

## Chapter 2

# Means and Variances

### 2.1 Genetically Narrow- vs. Broad-Based Reference Populations

Choice of germplasm as source of elite inbred lines is the most important decision the breeder takes. No tool or breeding methodology will be successful if a poor choice is made on source populations.

A population of maize can be characterized by the following properties: diploid ( $2n = 20$ ), panmictic (random mating with more than 95% of cross-pollination), monoecious (both sexes in the same individual but in different inflorescences), a tendency for protandry, and general assumptions for no maternal effects, linkage equilibrium, normal fertilization (non-competing gametes), normal meiosis, and normal segregation.

Both means and genetic variances are important factors to consider when choosing populations to be used as sources of inbred lines and hybrids. Choosing breeding populations with a high mean performance is straightforward. However, the study of genetic variation of plant populations includes different approaches for different types of populations. The reference population of genotypes may result from genetically narrow-based populations derived from a cross between two homozygous inbred lines or from genetically broad-based populations derived from improved and/or unimproved populations. Broad-based populations can be a result of crosses among a set of homozygous inbred lines (synthetic varieties), an open-pollinated variety, or a mixture of varieties and races (composites). General theories, however, make no distinction about the origin of the population unless it does not fill some of the basic requirements.

Populations derived from crosses of two elite pure lines are commonly used in plant breeding. Consequently, we can determine the genetic composition of different generations derived from crossing two pure lines, including backcross populations. The introduction to the estimation of genetic variances in these generations has the advantages that, assuming two alleles per locus, expected gene frequencies ( $p$  and  $q$ ) are known and have the same value ( $p = q = 0.5$ ) for segregating loci, which makes their derivations easy to interpret unlike genetically broad-based populations. The estimation of genetic variation within genetically broad-based populations in which the allele frequencies are not known is based on mating designs to develop progenies for evaluation. These progenies are based on the genetic composition for covariance

of relatives (see Chapter 3). Analyses of variance of the progenies derived from mating designs are used to evaluate additive and dominance genetic effects, average level of dominance, epistasis, and relative heritability as well as expected genetic gain. Public breeding programs allow growing progenies for not only estimating genetic variances but also for selection without relying on just the coefficient of co-ancestry. Estimating genetic variances is useful for designing breeding programs, predicting response to selection, constructing selection indices, predicting hybrid performance, and allocating breeding resources more efficiently (Bernardo, 2002).

The concepts of population means and variances in current quantitative genetics theory are based on gene effects and frequencies or, in other words, on the genetic structure of the population under study. The population structure, however, depends on several other factors such as ploidy level, linkage, mating system, and a number of environmental and genetic factors. Therefore, either some of these factors must be known or restrictions must be imposed about their effects to be able to establish a theoretical model for study.

Estimated parameters refer to a specific population from which the experimental material is a sample for a specific set of environmental conditions (Cockerham, 1963). Thus one must specify the reference population for both genotypes and environments because inferences cannot generally be translated from one population to another especially after selection. In genome-wide selection, for instance, molecular markers need to be ‘re-trained’ (Hammond, personal communication) after each time selection is conducted even within populations (e.g., across recurrent selection cycles). More detailed descriptions of the population means and variances were given by Kempthorne (1957) and Falconer (1960).

## 2.2 Hardy–Weinberg Equilibrium

Assume the reference population is in Hardy–Weinberg equilibrium. In 1908 Hardy and Weinberg independently demonstrated that in a large random mating population both gene frequencies and genotypic frequencies remain constant from generation to generation in the absence of mutation, migration, and selection. Such a population is said to be in Hardy–Weinberg equilibrium and remains so unless any disturbing force changes its gene or genotypic frequency.

This concept can be translated to a single locus as any population will attain its equilibrium after one generation of random mating. The Hardy–Weinberg equilibrium law can be demonstrated by taking one locus with two alleles ( $A_1$  and  $A_2$ ) in a diploid organism such as maize. Let us consider a population whose genotypic frequencies are as follows:

Genotypes	$A_1A_1$	$A_1A_2$	$A_2A_2$	
Number of individuals	$n_1$	$n_2$	$n_3$	$n_1 + n_2 + n_3 = N$
Frequency	$P = n_1/N$	$Q = n_2/N$	$R = n_3/N$	$P + Q + R = 1$

The total number of genes relative to locus A in this population is  $2N$ , i.e., two genes in each diploid individual. Thus the numbers of  $A_1$  and  $A_2$  genes are  $2n_1 + n_2$  and  $2n_3 + n_2$ , respectively, and their frequencies are

$$p(A_1) = \frac{2n_1+n_2}{2N} = \frac{n_1+(\frac{1}{2})n_2}{N} = P + \frac{1}{2}Q$$

$$q(A_2) = \frac{2n_3+n_2}{2N} = \frac{n_3+(\frac{1}{2})n_2}{N} = R + \frac{1}{2}Q$$

Because gametes unite at random in a population under random mating, the genotypic array and its frequency in the next generation will be

Genotypes	Male gametes		Frequencies	Male gametes	
	$A_1$	$A_2$		$p$	$q$
Female gametes	$A_1$	$A_1A_1$ $A_1A_2$	Female gametes	$p$	$p^2$ $pq$
	$A_2$	$A_1A_2$ $A_2A_2$		$q$	$pq$ $q^2$

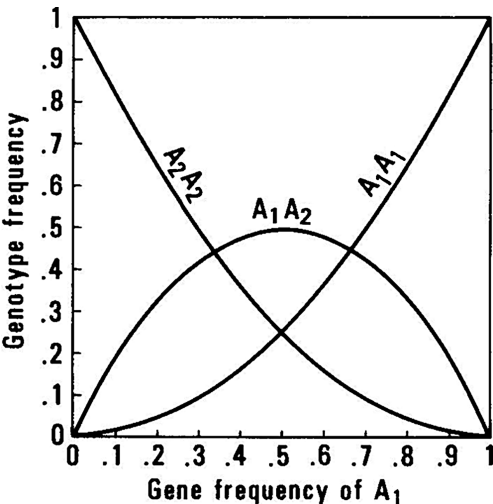
So the genotypic frequencies are  $p^2(A_1A_1) : 2pq(A_1A_2) : q^2(A_2A_2)$ , and this population is said to be in Hardy–Weinberg equilibrium since genotypic frequencies are expected to be unchanged in the next generation. Figure 2.1 shows the variation of genotypic frequencies for gene frequencies in the range from 0 to 1.

