural Selection in Composite Cross CCXXI of Barley, *Hordeum vulgare* (L). Department of Plant Breeding, Agricultural University, Wageningen.

Barrière, Y. and Argillier, O. (1993). Brown-midrib genes of maize: a review. *Agronomie* **13**, 865–876.

Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Stat.*, **25**, 16–39.

Becker, H.C. (1981). Biometrical and empirical relations between different concepts of phenotypic stability, in *Quantitative Genetics and Breeding Methods* (ed. A. Gallais), INRA, Versailles, pp. 307–314.

Becker, H.C. (1982). Züchtung synthetischer Sorten. II. Leistungsvorhersage und Selektion der Eltern. *Vortr. Pflanzenzüchtg.*, **1**, 23–40.

Becker, H.C., Geiger, H.H. and Morgenstern, K. (1982). Performance and phenotypic stability of different hybrid types in winter rye. *Crop Sci.*, **22**, 340–344.

Becker, H.C. (1988) Breeding synthetic varieties of crop plants. *Plant Genetics and Breeding Reviews*, **1**:31–54.

Bennett, J.H. (1954) On the theory of random mating. *Ann. Eugen.*, **18**, 311–317.

Blakeslee, A.F., Belling, J. and Farnham, M.E. (1923). Inheritance in tetraploid *Datura*. *Bot. Gaz.*, **76**, 329–373.

Bos, I. (1977). More arguments against intermating F2 plants of a self-fertilizing crop. *Euphytica*, **26**, 33–46.

Bos, I. (1980) About the advisable number of backcrosses for autotetraploid crops. *Euphytica*, **29**, 9–15.

Bos, I. (1981). The relative efficiency of honeycomb selection and other procedures for mass selection in winter rye (*Secale cereale* L.). Ph.D. Thesis Agricultural University, Wageningen.

Bos, I. (1983a). The optimum number of replications when testing lines or families on a fixed number of plots. *Euphytica*, **32**, 311–318.

Bos, I. (1983b). About the efficiency of grid selection. *Euphytica*, **32**, 885–893.

Bos, I. (1987). How to develop from three parents a new variety of a self-fertilizing crop? *Euphytica*, **36**, 455–466.

Bos, I. and De Pauw, R.M. (1984). Moving standardized yields as a criterion for selection for yield per plot. *Vortr. Pflanzenzüchtg.*, **7**, 243–258.

Bos, I. and Hennink, S. (1991). A comparison of several procedures for mass selection in winter rye. II. What are the merits of adjusting phenotypic values? *Euphytica*, **52**, 57–64.

Bos, I. and Kleikamp, A. (1985). Reduction of micro-environmental variation in a selection field of rye. *Euphytica*, **34**, 1–6.

Bos, I. and Qi, X. (1997). The relation between agronomic performances of pure lines of barley (*Hordeum vulgare* L.) and AFLP-marker data of these lines, in *Advances in Biometrical Genetics*. Proc. Xth Meeting of the Section Biometrics in Plant Breeding, Eucarpia (eds. P. Krajewski and Z. Kaczmarek), Poznan, pp. 73–79.

Bos, I. and Sparnaaij, L.D. (1993). Component analysis of complex characters in plant breeding. II. The pursuit of heterosis. *Euphytica*, **70**, 237–245.

Bradshaw, J.E. (1986). Competition between cultivars of fodder kale (*Brassica oleracea* L.) in yield trials with single-row plots. *Euphytica*, **35**, 433–439.

Breure, C.J. (1986). Parent selection for yield and bunch index in the oil palm of West New Britain. *Euphytica*, **35**, 65–72.

Breure, C.J. and Verdooren, L.R. (1995). Guidelines for Testing and Selecting Parent Palms in Oil Palm. Practical Aspects and Statistical Methods. ASD oil palm papers **9**, Costa Rica.

Briggs, K.G. and Shebeski, L.H. (1968). Implications concerning the frequency of control plots in wheat breeding nurseries. *Can. J. Plant Sci.*, **48**, 149–153.

Briggs, K.G. and Shebeski, L.H. (1971). Early generation selection for yield and breadmaking quality of hard red spring wheat (*Triticum aestivum* L. em. Thell.). *Euphytica*, **20**, 453–463.

Brim, C.A., Johnson, H.W. and Cockerham, C.C. (1959). Multiple selection criteria in soybeans. *Agron. J.*, **51**, 42–46.

Brown, J. and Caligari, P.D.S. (1986). The efficiency of seedling selection for yield and yield components in a potato breeding programme. *Z. Pflanzenzüchtg.*, **96**, 53–62.

Brown, J. and Caligari, P.D.S. (1988). The use of multivariate cross prediction methods in the breeding of a clonally reproduced crop (*Solanum tuberosum*). *Heredity*, **60**, 147–153.

Brown, J. and Caligari, P.D.S. (1989). Cross prediction in a potato breeding programme by evaluation of parental material. *Theor. Appl. Genet.*, **77**, 246–252.

Brown, J., Caligari, P.D.S. and MacKay, G.R. (1987). The repeatability of progeny means in the early generations of a potato breeding programme. *Ann. Appl. Biol.*, **110**, 365–370.

Brown, J., Caligari, P.D.S., Dale, M.F.B., Swan, G.E.L. and MacKay, G.R. (1988). The use of cross prediction methods in a practical potato breeding programme. *Theor. Appl. Genet.*, **76**, 33–38.

Brown, J., Caligari, P.D.S., MacKay, G.R. and Swan, G.E.L. (1984). The efficiency of seedling selection by visual preference in a potato breeding programme. *J. Agric. Sci.*, **103**, 339–346.

Busbice, T.H. (1969). Inbreeding in synthetic varieties. *Crop Sci.*, **9**, 601–604.

Busbice, T.H. (1970). Predicting yield of synthetic varieties. *Crop Sci.*, **10**, 265–269.

Busbice, T.H. and Wilsie, C.P. (1966). Inbreeding depression and heterosis in autotetraploids with application to *Medicago sativa* L. *Euphytica*, **15**, 52–67.