The Hardy–Weinberg law can also be extended to multiple alleles. In general, if  $p_i$  is the frequency of the  $i$ th allele at a given locus, the genotypic frequency array is given by

$$\sum_i p_i^2 \quad \text{for homozygotes } (A_iA_i)$$

$$2 \sum_{i < i'} p_i p_{i'} \quad \text{for heterozygotes } (A_iA_{i'})$$

**Fig. 2.1** Distributions of genotypic frequencies for gene frequencies ranging from 0 to 1.0 for one locus with two alleles in a population in Hardy–Weinberg equilibrium



With two alleles per locus the gene frequency that gives the maximum frequency of heterozygotes ( $Q = 2pq$ ) is found when  $p = 0.5$ . Therefore, in  $F_2$  populations derived from elite  $\times$  elite pure line crosses we expect maximum frequency of heterozygotes.

## 2.3 Means of Non-inbred Populations and Derived Families

A population of *phenotypes* (Fig. 2.2) can be characterized in terms of not only its gene and genotypic frequencies but also its mean and variance for a quantitative trait. Environmental factors largely influence the expression of these traits. These traits are studied by measures of central tendency and dispersion instead of phenotypic ratios. Genomic tools may provide additional information on gene information and the genetic architecture of quantitative traits as long as sample sizes are representative and a random set of populations is involved.

Base  
Pairs>Genes>Chromosomes>Genotype>Environment>**PHENOTYPE>POPULATION**

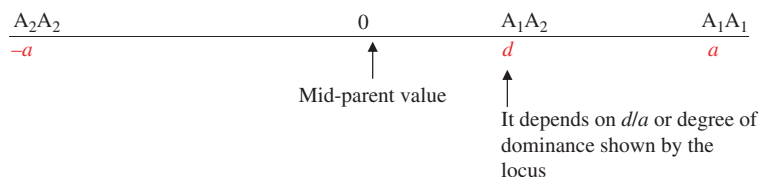
**Fig. 2.2** A population of phenotypes is made up of genotype and environment

A *phenotypic value* is an observed measure of its effect on the quantitative trait and can be measured. The values associated with genotypes are measured indirectly from the corresponding phenotypic values.

The phenotypic value can be divided into genotypic value and its environmental deviation as follows:

$$P \text{ (phenotypic value)} = G \text{ (genotypic value)} + E \text{ (environmental deviation)}$$

Therefore, phenotypic values are due to genetic and non-genetic circumstances. It is still challenging to accurately measure the *genotypic value* of an individual. However, it can be measured if we use a simple genetic model (one locus and two alleles) where the genotypes are distinguishable in their phenotype (e.g., inbred lines). So if we assign arbitrary values to the genotypes we can build a scale of genotypic values:



where  $d$  and  $d/a$  are related to the level of dominance

The degree of dominance for genes affecting plant height in maize is different from the level of dominance, if any, for genes affecting plant height in a self-pollinating crop like wheat. Hybrid vigor in maize is important and the difference

If $d = 0$	$d/a = 0$	No dominance
If $d > 0$ but $< a$	$0 < d/a < 1$	Partial dominance
If $d = a$ (or $-a$ )	$d/a = 1$	Complete dominance
If $d > a$	$d/a > 1$	Overdominance

between inbred lines and hybrids for plant height is significant and so is the dominance level for genes affecting these types of traits.

How do gene frequencies affect the mean of a trait in a population?

Considering one locus with two alleles,  $A_1$  and  $A_2$ , it is assumed that each locus has a particular effect on the total individual phenotype. Arbitrarily assuming  $A_1$  to be the allele that increases the value, we can denote by  $+a$ ,  $-a$ , and  $d$  the effects of genotypes  $A_1A_1$ ,  $A_2A_2$ , and  $A_1A_2$ , respectively. Such effects are taken as deviations from the mean of the two homozygotes, as shown on the linear scale earlier.

The population mean is thus calculated considering both the genotypic frequencies and genotypic effects (coded values), as shown in Table 2.1. Let gene frequencies for ‘ $A_1$ ’ and ‘ $A_2$ ’ be  $p$  and  $q$ , respectively.

**Table 2.1** Genotypic values and frequencies in a population in Hardy–Weinberg equilibrium for one locus with two alleles

Genotypes	Frequency ( $F_i$ )	# of ‘ $A_1$ ’ alleles	Genotypic values <sup>a</sup> ( $X_i$ )			
$A_1A_1$	$p^2$	2	<b>a</b>	$d$	$u$	$z$
$A_1A_2$	$2pq$	1	<b>d</b>	$\hat{h}$	$au$	$\hat{h}$
$A_2A_2$	$q^2$	0	<b>–a</b>	$–d$	$–u$	$o$

The first two columns show the three genotypes and their frequencies in a random mating population

<sup>a</sup>Different symbols across literature. We will use  $a$ ,  $d$ , and  $–a$

The mean value of a population is obtained by multiplying the values of each genotype ( $X_i$  or genotypic effect) by their frequencies ( $f_i$ ). Then we sum over the three genotypes.