Bussemakers, A. and Bos, I. (1999). The effect of interplant distance on the effectiveness of honeycomb selection in spring rye. III. Accumulated results of five selection cycles. *Euphytica*, **105**, 229–237.

Caligari, P.D.S. (1980). Competitive interactions in *Drosophila melanogaster*. I. Monocultures. *Heredity*, **45**, 219–231.

Caligari, P.D.S. (1993). *G* × *E Studies in Perennial Tree Crops: Old, Familiar Friend or Awkward, Unwanted Nuisance*. Proceedings of the 1991 International Society of Oil Palm Breeders Workshop, Malaysia. pp. 1–11.

Caligari, P.D.S. and Brown, J. (1986). The use of univariate cross prediction methods in the breeding of a clonally reproduced crop (*Solanum tuberosum*). *Heredity*, **57**, 395–401.

Caligari, P.D.S. and Powell, W. (1986). The effects of competitive interactions on variances and on seed germination in spring barley (*Hordeum vulgare*). *Heredity*, **57**, 331–334.

Caligari, P.D.S., Brown, J. and Manhood, C.A. (1985). The effect of varying the number of drills per plot and the amount of replication on the efficiency of potato yield trials. *Euphytica*, **34**, 291–296.

Caligari, P.D.S., Powell, W. and Jinks, J.L. (1987). A comparison of inbred lines derived by doubled haploidy and single seed descent in spring barley (*Hordeum vulgare*). *Ann. Appl. Biol.*, **111**, 667–675.

Casler, M. (1982). Genotype × environment interaction bias to parent-offspring regression heritability estimates. *Crop Sci.*, **22**, 540–542.

Casler, M. (1992). Usefulness of the grid system in phenotypic selection for smooth bromegrass fiber concentration. *Euphytica*, **63**, 239–243.

Castleberry, R.M., Crum, C.W. and Krull, C.F. (1984). Genetic yield improvement of U.S. maize cultivars under varying fertility and climatic environments. *Crop Sci.*, **24**, 33–36.

Ceccarelli, S., Grando, S. and Impiglia, A. (1998). Choice of selection strategy in breeding barley for stress environments. *Euphytica*, **103**, 307–318.
Cochran, W.G. and Cox, G.M. (1957). *Experimental Designs*. 2nd edn, Wiley, New York.
Cockerham, C.C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis present. *Genetics*, **39**, 859–882.
Compton, W.A. and Comstock, R.E. (1976). More on modified ear-to-row selection in corn. *Crop Sci.*, **16**, 122.
Comstock, R.E. and Moll, R.H. (1963). Genotype-environment interactions, in *Statistical Genetics and Plant Breeding* (eds. W.D. Hanson and H.F. Robinson), Publn 982, National Academy of Sciences, Washington D.C, pp. 164–194.
Comstock, R.E. and Robinson, H.F. (1948). The components of genetic variance in populations of biparental progenies and their use in estimating the average degree of dominance. *Biometrics*, **4**, 254–266.
Comstock, R.E and Robinson, H.F. (1952). Estimation of average dominance of genes, in *Heterosis* (ed. J.W. Gowen), Iowa State College Press, pp. 494–516.
Cotterill, P.P. and Jackson, N. (1985). On index selection. I. Methods of determining economic weight. *Silvae Genetica*, **34**, 56–63.
Crews, W.C., Jones, G.L. and Mason, D.D. (1963). Field plot technique studies with flue-cured tobacco. I. Optimum plot size and shape. *Agron. J.*, **55**, 197–199.
Crosbie, T.M. and Mock, J.J. (1979). Evaluation of plant density tolerance of five maize populations developed by recurrent selection for grain yield at low plant density. *Maydica*, **24**, 141–153.
Crow, J.F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory*, Harper & Row, New York.
Crumpacker, D.W. (1967). Genetic loads in maize (*Zea mays* L.) and other cross-fertilized plants and animals. *Evol. Biol.*, **1**, 306–420.
DePauw, R.M. and Shebeski, L.H. (1973). An evaluation of an early generation yield testing procedure in *Triticum aestivum. Can. J. Plant Sci.*, **53**, 465–470.
De Wolff, F. (1972). Mass selection in maize composites by means of selection indices. *Meded. Landbouwhogeschool* 72–1, Wageningen.
Dickinson-Gibbons, J., Olkin, I. and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology*, Wiley, New York.
Dudley, J.W. (1993). Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. *Crop Sci.*, **33**, 660–668.
Dudley, J.W., Lambert, R.J. and Alexander, D.E. (1974). Seventy generations of selection for oil and protein concentration in the maize kernel, in *Seventy Generations of Selection for Oil and Protein in Maize*, (ed. J.W. Dudley), Crop Science Society of America, Madison, Wisconsin, pp. 181–212.
Dudley, J.W. and Lambert, R.J. (1992). Ninety generations of selection for oil and protein in maize. *Maydica*, **37**, 81–87.
Duvick, D.N. (1992). Genetic contributions to advances in yield of U.S. maize. *Maydica*, **37**, 69–79.
Dyke, G.V. and Shelley, C.F. (1976). Serial designs balanced for effects of neighbours on both sides. *J. Agric. Sci.*, **87**, 303–305.
East, E.M. (1910). A mendelian interpretation of variation that is apparently continuous. *Am. Naturalist*, **44**, 65–82.
East, E.M. (1916). Studies on size inheritance in *Nicotiana*. *Genetics*, **1**, 164–176.
Eberhart, S.A. and Russell, W.A. (1966). Stability parameters for comparing varieties. *Crop Sci.*, **6**, 36–40.
Eenink, A.H. (1974). Linkage in *Spinacia oleracea* L. between the locus for resistance to *Peronospora spinaciae* Laub. and the locus for tolerance for cucumber virus 1. *Euphytica*, **23**, 485–487.
Elgin, J.H., Hill, R.R. and Zeiders, K.E. (1970). Comparison of four methods of multiple trait selection for five traits in alfalfa. *Crop Sci.*, **10**, 190–193.
El Sayed, M.N. and John, C.A. (1973). Heritability studies of tomato emergence at different temperatures. *J. Am. Soc. Hort. Sci.*, **98**, 440–443.

Emerson, R.A. and Smith, H.H. (1950). Inheritance of number of kernel rows in maize. *Cornell Univ. Agric. Exp. St. Memoir* 296, p. 30.
Evans, L.E. and Bhatt, G.M. (1977). Influence of seed size, protein content and cultivar on early seedling vigor in wheat. *Can. J. Plant Sci.*, **57**, 929–935.
Ewens, W.J. (1969). *Population Genetics*. Methuen, London.
Fairfield Smith, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *J. Agric. Sci.*, **28**, 1–23.
Falconer, D.S. (1985). A note on Fishers 'average effect' and 'average excess'. *Genet. Res.* **46**, 337–347.
Falconer, D.S. (1989). *Introduction to Quantitative Genetics*. 3rd edn, Longman, London.
Falconer, D.S. and Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics*, 4th edn, Longmans Green, Harlow.
Faris, D.G. and De Pauw, R.M. (1981). Effects of seeding rate on growth and yield of three spring wheat cultivars. *Field Crops Res.*, **3**, 289–301.
Faris, M.A., de Araujo, M.R.A. and Lira, M. de A. (1981). Yield stability in forage sorghum in Northeastern Brazil. *Crop Sci.*, **21**, 132–134.
Fasoulas, A. (1973). *A New Approach to Breeding Superior Yielding Varieties*. Publn 3, Aristotle University, Thessaloniki.
Fasoulas, A. (1981). *Principles and Methods of Plant Breeding*. Publn 11, Aristotle University, Thessaloniki.
Fasoulas, A. and Tsaftaris, A. (1975). *An Integrated Approach to Plant Breeding and Field Experimentation*. Publn 5, Aristotle University, Thessaloniki.
Fatunla, T. and Frey, K.J. (1976). Repeatability of regression stability indexes for grain yield of oats (*Avena sativa* L.). *Euphytica*, **25**, 21–28.
Federer, W.T. (1956). A method for evaluating genetic progress in a sugar cane breeding program. *Hawaiian Planter's Record*, **55**, 177–189.
Ferguson, J.H.A. (1962). Random variability in horticultural experiments. *Euphytica*, **11**, 213–220.
Finlay, K.W. and Wilkinson, G.N. (1963). The analysis of adaptation in a plant breeding program. *Austr. J. Agric. Res.*, **14**, 742–754.
Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc., Edinburgh*, **52**, 399–433.
Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd, London.
Forkman, G. and Seyffert, W. (1977). Simulation of quantitative characters by genes with biochemically definable action. VI. Modifications of a simple model. *Genetics*, **85**, 557–572.
Foster, C.A. (1971). A study of the theoretical expectation of F1 hybridity resulting from bulk interpopulation hybridization in herbage grasses. *J. Agric. Sci.*, **76**, 295–300.
Frey, K.J. (1971). Improving crop yields through plant breeding, in *Moving off the Yield Plateau* (eds. J.D. Eastin and R.D. Munson), Publn 20, *Am. Soc. Agron.*, Madison, Wisconsin, pp. 15–58.
Frey, K.J. and Horner, T. (1957). Heritability in standard units. *Agron. J.*, **49**, 59–62.
Gallais, A. (1967). Evolution de la vigueur des variétés synthétiques diploïdes au cours des générations de multiplication. I. En panmixie, influence du nombre de parents et du coefficient de consanguinité de départ. *Ann. Amél. Pl.*, **17**, 291–301.
Gallais, A. (2003). Quantitative genetics and breeding methods in autopolypoid plants. INRA Editions, Paris.
Gardner, C.O. (1961). An evaluation of effects of mass selection and seed irradiation with thermal neutrons on yield of corn. *Crop Sci.*, **1**, 241–245.
Gardner, C.O. (1978). Population improvement in maize, in *Maize Breeding and Genetics* (ed. D.B. Walden), Wiley, New York, pp. 207–228.
Gardner, C.O. and Lonnquist, J.H. (1966). Statistical genetic theory and procedures useful in studying varieties and intervarietal crosses in maize, in *Heterosis in Intervarietal Crosses of Maize*, Cimmyt Research Bulletin No. 2.
Gauch, H.G. and Zobel, R.W. (1989). Accuracy and selection success in yield trial analysis. *Theor. Appl. Genet.*, **77**, 473–481.

Geiger, H.H., Diener C. and Singh, R.K. (1981). Influence of self-fertility on the performance of synthetic populations in rye (*Secale cereale* L.), in *Quantitative Genetics and Breeding Methods* (ed. A. Gallais), INRA, Versailles, pp. 169–177.

Genter, C.F. (1967). Inbreeding without inbreeding depression. *Corn Ind. Res. Conf.*, **22**, 82–90.

Genter, C.F. (1982). Recurrent selection for high inbred yields from the F_2 of a maize single cross. *Corn Sorghum Res. Conf.*, **37**, 67–76.

Genter, C.F. and Alexander, M.W. (1962). Comparative performance of S_1 progenies and testcrosses of corn. *Crop Sci.*, **2**, 516–519.

Gomez, K.A. and Gomez, A. (1976). *Statistical Procedures for Agricultural Research*, 2nd edn, International Rice Research Institute, Los Baños.

Gotoh, K. and Osanai, S.I. (1959). Efficiency of selection for yield under different densities in a wheat cross. *Jap. J. Breed.*, **9**, 7–11.

Goulden, C.H. (1939). Problems in plant selection, in *Proceedings of the 7th International Genetics Congress* (ed. R.C. Bunnett), Cambridge University Press, Cambridge, pp. 132–133.

Griffing, B. (1956). Concept of general and specific combining ability in relation to diallel crossing systems. *Austr. J. Biol. Sci.*, **9**, 463–493.

Gupta, S.S. (1956). On a decision rule for a problem in ranking means. Ph.D. thesis (and Mimeograph Series No. 150), Institute of Statistics, University of North Carolina, Chapel Hill.

Hadjichristodoulou, A. and Della, A. (1976). Frequency of control plots in screening nurseries for protein content. *Euphytica*, **25**, 387–391.

Haldane, J.B.S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.*, **8**, 299–309.

Haldane, J.B.S. (1932). *The causes of evolution*. Longman, Green and Co., London.