$$\bar{X} = \sum (X_i F_i) \text{ or } \sum X_i / n$$

Since the sum of frequencies is 1 ( $p + q = 1$ ) the sum of values multiplied by frequencies is the mean value:

$$\begin{aligned} \bar{X} &= p^2 a + 2pqd - q^2 a \\ &= (p^2 - q^2)a + 2pqd \\ &= (p+q)(p-q)a + 2pqd \quad \text{since } p+q=1 \\ &= (p-q)a + 2pqd \end{aligned}$$

$$\bar{X} = \sum (p-q)a + 2 \sum pqd$$

After the contribution of several loci

As seen above, the mean will vary according to the level of dominance, the gene frequencies, and/or if genes become fixed. You could ask what would happen to the mean if  $d = 0$ , if  $A_1$  was fixed, if  $d = a$ , or if the population had frequencies in equilibrium. The contribution of any locus to the population mean has one term for homozygotes and another term for heterozygotes. The formula assumes that the combination of loci produces a joint *additive* effect on the trait. The ‘*additive action*’ is, therefore, associated not only with alleles at one locus but also with alleles at different loci. Alleles at a locus have additive action in the absence of dominance and across loci if epistatic deviations are not present. Since environmental effects are taken as deviations from the general mean over the whole population, they add to zero, and then also expresses the mean phenotypic value.

2.3.1 Half-Sib Family Means

A half-sib family is obtained from seeds produced by one plant (female common parent) that was pollinated by a random sample of pollen from the population (Table 2.2).

**Table 2.2** Genotypic values and frequencies of half-sib families from a population in Hardy–Weinberg equilibrium for one locus with two alleles

Female parent	Frequency	Family genotypes <sup>a</sup>			Coded half-sib family values
		A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>	
A <sub>1</sub> A <sub>1</sub>	$p^2$	$p$	$q$	—	$pa + qd$
A <sub>1</sub> A <sub>2</sub>	$2pq$	$(\frac{1}{2})p$	$\frac{1}{2}$	$(\frac{1}{2})q$	$(\frac{1}{2})[(p-q)a + d]$
A <sub>2</sub> A <sub>2</sub>	$q^2$	—	$p$	$q$	$pd - qa$

<sup>a</sup>Produced after pollination by  $p(A_1)$  and  $q(A_2)$  male gametes.

The mean of the population of half-sib families is

$$\begin{aligned}\bar{X}_{\text{HS}} &= p^2 (pa - qd) + 2pq \left(\frac{1}{2}\right) [(p - q) a + \left(\frac{1}{2}\right) d] + q^2 (pq - qa) \\ &= (p - q)a + 2pqd\end{aligned}$$

This is equal to the original population mean.

2.3.2 Full-Sib Families

A full-sib family is obtained by crossing a random pair of plants (both parents in common) from the population. The probability of each cross is obtained by the product of genotypic frequencies, as shown in Table 2.3.

**Table 2.3** Genotypic values and frequencies of full-sib families from a population in Hardy–Weinberg equilibrium for one locus with two alleles

Female parent	Male parent	Probability of cross	Family genotypes			Coded full-sib family values
			A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>	
A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>1</sub>	$p^4$	1	—	—	$a$
	A <sub>1</sub> A <sub>2</sub>	$2p^3q$	$\frac{1}{2}$	$\frac{1}{2}$	—	$(\frac{1}{2})(a + d)$
	A <sub>2</sub> A <sub>2</sub>	$p^2q^2$	—	1	—	$d$
A <sub>1</sub> A <sub>2</sub>	A <sub>1</sub> A <sub>1</sub>	$2p^3q$	$\frac{1}{2}$	$\frac{1}{2}$	—	$(\frac{1}{2})(a + d)$
	A <sub>1</sub> A <sub>2</sub>	$4p^2q^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$(\frac{1}{2})d$
	A <sub>2</sub> A <sub>2</sub>	$2pq^3$	—	$\frac{1}{2}$	$\frac{1}{2}$	$(\frac{1}{2})(d - a)$
A <sub>2</sub> A <sub>2</sub>	A <sub>1</sub> A <sub>1</sub>	$p^2q^2$	—	1	—	$d$
	A <sub>1</sub> A <sub>2</sub>	$2pq^3$	—	$\frac{1}{2}$	$\frac{1}{2}$	$(\frac{1}{2})(d - a)$
	A <sub>2</sub> A <sub>2</sub>	$q^4$	—	—	1	$-a$

The mean of the population for full-sib families is

$$\begin{aligned}\bar{X}_{\text{FS}} &= p^4(a) + 2p^3q\left(\frac{1}{2}\right)\left[(a + b) + \cdots + q^4\right](-a) \\ &= (p - q)a + 2pqd\end{aligned}$$

Results so far obtained show that the expected value of half-sib families as well as of full-sib families equals the mean of the reference population.

2.3.3 Inbred families

Selfing is the most common system of inbreeding used in practical maize breeding for inbred line development during pedigree selection. Considering a non-inbred parent population in Hardy–Weinberg equilibrium from which selfed lines will be drawn, we have the family structure as shown in Table 2.4 for S<sub>1</sub> families, i.e., families developed by one generation of selfing. This assumes that the F<sub>2</sub> population equals an S<sub>0</sub> and we will follow this nomenclature throughout the book. (Note that there are maize breeding programs assuming an F<sub>2</sub> population equals an S<sub>1</sub>.)

**Table 2.4** Genotypic values and frequencies of inbred (S<sub>1</sub>) families from a non-inbred population in Hardy–Weinberg equilibrium for one locus with two alleles

Parent genotypes	Frequency	Family genotypes <sup>a</sup>			Coded S <sub>1</sub> family values
		A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>	
A <sub>1</sub> A <sub>1</sub>	$p^2$	1	—	—	$a$
A <sub>1</sub> A <sub>2</sub>	$2pq$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$(\frac{1}{2})d$
A <sub>2</sub> A <sub>2</sub>	$q^2$	—	—	1	$-a$

<sup>a</sup>After one generation of selfing

The mean of the population for inbred families is

$$\begin{aligned}\bar{X}_{S_1} &= p^2 a + 2pq \left(\frac{1}{2}\right) d + q^2 (-a) \\ &= (p - q)a + 2pqd\end{aligned}$$

which equals the reference population mean when  $d = 0$ , i.e., when there are no dominance effects. If dominance effects are present, the mean is reduced (see below). If the gene frequencies are the same for the reference and  $S_1$  populations, the mean of the  $S_1$  population will be halfway between the mean of the  $S_0$  and  $S_\infty$  generations.