Hallauer, A.R. and Lopez-Perez, E. (1979). Comparisons among testers for evaluating lines of corn. *Corn Sorghum Res. Conf.*, **34**, 57–75.

Hallauer, A.R. and Miranda, FO, J.B. (1981). *Quantitative Genetics in Maize Breeding*, Iowa State University Press, Ames.

Hallauer, A.R. and Sears, J.H. (1969). Mass selection for yield in two varieties of maize. *Crop Sci.*, **9**, 47–50.

Hallauer, A.R. and Sears, J.H. (1973). Changes in quantitative traits associated with inbreeding in a synthetic variety of maize. *Crop Sci.*, **13**, 327–331.

Hamblin, J. and Donald, C.M. (1974). The relationships between plant form, competitive ability and grain yield in a barley cross. *Euphytica*, **23**, 535–542.

Hardwick, R.C. (1981). The analysis of genotype × environment interactions: what does it mean if varietal stability is linearly related to varietal performance. *Euphytica*, **30**, 217–221.

Hardy, G. (1908). Mendelian proportions in a mixed population. *Science*, **28**, 49–50.

Harper, J.L. (1977). *Population Biology of Plants*. Academic Press, London.

Hatheway, W.H. (1958). Covenient plot size. *Agron. J.*, **53**, 279–280.

Hayes, H.K., Immer F.R. and Smith, D.C. (1955). *Methods of Plant Breeding*, 2nd edn, McGraw-Hill, New York.

Hayman, B.I. (1954). The theory and analysis of diallel crosses. *Genetics*, **39**, 789–809.

Hayward, M.D. and Vivero, J.L. (1984). Selection for yield in *Lolium perenne*. II. Performance of spaced plant selections under competitive conditions. *Euphytica*, **33**, 787–800.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–252.

Hill, J., Mather, K. and Caligari, P.D.S. (1987). Analysis of competitive ability among genotypes of perennial ryegrass. I. Effect upon survival. *Euphytica*, **36**, 99–107.

Horner, E.S., Lundy, H.W., Lutvick, M.C. and Chapman, W.H. (1973). Comparison of three methods of recurrent selection in maize. *Crop Sci.*, **13**, 485–489.

Hühn, M. and Zimmer, E.W. (1983). Einige experimentelle Ergebnisse zur phänotypischen Stabilität von Doppel- und Dreiweghybriden bei Mais. *Z. Pflanzenzüchtg.*, **91**, 246–252.

Inoue, Y. and Kaneko, K. (1976). *Studies on the breeding method of the synthetic varieties from the inbred lines in maize (Zea mays* L.). Research Bulletin of the Hokkaido National Agricultural Experimental Station 114, pp. 195–253.

Jaggard, K.W. (1975). The size and shape of plots in sugar beet experiments. *Ann. Appl. Biol.*, **80**, 351–357.

Jain, S.K. and Allard, R.W. (1960). Population studies in predominantly self-pollinated species. II. Evidence for heterozygote advantage in a closed population of barley. *Proc. Nat. Acad. Sci. USA*, **46**, 1371–1377.

Jain, S.K. and Suneson, C.A. (1964). Population studies in predominantly selfpollinated species. VII. Survival of male-sterility gene in relation to heterozygosis in barley populations. *Genetics*, **50**, 905–913.

Jana, S. (1971). Simulation of quantitative characters from qualitatively acting genes. I. Nonallelic gene interactions involving 2 or 3 loci. *Theor. Appl. Genet.*, **41**, 216–226.

Jana, S. (1975). Genetic analysis by means of diallel graph. *Heredity*, **35**, 1–19.

Jana, S. and Seyffert, W. (1971). Simulation of quantitative characters by genes with biochemically definable action. III. The components of genetic effects in the inheritance of anthocyanins in *Matthiola incana*. R. Br. *Theor. Appl. Genet.*, **41**, 329–337.

Jana, S. and Seyffert, W. (1972). Simulation of quantitative characters by genes with biochemically definable action. IV. The analysis of heritable variation by the diallel technique. *Theor. Appl. Genet.*, **42**, 16–24.

Jenkins, M.T. (1934). Methods of estimating the performance of double crosses in corn. *J. Am. Soc. Agron.*, **26**, 199–204.

Jenkins, M.T. (1935). The effect of inbreeding and of selection within inbred lines of maize upon the hybrids made after successive generations of selfing. *Iowa State Coll. J. Sci.*, **9**, 429–450.

Jensen, N.F. (1970). A diallel selective mating system for cereal breeding. *Crop Sci.*, **10**, 629–635.

Jensen, N.F. (1988). *Plant Breeding Methodology*, Wiley, New York.

Jinks, J.L. (1954). The analysis of continuous variation in a diallel cross of *Nicotiana rustica* varieties. *Genetics*, **39**, 767–788.

Jinks, J.L. (1981). The genetic framework of plant breeding. *Phil. Trans. R. Soc. London B*, **292**, 407–419.

Jinks, J.L. and Perkins, J.M. (1972). Predicting the range of inbred lines. *Heredity*, **28**, 399–403.

Jinks, J.L. and Pooni, H.S. (1976). Predicting the properties of recombinant inbred lines derived by single seed descent. *Heredity*, **36**, 253–266.

Johannsen, W. (1909). *Elemente der exakten Erblichkeitslehre*, Fischer, Jena.

John, P.W.M. (1971). *Statistical Design and Analysis of Experiments*, Macmillan, New York.

Jones, D.F. (1917). Dominance of linked factors as a means of accounting for heterosis. *Genetics*, **2**, 466– 479.

Jones, D.F. (1924). The attainment of homozygosity in inbred strains of maize. *Genetics* **9**, 405–418.

Jones, D.F. (1939). Continued inbreeding in maize. *Genetics*, **24**, 462–473.

Josephson, L.M. (1962). Gamete deletion in male-sterile crosses. *Maize Genet. Coop. News Letter*, **36**, 92–93.

Julén, G. (1959). Weisklee, *Trifolium repens* L., in *Handbuch der Pflanzenzüchtung IV* (eds. H. Kappert and W. Rudorf), Paul Parey, Berlin, pp. 306–320.