Under a regular system of selfing the general mean decreases in each generation due to decreases in the frequency of heterozygotes. The general formula for the  $n$ th generation of inbreeding is

$$\bar{X}_{S_n} = (p - a)a + \left(\frac{1}{2}\right)^{n-1} pqd$$

which equals the non-inbred population mean for  $n = 0$ .

The above formula may also be expressed as a function of  $F_n$ , the coefficient of inbreeding, of progenies in the  $n$ th generation of selfing:

$$\bar{X}_{S_1} = (p - q)a + 2(1 - F_n)pqd$$

which equals the non-inbred population mean when  $F = 0$ .

Most reference or base populations (first segregating population,  $S_0$  in maize) is derived from elite (pure line)  $\times$  elite (pure line) crosses. Therefore, average gene frequency at all segregating loci is expected to be  $\frac{1}{2}$  and, therefore, we may assume that  $F_2$  populations can be represented with loci having gene frequencies in equilibrium ( $p = q = \frac{1}{2}$ ). Linkage, however, could be a serious bias. For example, most commercial breeding programs are represented within this scheme. However, the maize-breeding program at NDSU also develops inbred lines from genetically broad-based populations with arbitrary allele frequencies.

If we cross two inbred lines and consider one locus (two alleles A and a in this case):

Parents	AA	$\times$	aa
F <sub>1</sub>		Aa	$\otimes$

The  $S_0$  population is considered to be the base population:

F <sub>2</sub> ( $S_0$ )	$(\frac{1}{4})$ AA	$(\frac{1}{2})$ Aa	$(\frac{1}{4})$ aa	Base population or reference population (alleles segregating), e.g., 200 individuals
	$\downarrow$	$\downarrow$	$\downarrow$	
	$\otimes$ 50	$\otimes$ 100	$\otimes$ 50	

The  $S_1$  population is considered to be the result of one generation of self-fertilization:

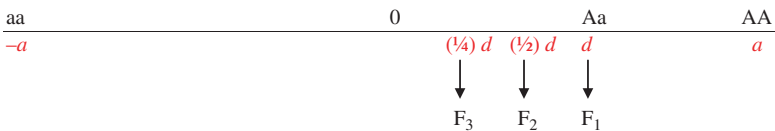
$$F_3 (S_1) \quad (\frac{1}{4}) AA \quad (\frac{1}{2}) \left[ (\frac{1}{4}) AA + (\frac{1}{2}) Aa + (\frac{1}{4}) aa \right] \quad (\frac{1}{4}) aa$$



So, if we calculate the mean for each generation we obtain

$$\begin{aligned}
 \bar{X}_{S_0} &= (1/4) AA + (1/2) Aa + (1/4) aa \\
 &= (1/4) a + (1/2) d - (1/4) a \\
 &= (1/2) d \\
 \bar{X}_{S_1} &= (1/4) AA + (1/2) [(1/4) AA + (1/2) Aa + (1/4) aa] + (1/4) aa \\
 &= (1/4) a + (1/2) [(1/2) d] - (1/4) a \\
 &= (1/4) d
 \end{aligned}$$

So, if we come back to our scale of genotypic values we see that the mean is reduced in the presence of dominance effects when selfing:



The examples of plant height and grain yield in maize seem to closely validate the theory.

## 2.4 Means of Inbred Populations and Derived Families

The main difference between inbred and non-inbred populations is in genotypic frequencies. Gene frequencies remain constant, but genotypic frequencies change under inbreeding because inbreeding decreases the frequency of heterozygotes and consequently increases the frequency of homozygous genotypes. Using Wright's coefficient  $F$  as a measure of inbreeding, genotypic frequencies are distributed according to the pattern shown in Table 2.5.

**Table 2.5** Genotypic values and frequencies in inbred populations (inbreeding measured by  $F$ ) for one locus with two alleles

Genotypes	Frequencies	Coded genotypic values
$A_1A_1$	$p^2 + Fpq$	$a$
$A_1A_2$	$2pq(1-F)$	$d$
$A_2A_2$	$q^2 + Fpq$	$-a$

Hence the mean of an inbred population is

$$\bar{X}_s = (p - q)a + 2pq(1 - F)d$$

When  $F = 1$  (completely homozygous population), then the inbred population mean equals  $(p - q)a$  because there will be no dominance effects expressed.

When  $F = 1/2$ , then the inbred population mean becomes  $(p - q)a + pqd$ , which equals the  $S_1$  family mean, as previously shown in Table 2.4.

Half-sib and full-sib families drawn from an inbred population result in non-inbred progenies, and their mean equals that of a non-inbred population  $(p - q)a + 2pqd$  because one generation of random mating is involved.

Kempthorne (1957) gives a general formulation for the changes in population mean under inbreeding, including epistatic effects. In his definition the lack of dominance and dominance types of epistasis do not change the population mean with inbreeding. If there are no dominance types of epistasis, the mean of the inbred population is linearly related to  $F$  even in the presence of additive types of epistasis.

## 2.5 Mean of a Cross Between Two Populations

Let  $P_1$  and  $P_2$  be two populations in Hardy–Weinberg equilibrium. Denoting by  $p$  and  $q$  the frequencies of both alleles,  $A_1$  and  $A_2$ , in population  $P_1$  and by  $r$  and  $s$  the frequencies of the same alleles in population  $P_2$ , we have the following structure in the crossed population (Table 2.6).