Karlin, S. (1968). *Equilibrium Behaviour of Population Genetic Models with Non-random Mating*, Gordon and Breach, New York.

Kearsey, M.J. (1984). A biometrical approach to vegetable breeding. *Vortr. Pflanzenzüchtg.*, **7**, 99–111.

Kearsey, M.J. and Pooni, H.S. (1996). The genetical analysis of quantitative traits. Chapman and Hall, London.

Keeler, C. (1968). Some oddities in the delayed appreciation of 'Castle's Law', *J. Heredity* **59**, 110–112.

Kelker, H.A. and Briggs, K.G. (1979). *The Effects of Intergenotypic Competition and Interplant Spacings in Simulated Segregating Rows of Wheat*, Proceedings of the 5th International Wheat Genetics Symposium, II, New Delhi, pp. 778–786.
Kempthorne, O. (1957). *An Introduction to Genetic Statistics*, Wiley, New York.
Kempton, R.A. (1984) The design and analysis of unreplicated field trials. *Vortr. Pflanzenzüchtg.*, **7**, 219–242.
Kempton, R.A., Gregory, R.S., Hughes, W.G. and Stoehr, P.J. (1986). The effect of interplot competition on yield assessment in *Triticale* trials. *Euphytica*, **35**, 257–265.
Kiesselbach, T.A. (1960). Performance of advanced generation corn hybrids. *Agron. J.* **52**, 29–32.
Kinman, M.L. and Sprague, G.F. (1945). Relation between number of parental lines and theoretical performance of synthetic varieties of corn. *J. Am. Soc. Agron.*, **37**, 341–351.
Kira, T., Ogawa, H. and Sakazaki, N. (1953). Intraspecific competition among higher plants. I. Competition-yield-density interrelationship in regularly dispersed populations. *J. Inst. Polytech. Osaka Cy Univ., Ser.* D, **4**, 1–16.
Knight, R. (1970). The measurement and interpretation of genotype-environment interactions. *Euphytica*, **19**, 225–235.
Knight, R. (1973). The relation between hybrid vigour and genotype-environment interactions. *Theor. Appl. Genet.*, **43**, 311–318.
Knight, R. (1983). Some factors causing variation in the yield of individual plants of wheat. *Aust. J. Agric. Res.*, **34**, 219–228.
Knott, D.R. (1972). Effects of selection for F2 plant yield on subsequent generations in wheat. *Can. J. Plant Sci.*, **52**, 721–726.
Koch, E.J. and Rigney, J.A. (1951). A method of estimating optimum plot size from experimental data. *J. Am. Soc. Agron.*, **43**, 17–21.
Koch, H.D. and Degner, P. (1977). Anwendung der Methode der Kreuzungsverbände in der Wintergerstenzüchtung, unter Verwendung genetisch bedingter männlicher Sterilität. *Tag. Ber. Akad. Landwirtsch. Wiss., DDR*, **158**, 271–275.
Kosambi, D.D. (1944). The estimation of map distances from recombination values. *Ann. Eugen.* **12**, 172–175.
Koutsika-Sotiriou, M., Bos, I. and Fasoulas, A. (1990). Hybrid reconstruction in maize. *Euphytica* **45**, 257– 266.
Kramer, Th. (1983). *Fundamental Considerations on the Density-dependence of the Selection Response to Plant Selection in Wheat*, Proceedings of the 6th International Wheat Genetics Symposium, Kyoto, pp. 719–724.
Kramer, Th., van Ooyen, J.W. and Spitters, C.J.T. (1982). Selection for yield in small plots of spring wheat. *Euphytica*, **31**, 549–564.
Kuehl, R.O. (2000). *Design of Experiments. Statistical Principles of Research Design and Analysis*, 2nd edn, Brooks/Cole, Pacific Grove.
Kyriakou, D.T. and Fasoulas, A.C. (1985). Effects of competition and selection pressure on yield response in winter rye (*Secale cereale* L.). *Euphytica*, **34**, 883–895.
LeClerg, E.L., Leonard, W.H. and Clark, A.G. (1962). *Field Plot Technique*, 2nd edn, Burgess, Minneapolis.
Lerner, I.M. (1950). *Population Genetics and Animal Improvement*. Cambridge University Press, Cambridge.
Lerner, I.M. (1958). *The Genetic Basis of Selection*, Wiley, New York.
Li, C.C. (1976). *First Course in Population Genetics*, Boxwood Press, Pacific Grove.
Lin, C.S. and Binns, M.R. (1984). Working rules for determining the plot size and number of plots per block in field experiments. *J. Agric. Sci.*, **103**, 11–15.
Lonnquist, J.H. (1964). A modification of the ear-to-row procedure for the improvement of maize populations. *Crop Sci.*, **4**, 227–228.
Lonnquist, J.H. (1967). Mass selection for prolificacy in maize. *Der Züchter*, **37**, 185–188.
Lupton, F.G.H. (1961). Studies in the breeding of self- pollinating cereals. 3. Further studies in cross prediction. *Euphytica*, **10**, 209–224.
Lush, J.L. (1945). *Animal Breeding Plans*, Iowa State College Press.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinnauer Associates.

MacArthur, J.W. and Butler, L. (1938). Size inheritance and geometric growth processes in the tomato fruit. *Genetics*, **23**, 253–268.

Mac Key, J. (1976). Genetic and evolutionary principles of heterosis. p. 17–33. Heterosis in plant breeding. Proc. 7th Congr. Eucarpia, Budapest.

Mak, C., Harvey, B.L. and Berdahl, J.D. (1978). An evaluation of control plots and moving means for error control in barley nurseries. *Crop Sci.*, **18**, 870–873.

Manneh, B. (2004). Genetic, physiological and modelling approaches towards tolerance to salinity and low nitrogen supply in rice (*Oryza sativa* L.), PhD Thesis, Wageningen University.

Mather, K. (1949). *Biometrical genetics*. Methuen, London.

Mather, K. (1973). *Genetical Structure of Populations*. Chapman, London.

Mather, K. and Caligari, P.D.S. (1981). Competitive interactions in *Drosophila melanogaster*. II. Measurement of competition. *Heredity*, **46**, 239–254.