**Table 2.6** Genotypic values and frequencies in a cross between two populations in Hardy–Weinberg equilibrium for one locus with two alleles

Genotypes	Frequency	Coded genotypic values
$A_1A_1$	$pr$	$a$
$A_1A_2$	$ps + qr$	$d$
$A_2A_2$	$qs$	$-a$

The population cross mean for one locus is

$$\bar{X}_{12} = (pr - qs)a + (ps + qr)d$$

The cross between two populations also may be obtained according to a family structure. If half-sib families are drawn with, for example,  $P_1$  as female parents, we have the family structure shown in Table 2.7.

The mean of half-sib families is then

$$\begin{aligned}\bar{X}_{HS_{12}} &= p^2 (ra + sd) + 2pq \left[ \left( \frac{1}{2} \right) (r - s)a + \left( \frac{1}{2} \right) d \right] + q^2 (rd - sa) \\ &= (pr - qs)a + (ps + qr)d\end{aligned}$$

which equals the randomly crossed population mean. Note that the mean will be the same whatever population is used as the female parent.

If the crossed population is structured as full-sib families, we have the genotypes and frequencies shown in Table 2.8.

**Table 2.7** Genotypic values and frequencies in a cross between two populations structured as half-sib families for one locus with two alleles

Female parent, P <sub>1</sub>	Frequencies	Family genotypes <sup>a</sup>			Coded half-sib family values
		A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>	
A <sub>1</sub> A <sub>1</sub>	$p^2$	$r$	$s$	—	$ra + sd$
A <sub>1</sub> A <sub>2</sub>	$2pq$	$(\frac{1}{2})r$	$(\frac{1}{2})(r + s)$	$(\frac{1}{2})s$	$(\frac{1}{2})(r - s)a + (\frac{1}{2})d$
A <sub>2</sub> A <sub>2</sub>	$q^2$	—	$r$	$s$	$rd - sa$

<sup>a</sup>After pollination by  $r(A_1)$  and  $s(A_2)$  male gametes (from P<sub>2</sub>)

**Table 2.8** Genotypic values and frequencies in a cross between two populations structured as full-sib families for one locus with two alleles

Female parent, P <sub>1</sub>	Male parent, P <sub>2</sub>	Frequency of crosses	Family genotypes			Coded full-sib family values
			A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>	
A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>1</sub>	$p^2r^2$	1	0	0	$a$
	A <sub>1</sub> A <sub>2</sub>	$2p^2rs$	$\frac{1}{2}$	$\frac{1}{2}$	0	$(\frac{1}{2})(a + d)$
	A <sub>2</sub> A <sub>2</sub>	$p^2s^2$	0	1	0	$d$
A <sub>1</sub> A <sub>2</sub>	A <sub>1</sub> A <sub>1</sub>	$2pqr^2$	$\frac{1}{2}$	$\frac{1}{2}$	0	$(\frac{1}{2})(a + d)$
	A <sub>1</sub> A <sub>2</sub>	$4pqrs$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$(\frac{1}{2})d$
	A <sub>2</sub> A <sub>2</sub>	$2pqs^2$	0	$\frac{1}{2}$	$\frac{1}{2}$	$(\frac{1}{2})(d - a)$
A <sub>2</sub> A <sub>2</sub>	A <sub>1</sub> A <sub>1</sub>	$q^2r^2$	0	1	0	$d$
	A <sub>1</sub> A <sub>2</sub>	$2q^2rs$	0	$\frac{1}{2}$	$\frac{1}{2}$	$(\frac{1}{2})(d - a)$
	A <sub>2</sub> A <sub>2</sub>	$q^2s^2$	0	0	1	$-a$

The mean becomes

$$\begin{aligned}\bar{X}_{FS12} &= p^2r^2a + 2p^2rs\left(\frac{1}{2}\right)(a + d) + \cdots + q^2s^2(-a) \\ &= (pr - qs)a + (ps + qr)d\end{aligned}$$

which again equals the randomly crossed population mean and has the same value whatever parent is used as female.

2.6 Average Effect

It is a value not associated with genotypes but rather associated with the genes carried by the individual and transmitted to its offspring. The *average effect* of an allele is the mean deviation from the population mean of individuals (Table 2.9).

**Table 2.9** Genotypic values, frequencies, and average effect of alleles in a one locus model

Genotypes	Frequency ( $F_i$ )	Genotypic values ( $Y_i$ )	'A <sub>1</sub> ' gametes	'A <sub>2</sub> ' gametes
A <sub>1</sub> A <sub>1</sub>	$p^2$	$a$	$p$	0
A <sub>1</sub> A <sub>2</sub>	$2pq$	$d$	$q$	$p$
A <sub>2</sub> A <sub>2</sub>	$q^2$	$-a$	0	$q$

So, the average effect of 'A<sub>1</sub>' alleles ( $\alpha_1$ ) is

$$\begin{aligned}\alpha_1 &= pa + qd - \bar{X} \\ &= pa + qd - [(p - q)a + 2pqd] \\ \alpha_1 &= q[a + (q - p)d]\end{aligned}$$

and the average effect of 'A<sub>2</sub>' alleles ( $\alpha_2$ ) is

$$\begin{aligned}\alpha_2 &= pd + qa - \bar{X} \\ &= pd - qa - [(p - q)a + 2pqd] \\ \alpha_2 &= -p[a + (q - p)d]\end{aligned}$$

The concept of 'average effect' of a gene is basic to the understanding of breeding value. The average effect of a gene is defined as the mean deviation from the population mean of a group of individuals that received the gene from the same parent, the other gene of such individuals being randomly sampled from the whole population as shown in Table 2.10.