Mather, K. and Caligari, P.D.S. (1983). Pressure and response in competitive interactions. *Heredity*, **51**, 435–454.

Mather, K. and Jinks, J.L. (1977). *Introduction to Biometrical Genetics*, Chapman, London.

Mather, K. and Jinks, J.L. (1982). *Biometrical Genetics*, 3rd edn, Chapman, London.

McGinnes, R.C. and Shebeski, L.H. (1968). The reliability of single plant selection for yield in F_2, in *Proceedings of the International Wheat Genetics Symposium*, Canberra (eds. K.W. Finlay and K.W. Shepherd), Butterworths, London, pp. 410–415.

McVetty, P.B.E. and Evans, L.E. (1980). Breeding methodology in wheat. II. Productivity, harvest index, and height measured on F_2 spaced plants for yield selection in spring wheat. *Crop Sci.*, **20**, 587–589.

Melchinger, A.E. (1984). Analysis of incomplete factorial mating designs. *Vortr. Pflanzenzüchtg.*, **7**, 131–150.

Mitchell, K.J. and Lucanus, R. (1962). Growth of pasture species under controlled environments. III. Growth at various levels of constant temperature with 8 and 16 hours of uniform light per day. *N.Z. J. Agric. Res.*, **5**, 135–144.

Mitchell, J.W., Baker, R.J. and Knott, D.R. (1982). Evaluation of honeycomb selection for single plant yield in durum wheat. *Crop Sci.*, **22**, 840–843.

Morgan, J.P. (1988). Polycross designs with complete neighbor balance. *Euphytica*, **39**, 59–63.

Nass, H.G. (1980). Harvest index as a selection criterion for grain yield in two spring wheat crosses grown at two population densities. *Can. J. Plant Sci.*, **60**, 1141–1146.

Neal, N.P. (1935). The decrease in yielding capacity in advanced generations of hybrid corn. *J. Am. Soc. Agron.*, **27**, 666–670.

Nilsson-Ehle, H. (1909). *Kreuzungsuntersuchungen an Hafer und Weizen*, Univ. Aarskr. Lund.

Oleson, K. (1976). A completely balanced polycross design. *Euphytica*, **25**, 485–488.

Oleson, K. and Oleson, O.J. (1973). A polycross pattern formula. *Euphytica*, **22**:500–502.

Omolo, E. and Russell, W.A. (1971). Genetic effects of population size in the reproduction of two heterogeneous maize populations. *Iowa St. Coll. J. Sci.*, **45**, 499–512.

Papadakis, J.S. (1937). *Méthode statistique pour des expériences sur champ.* Bull. Inst. d'Amélioration des Plantes à Salonique, No. 23.

Pasini, R.J. and Bos, I. (1990a). The effect of interplant distance on the effectiveness of honeycomb selection. I. Results of the first selection cycle. *Euphytica*, **49**, 121–130.

Pasini, R.J. and Bos, I. (1990b). The effect of interplant distance on the effectiveness of honeycomb selection. II. Results of the second selection cycle. *Euphytica*, **50**, 147–153.

Patanothai, A. and Atkins, R.E. (1971). Heterotic response for vegetative growth and fruiting development in grain Sorghum, *Sorghum bicolor* (L) Moench, *Crop Sci.*, **11**, 839–843.

Patterson, H.D., Silvey, V., Talbot, M. and Weatherup, S.T.C. (1977). Variability of yields of cereal varieties in U.K. trials. *J. Agric., Sci.*, **89**, 239–245.

Patterson, H.D., Williams, E.R. and Hunter, E.A. (1978). Block designs for variety trials. *J. Agric. Sci.*, **90**, 395–400.

Pearson, E.S. and Hartley, H.O. (1970). *Biometrika Tables for Statisticians*, vol. I, Cambridge University Press, Cambridge.

Pearson, E.S. and Hartley, H.O. (1972). *Biometrika Tables for Statisticians,* vol. II, Cambridge University Press, Cambridge.

Pfeffer, C., Werner, E., Effmert, B. and Reda, S. (1982). Die Selektion identischer Kartoffelpopulationen in zwei Umwelten, *Arch. Züchtungsforsch.*, **12**, 359–365.

Pijper, G.T. (1981). Enkele correctiemethoden op vruchtbaarheidsverloop in het proefveld. Department of Plant Breeding, Agricultural University, Wageningen.

Pinthus, M.J. (1972). A suggested method to estimate the economic value of plant breeding programmes. *Z. Pflanzenzüchtg.*, **68**, 258–260.

Plaisted, R.L. and Peterson, L.C.A. (1959). Technique for evaluating the ability of selections to yield consistency in different locations or seasons. *Am. Potato J.*, **36**, 381–385.

Poehlman, J.M. and Sleper, D.A. (2006). *Breeding Field Crops*, 5th edn, Blackwell Publishing. Ames, IA, USA.

Potz, H. (1987). Untersuchungen zur genetischen Konstitution von *Allium porrum* L., Inst. angewandte Genetik, University of Hannover.

Powell, W., Caligari, P.D.S. and Thomas, W.T.B. (1986). Comparison of spring barley lines produced by single seed descent, pedigree inbreeding and doubled haploidy. *Plant Breeding*, **97**, 138–146.

Powell, W., Caligari, P.D.S., Goudappel, P.H. and Thomas, W.T.B. (1985a). Competitive effects in monocultures and mixtures of spring barley (*Hordeum vulgare*). *Theor. Appl. Genet.*, **71**, 443–450.

Powell, W., Caligari, P.D.S., McNicol, J.W. and Jinks, J.L. (1985b). The use of doubled haploids in barley breeding. 3. An assessment of multivariate cross prediction methods. *Heredity*, **55**, 249–254.

Powell, W., Caligari, P.D.S., Phillips, M.S. and Jinks, J.L. (1986). The measurement and interpretation of genotype by environment interaction in spring barley (*Hordeum vulgare*). *Heredity*, **56**, 255–262.

Powers, L. (1944). An expansion of Jones' theory for the explanation of heterosis, *Am. Naturalist*, **78**, 275–280.

Quiros, C.E. (1982). Tetrasomic segregation for multiple alleles in alfalfa, *Genetics*, **101**, 117–127.

Rasmusson, D.C. (1987). An evaluation of ideotype breeding. *Crop Sci.*, **27**, 1140–1146.

Rasmusson, D.C. and Lambert, J.W. (1961). Variety × environment interactions in barley variety tests. *Crop Sci.*, **1**, 261–262.

Rasmusson, J. (1933). A contribution to the theory of quantitative character inheritance. *Hereditas*, **18**, 245–261.

Reich, V.H. and Atkins, R.E. (1970). Yield stability of four population types of grain sorghum, *Sorghum bicolor* (L.) Moench., in different environments. *Crop Sci.*, **10**, 511–517.

Rieger, R., Michaelis, A. and Green, M.M. (1991). *Glossary of Genetics. Classical and Molecular*, 5th edn, Springer, Berlin.

Rosielle, A.A. and Frey, K.J. (1975). Estimates of selection parameters associated with harvest index in oat lines derived from a bulk population. *Euphytica*, **24**, 121–131.

Rosielle, A.A. and Hamblin, J. (1981). Theoretical aspects of selection for yield in stress and non-stress environments. *Crop Sci.*, **21**, 943–946.

Sakai, K.J. (1961). Competitive ability in plants: its inheritance and some related problems. Symposium of the Society for Experimental Biology, **15**, 245–263.

Satorre, E.H. and Snaydon, R.W. (1992). A comparison of root and shoot competition between spring cereals and *Avena fatua* L. *Weed Research*, **32**, 45–55.

Schnell, F.W. and Becker, H.C. (1986). Yield and yield stability in a balanced system of widely differing population structures in *Zea mays* L. *Plant Breeding*, **97**, 30–38.

Schut, J.W. (1998). Prediction of cross performance in barley. Thesis Wageningen Agricultural University.
Schutz, W.M. and Bernard, R.L. (1967). Genotype × environment interactions in the regional testing of soybean strains. *Crop Sci.*, **7**, 125–130.
Seyffert, W. (1959). Theoretische Untersuchungen über die Zusammensetzung tetrasomer Populationen. II. Selbstbefruchtung. *Z. f. Vererbungslehre*, **90**, 356–374.
Seyffert, W. (1960). Theoretische Untersuchungen über die Zusammensetzung tetrasomer Populationen. I. Panmixie. *Biom. Zeitschr.*, **2**, 1–44.
Shebeski, L.H. (1970). *Wheat and breeding.* In: K.F. Nielsen (ed.), Proc. Canad. Centennial Wheat Symp., Modern Press. Saskatoon.
Shorter, R. and Butler, D. (1985). Effect of moving mean covariance adjustments on error and genetic variance estimates and selection of superior lines in peanuts (*Arachis hypogaea* L.). *Euphytica*, **35**, 185–192.
Shull, G.H. (1909). A pure line method of corn breeding. *Am. Breed. Assoc. Rept.*, **5**, 51–59.
Simmonds, N.W (Ed.) (1976). Evolution of Crop Plants, Longman, London.
Singh, M., Khehra, A.S. and Dhillon, B.S. (1986). Direct and correlated response to recurrent full-sib selection for prolificacy in maize. *Crop Sci.*, **26**, 275–278.
Smith, H.F. (1936). A discriminant function for plant selection. *Ann. Eugenics*, **7**, 240–250.
Smith, J.D. and Kinman, M.L. (1965). The use of parent-offspring regression as an estimator of heritability. *Crop Sci.*, **5**, 595–596.
Snape, J.W. (1997). Applications of doubled haploid lines in plant breeding and genetical research: current issues and approaches, in *Advances in Biometrical Genetics.* Proc. Xth Meeting of the Section Biometrics in Plant Breeding, Eucarpia (eds. P. Krajewski and Z. Kaczmarek), Poznan.
Snedecor, G.W. and Cochran, W.G. (1980). *Statistical Methods*, 7th edn, Iowa State University Press, Ames.
Sneep, J. (1977). Selection for yield in early generations of self-fertilizing crops. *Euphytica*, **26**, 27–30.
Sneep, J. and Hendriksen, A.J.T. (eds.) (1979). *Plant breeding perspectives*, Pudoc, Wageningen.
Soetono and Donald, C.M. (1980). Emergence, growth and dominance in drilled and square-planted barley crops. *Austr. J. Agric. Res.*, **31**, 455–470.
Soliman, K.M. and Allard, R.W. (1991). Grain yield of composite cross populations of barley: effects of natural selection. *Crop Sci.*, **31**, 707–708.
Sparnaaij, L.D. and Bos, I. (1993). Component analysis of complex characters in plant breeding. I. Proposed method for quantifying the relative contribution of individual components to variation of the complex character. *Euphytica*, **70**, 225–235.
Spitters, C.J.T. (1979). Competition and its consequences for selection in barley breeding. *Agricultural Research Reports* 893, Wageningen.
Sprague, G.F. and Tatum, L.A. (1942). General vs. specific combining ability in single crosses of corn. *J. Am. Soc. Agron.*, **34**, 923–932.
Stam, P. (1977). Selection response under random mating and under selfing in the progeny of a cross of homozygous parents. *Euphytica*, **26**, 169–184.
Stam, P. (1984). Estimation of genotypic values without replication in field trials. *Euphytica*, **33**, 841–852.
Stam, P. (1998). Crop physiology, QTL analysis and plant breeding, in *Inherent variation in plant growth. Physiological mechanisms and ecological consequences* (eds. H. Lambers, H. Poorter and M.M.I. van Vuuren), Backhuys Publishers, Leiden, pp. 429–440.
Stam, P. and van Ooijen, J.W. (1995). JoinMap (tm) version 2.0: Software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen.
Steel, R.G.D. and Torrie, J.H. (1980). *Principles and Procedures of Statistics*, 2nd edn, McGraw-Hill, New York.
Strickberger, M.W. (1976). *Genetics*, 2nd edn, MacMillan, New York.
Suneson, C.A. (1956). An evolutionary plant breeding method. *Agron. J.*, **48**, 188–191.