**Table 2.10** Genotypes and average effects of progenies having a common parental gamete for one locus (after Falconer, 1960)

Gamete	Genotypes in progenies <sup>a</sup>			Progeny effects	Population mean	Average effect of a gene
	A <sub>1</sub> A <sub>1</sub>	A <sub>1</sub> A <sub>2</sub>	A <sub>2</sub> A <sub>2</sub>			
A <sub>1</sub>	<i>p</i>	<i>q</i>	—	<i>pa + qd</i>	<i>(p - q)a + 2pqd</i>	$\alpha_1 = q[a + (q - p)d]$
A <sub>2</sub>	—	<i>p</i>	<i>q</i>	<i>pd - qa</i>	<i>(p - q)a + 2pqd</i>	$\alpha_2 = -p[a + (q - p)d]$

<sup>a</sup>After pollination with a random sample of gametes: *p*(A<sub>1</sub>) and *q*(A<sub>2</sub>)

If two alleles are present per locus we can define the *average effect of gene substitution* ( $\alpha$ ) as the difference between the average effects of the two alleles:

$$\begin{aligned}\alpha &= \alpha_1 - \alpha_2 \\ &= p + q[a + (q - p)d] \\ \alpha &= a + (q - p)d\end{aligned}$$

The 'average effect of a gene substitution' is the average deviation due to the substitution of one gene by its allele in each genotype. Consider the A<sub>2</sub> gene being substituted by its gene A<sub>1</sub> at random in the population. Since the A<sub>1</sub>A<sub>1</sub>, A<sub>1</sub>A<sub>2</sub>, and A<sub>2</sub>A<sub>2</sub> genotypes have frequencies *p*<sup>2</sup>, 2*pq*, and *q*<sup>2</sup>, respectively, then genes that will be substituted are found in genotypes A<sub>1</sub>A<sub>2</sub> and A<sub>2</sub>A<sub>2</sub> with frequency *pq* + *q*<sup>2</sup> = *q*. Proportionally, we have *pq*/*q* = *p* A<sub>1</sub>A<sub>2</sub> genotypes: *q*<sup>2</sup>/*q* = *q* A<sub>2</sub>A<sub>2</sub> genotypes; i.e., A<sub>2</sub> genes will be substituted in A<sub>1</sub>A<sub>2</sub> and A<sub>2</sub>A<sub>2</sub> genotypes at frequencies *p* and *q*,

respectively. When the substitution is in  $A_1A_2$  genotypes, the change in genotypic value will be from  $d$  to  $a$ , and when substitution takes place in  $A_2A_2$ , the change is from  $-a$  to  $d$ . The change in the population is

$$\begin{aligned}\alpha &= p(a - d) + q(d + a) \\ &= a + (q - p)d\end{aligned}$$

This is the definition of the average effect of a gene substitution. It can be seen that the average effect of a gene substitution  $\alpha$  is the difference between the average effects of genes involved in the substitution; i.e.,  $\alpha = \alpha_1 - \alpha_2$ , as shown in Table 2.10. Both the average effect of a gene and the average effect of a gene substitution depend on gene effects and gene frequency; therefore, both are a property of the population and of the gene.

2.7 Breeding Value

When panmictic populations are under consideration, one must consider that the genotypes of any offspring are not identical to their parents. The relationship between any individual in the offspring and one of its parents is established by the gamete received from that parent. It is known that gametes are haploid entities and carry genes and not genotypes. So for the understanding of the inheritance of a quantitative trait in a panmictic population it is valuable to have an individual measure associated with its genes and not its genotype. Such a value is designated by Falconer (1960) as ‘breeding value,’ which is ‘the value of an individual, judged by the mean value of its progeny.’

The breeding value of an individual can, therefore, be measured. It is twice the mean deviation of the progeny from the population mean since only one half of the genes are passed to the progeny.

The breeding value of an individual is equal to the sum of average effects of the genes it carries (Table 2.11). If all loci are taken into account, the breeding value of a particular genotype is the sum of breeding values from each locus (‘additive genotype’).

Extending the concept of average effect of genes to the individual genotype gives the concept of breeding value of the individual. At the gene level the breeding value is the sum of average effects of genes, summation being over all alleles and over all loci. Similar to the average effect of a gene, breeding value is a property of the

Table 2.11 Genotypic values, frequencies, and breeding values in a one locus model

Genotypes	Frequency	Genotypic values	Breeding value
$A_1A_1$	$p^2$	$a$	$2 \alpha_1$
$A_1A_2$	$2pq$	$d$	$\alpha_1 + \alpha_2$
$A_2A_2$	$q^2$	$-a$	$2 \alpha_2$

individual as well as of the population but the breeding value can be measured experimentally. The breeding value is, therefore, a measurable quantity and is of much relevance in animal breeding where the individual value is an important criterion. On the other hand, individual values are less important in crop species like maize, since the whole population is concerned. In this case individuals are looked upon as ephemeral representatives of the whole population and its gene pool. The average effect of a gene and individual breeding value concepts, however, are closely related to genotype evaluation procedures like topcross tests in maize.

## 2.8 Genetic Variance

Breeders choose not only populations with high phenotypic means but also populations having large and useful genetic variance.

The variation among phenotypic values (phenotypic variance) can be partitioned into observational components of variance:

$$\hat{\sigma}_P^2 = \hat{\sigma}_G^2 + \hat{\sigma}_E^2 + \hat{\sigma}_{GE}^2$$

Even though it is the goal of breeders to separate the genetic variance ( $\hat{\sigma}_G^2$ ) from the environmental variance ( $\hat{\sigma}_E^2$ ), the variance due to crossover (e.g., rank) and non-crossover (e.g., magnitude) interactions between genotypes and environments ( $\hat{\sigma}_{GE}^2$ ) is the most difficult to manage.

Fisher (1918) first demonstrated that the hereditary variance in a random mating population can be partitioned into three parts: (1) an additive portion associated with average effects of genes, (2) a dominance portion due to allelic interactions, and (3) a portion due to non-allelic interactions or epistatic effects. Therefore, the genetic proportion of variance has the following components:

$$\hat{\sigma}_G^2 = \hat{\sigma}_A^2 + \hat{\sigma}_D^2 + \hat{\sigma}_I^2$$

*Epistatic* interactions give rise to the component of variance  $\hat{\sigma}_I^2$ , which is the variance due to the interaction deviations involving more than one locus. This is subdivided into components according to the number of loci involved (e.g., two-factor interaction, three-factor interaction). Another subdivision can be done based upon the type of interaction present. If the interaction involves breeding values then the additive  $\times$  additive interaction variance is present ( $\hat{\sigma}_{AA}^2$ ). If the interaction is between the breeding value of one locus and the dominance deviation of the other then the additive  $\times$  dominance interaction variance is present ( $\hat{\sigma}_{AD}^2$ ). Finally, if the interaction is between dominance deviations from two loci then the dominance  $\times$  dominance interaction variance is present ( $\hat{\sigma}_{DD}^2$ ).