Swallow, W.H. and Wehner, T.C. (1986). Optimum plot size determination and its application to cucumber yield trials. *Euphytica*, **35**, 421–432.

Thoday, J.M. (1976). Effects of specific genes, in *Proceedings of the International Conference on Quantitative Genetics* (eds. E. Pollak, O. Kempthorne and J.B. Bailey), Iowa State University Press, Ames, pp. 141–159.

Thoday, J.M. and Thompson, J.N. (1976). The number of segregating genes by continuous variation. *Genetica*, **46**, 335–344.

Townley-Smith, T.F. and Hurd, E.A. (1973). Use of moving means in wheat yield trials. *Can. J. Plant Sci.*, **53**, 447–450.

Townley-Smith, T.F., Hurd, E.A. and McBean, D.S. (1973). Techniques of selection for yield in wheat, in *Proceedings of the International Wheat Genetics Symposium, Missouri Agricultural Experimental Station, Columbia* (eds. E.R. Sears and L.M.S. Sears), Kimber, Columbia, Mo, pp. 605–609.

Van Cruchten, C.J.M. (1973). Verband tussen inteeltlijnen en hun kruisingsprodukten bij mais. Department of Plant Breeding, Agricultural University, Wageningen.

Van der Laan, P. and Verdooren, L.R. (1990). A review with some applications of statistical selection procedures for selecting the best variety. *Euphytica*, **51**, 67–75.

Van der Vossen, H.A.M. (1974). *Towards more Efficient Selection for Oil Yield in the Oil Palm* (Elaeis guineensis *Jacquin*). Pudoc, Wageningen.

Van Hintum, T.J.L. and van Adrichem, B.N.M. (1986). De effecten van directe en indirecte-directe massaselectie op biomassa bij mais (*Zea mays* L.). Department of Plant Breeding, Agricultural University, Wageningen.

Van Oeveren, A.J. (1993) Efficiency of single seed descent and early selection in the breeding of self-fertilizing crops. Wageningen Agricultural University dissertation 1578.

Van Ooijen, J.W. (1989). Estimation of additive genotypic variance with the F_3 of autogamous crops. *Heredity*, **63**, 73–81.

Van Ooijen, J.W. and Maliepaard, C. (1995). MapQTL (tm) version 3.0: Software for the calculation of QTL positions on genetic maps, CPRO-DLO, Wageningen.

Vela-Cardenas, M. and Frey, K.J. (1972). Optimum environment for maximizing heritability and genetic gain from selection. *Iowa State Coll. J. Sci.*, **46**, 381–394.

Vencovsky, R. and Godoi, C.R.M. (1976). Immediate response and probability of fixation of favourable alleles in some selection schemes, in *Proceedings of the International Biometrics Conference, Boston*, The Biometric Society, Raleigh, pp. 292–297.

Verhalen, L.M., Baker, J.L. and McNew, R.W. (1975). Gardner's grid system and plant selection efficiency in cotton. *Crop Sci.*, **15**, 588–591.

Ward, S.M. (2000). Allotetraploid segregation for single-gene morphological characters in quinoa (*Chenopodium quinoa* Willd.). *Euphytica*, **116**, 11–16.

Webel, O.D. and Lonnquist, J.H. (1967). An evaluation of modified ear-to-row selection in a population of corn (*Zea mays* L.). *Crop Sci.*, **7**, 651–655.

Weber, C.R. and Moorthy, B.R. (1952). Heritable and non-heritable relationships and variability of oil content and agronomic characters in F_2 generation of soybean crosses. *Agron. J.*, **44**, 202–209.

Weber, E. (1978). *Mathematische Grundlagen der Genetik.* vol 2, Aufl. Fischer, Jena.

Weber, W.E. and Stam, P., (1988). On the optimum grid size in field experiments without replications. *Euphytica*, **39**, 237–247.

Weber, W.E. and Wricke, G. (1994). Genetic markers in plant breeding. *Adv. Plant Breeding* **16**.

Weinbaum, S.A., Shaw, D.V., Azari, R. and Muraoka, T.T. (1990). Mass selection of walnut rootstocks using response surface methods to correct for environmental trends. *Euphytica*, **46**, 227–235.

Weinberg, W. (1908). Über den Nachweis der Vererbung beim Menschen. *Verein Naturk.*, **64**, 368–382.

Weir, B.S. and Cockerham, C.C. (1973). Mixed self and random mating at two loci. *Genet. Res.*, **21**, 247–262.

Whan, B.R., Knight, R. and Rathjen, A.J. (1982). Response to selection for grain yield and harvest index in F_2, F_3 and F_4 derived lines of two wheat crosses. *Euphytica*, **31**, 139–150.

Whan, B.R., Rathjen, A.J. and Knight, R. (1981). The relation between wheat lines derived from the F_2, F_3, F_4 and F_5 generations for grain yield and harvest index. *Euphytica*, **30**, 419–430.

Wilkinson, G.N., Eckert, S.R., Hancock, T.W. *et al.* (1983). A new statistical methodology for design and analysis of plant breeding and varietal field trials, in *Australian Plant Breeding Conference, Adelaide*, (ed. C.J. Driscoll), University of Adelaide, pp. 59–65.

Williams, E.R. and Talbot, M. (1993). ALPHA + Experimental designs for variety trials, CSIRO, Canberra, and BioSS, Edinburgh.

Workman, P.L. and Allard, R.W. (1962). Population studies in predominantly self-pollinated species. III. A matrix model for mixed selfing and random outcrossing. *Proc. Nat. Acad. Sci. USA*, **48**, 1318–1325.

Wricke, G. (1964). Zur Berechnung der Ökovalenz bei Sommerweizen und Hafer. *Z. Pflanzenzüchtg.*, **52**, 127–138.

Wright, S. (1921). Systems of mating. *Genetics*, **6**, 111–178.

Wright, S. (1922). *The Effects of Inbreeding and Cross-breeding on Guinea Pigs. III. Crosses Between Highly Inbred Families*. US Department of Agriculture Bulletin 1121.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.*, **15**, 323–354.

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci.*, **26**, 424–455.

Index