A general theory for the partition of hereditary variance was further developed by Cockerham (1954) and Kempthorne (1954). Thus, in general, the total genetic variance  $\hat{\sigma}_G^2$  can be partitioned into the following components:

- $\hat{\sigma}_A^2$  additive variance due to the average effects of alleles (additive effects, same locus)
- $\hat{\sigma}_D^2$  dominance variance due to interaction of average effects of alleles (dominance effects, same locus)
- $\hat{\sigma}_{AA}^2, \hat{\sigma}_{AAA}^2, \dots$  = epistatic variances due to interaction of additive effects of two or more loci
- $\hat{\sigma}_{DD}^2, \hat{\sigma}_{DDD}^2, \dots$  = epistatic variances due to interaction of dominance effects of two or more loci
- $\hat{\sigma}_{AD}^2, \hat{\sigma}_{AAD}^2, \hat{\sigma}_{ADD}^2, \dots$  = epistatic variances due to interaction of additive and dominance effects involving two or more loci

Collecting all components together, the total genetic variance is

$$\hat{\sigma}_G^2 = \hat{\sigma}_A^2 + \hat{\sigma}_D^2 + \hat{\sigma}_{AA}^2 + \hat{\sigma}_{DD}^2 + \hat{\sigma}_{AD}^2 + \hat{\sigma}_{AAA}^2 + \hat{\sigma}_{AAD}^2 + \dots$$

The interaction between loci (located within or between chromosomes) controlling the expression of quantitative traits is assumed to be frequent. However, the estimation of the amount of variance generated by interactions is challenging even at the molecular level.

An additional source of genetic variance is the one due to *disequilibrium*. In this case, genotypic frequencies at several loci cannot be predicted by allele frequencies. If there is no epistasis between two loci we can estimate the total genotypic variance caused by the two loci together as follows:

$$\hat{\sigma}_{TG}^2 = \hat{\sigma}_G^2 (\text{first locus}) + \hat{\sigma}_D^2 (\text{second locus}) + 2\widehat{\text{Cov}} (\text{both loci})$$

The covariance term is the correlation between the genotypic values at the two loci in different individuals. This correlation can be positive or negative. Therefore, linkage disequilibrium can either decrease or increase the variance depending on the linkage phase present. Coupling phase linkage will cause an upward bias for the additive and dominance genetic variances. On the other hand, repulsion phase linkage will only cause the dominance genetic variance to increase; the additive genetic variance is expected to decrease. No covariance term is present if there is random mating equilibrium.

Linkage of traits with molecular markers became a popular scientific research targeted initially at improvement of quantitative traits. But quantitative traits are dependent upon a large number of genes each having a relatively minor effect as compared with environmental effects (Lonnquist, 1963). Based on this definition, quantitative traits have been explained by polygenes (Mather, 1941) and quantitative trait loci (QTL) (Geldermann, 1975) or chromosome segments affecting the quantitative trait (Falconer and Mackay, 1996). Rather than QTL mapping in bi-parental populations breeding plans that currently utilize molecular marker information for germplasm (e.g., association mapping on relevant breeding germplasm, genome-wide selection) could be assessed depending on the amount of linkage between markers and loci and may generate useful information that is relevant to improving elite germplasm

(Sorrells, 2008). All these approaches, however, rely on the maintenance of strong applied breeding programs and need to be proven useful for developing cultivars in a more efficient way. Alternative approaches such as ‘meta-QTL analysis’ focused on major QTLs that are stable across numerous populations also have potential (Snape et al., 2008).

### 2.8.1 Total Genetic Variance

Total genetic variance of a population in Hardy–Weinberg equilibrium is obtained from a modified version of Table 2.1 as follows:

Genotypes	Frequency ( $F_i$ )	Genotypic values (GV)	$F_i \times \text{GV}$
$A_1A_1$	$p^2$	$a$	$p^2a$
$A_1A_2$	$2pq$	$d$	$2pqd$
$A_2A_2$	$q^2$	$-a$	$q^2(-a)$

We could, therefore, estimate the variance statistically and use this information to estimate the genetic variance of a population:

$$\hat{\sigma}_G^2 = \sum F_i X_i^2 - \sum (F_i X_i)^2 \text{ or } \sum X_i^2 - \left( \frac{(\sum X_i)^2}{n} \right) \text{ or } \sum (X_i - \bar{X})^2$$

In this case, the mean is the corrector factor in the working formula. Therefore,

$$\hat{\sigma}_G^2 = [p^2a^2 + 2pqd^2 + q^2(-a)^2] - \bar{X}^2 \quad \text{and} \quad \bar{X}^2 = [(p-q)^2a^2 + 4pq(p-q)ad + 4p^2q^2d^2]$$

Then, the total genetic variance for one locus is

$$\begin{aligned} \hat{\sigma}_G^2 &= \cancel{p^2a^2} + 2pqd^2 + \cancel{q^2a^2} - \cancel{p^2a^2} + 2pqa^2 - \cancel{q^2a^2} - 4pq(p-q)ad - 4p^2q^2d^2 \\ \hat{\sigma}_G^2 &= 2pqd^2 + 2pqa^2 - apq(p-q)ad - 4p^2q^2d^2 \\ \hat{\sigma}_G^2 &= 2pq[a^2 + b^2 - 2(p-q)ad - 2pqd^2] \end{aligned}$$

Then, the total genetic variance for one locus is

$$\boxed{\hat{\sigma}_G^2 = 2pq[a^2 + 2(q-p)ad + (1-2pq)d^2]}$$

And if we have a population with frequencies in equilibrium ( $p = \frac{1}{2}$  and  $q = \frac{1}{2}$ ) then (e.g.,  $F_2$  populations)



$$\hat{\sigma}_G^2 = (1/2) a^2 + (1/4) d^2$$

The first term of the formula is the most important for breeders while the second term is the one breeders are not able to fix.

### 2.8.2 Additive Genetic Variance

The additive genetic variance is obtained as follows:

Genotypes	Frequency ( $F_i$ )	Breeding values
$A_1A_1$	$p^2$	$2\alpha_1 = 2q\alpha$
$A_1A_2$	$2pq$	$\alpha_1 + \alpha_2 = (q-p)\alpha$
$A_2A_2$	$q^2$	$2\alpha_2 = -2p\alpha$

The additive genetic variance ( $\hat{\sigma}_A^2$ ) is, therefore derived as follows:

$$\begin{aligned}\hat{\sigma}_A^2 &= p^2 (2q\alpha)^2 + 2pq [(q-p)\alpha]^2 + q^2 [(-2p\alpha)]^2 \\ \hat{\sigma}_A^2 &= 2pq\alpha^2 [2pq + q^2 + p^2]\end{aligned}$$

Then, the additive genetic variance for one locus is

$$\hat{\sigma}_A^2 = 2pq\alpha^2 \quad \text{or} \quad 2pq [a + (q-p)d]^2$$

Clearly,  $\hat{\sigma}_A^2$  depends on gene frequencies. Therefore, for segregating alleles in equilibrium (e.g.,  $p = q = 0.5$ ) then

$$\begin{aligned}\hat{\sigma}_A^2 &= 2pq [a + (q-p)d]^2 \\ &= 2pqa^2 \\ \hat{\sigma}_A^2 &= (1/2) a^2\end{aligned}$$

This is the additive genetic variance for the special case of  $F_2$  populations.

In the general case of arbitrary gene frequencies (e.g., genetically broad-based populations) the following formula applies:

$$\hat{\sigma}_{S_0}^2 = \hat{\sigma}_A^2 \quad \text{Note the } \hat{\sigma}_A^2 \text{ has a level of dominance between alleles in the additive portion of this segregating population}$$

### 2.8.3 Dominance Genetic Deviations

The dominance variance is the remainder from the total variance and is calculated by subtraction as

$$\begin{aligned}\hat{\sigma}_D^2 &= \hat{\sigma}_G^2 - \hat{\sigma}_A^2 \\ &= 2pq [a^2 + 2(q-p)ad + (1-2pq)d^2] - 2pq [a + (q-p)d]^2 \\ &= 2pqd^2 (2pq)\end{aligned}$$

Then, the dominance genetic variance for one locus is

$$\hat{\sigma}_D^2 = 4p^2q^2d^2$$

$\hat{\sigma}_D^2$  also depends on gene frequencies. Therefore, for segregating alleles in equilibrium (e.g.,  $p = q = 0.5$ ) then

$$\hat{\sigma}_D^2 = 4(1/4)(1/4)d^2$$

$$\hat{\sigma}_D^2 = (1/4)d^2$$

This is the dominance genetic variance for the special case of  $F_2$  populations.

In the general case of arbitrary gene frequencies (e.g., genetically broad-based populations) the following formula applies:

$$\hat{\sigma}_{S_0}^2 = \hat{\sigma}_D^2$$

Note the  $\hat{\sigma}_D^2$  does not have additive effects in the dominance portion of this segregating population

Both additive and dominance genetic variances ( $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$ ) can also be explained by regression analyses (Falconer and Mackay, 1996).  $\hat{\sigma}_A^2$  is defined as the variance due to linear regression of genotypic values on gene frequencies in individual genotypes explained by the sum squares of the regression ANOVA while  $\hat{\sigma}_D^2$  represents the variance due to deviations from the regression. Hence the dominance variance is the deviation from the regression of genotypic values on the gene content of the genotypes.

Therefore, if we look at the example for the additive genetic variance we obtain

$$\begin{aligned}\hat{\sigma}_A^2 &= [\widehat{\text{Cov}}]^2 / \widehat{\text{Var}} \\ &= \{2pq [a + (q-p)d]\}^2 / 2pq\end{aligned}$$

$$\hat{\sigma}_A^2 = 2pq [a + (q-p)d]^2$$

**Table 2.12** Gene frequency ( $p$ ) at maximum values for  $\hat{\sigma}_G^2$ ,  $\hat{\sigma}_A^2$ , and  $\hat{\sigma}_D^2$  for no dominance and complete dominance for one locus with two alleles

Gene action	$\hat{\sigma}_G^2$	$\hat{\sigma}_A^2$	$\hat{\sigma}_D^2$
No dominance	$1/2$	$1/2$	—
Complete dominance	$1 - \sqrt{1/2}$	$1/4$	$1/2$

Gene frequencies, which give the maximum values for  $\hat{\sigma}_G^2$ ,  $\hat{\sigma}_A^2$ , and  $\hat{\sigma}_D^2$ , are found by taking the first derivative of their respective expressions equal to zero. The simplest cases are those for no dominance and complete dominance, as shown in Table 2.12.

### 2.8.4 Variance Among Non-inbred Families

Genetic variance among half-sib families from a population in Hardy–Weinberg equilibrium is obtained from Table 2.2 as follows:

$$\begin{aligned}\hat{\sigma}_{\text{HS}}^2 &= \sum F_i X_i^2 - \bar{X}_{\text{HS}} \\ &= (1/2) pq [a + (1 - 2p)d]^2\end{aligned}$$

Since,

$$\hat{\sigma}_A^2 = 2pq [a + (q - p)d]^2$$

then,

$$\boxed{\hat{\sigma}_{\text{HS}}^2 = (1/4)\hat{\sigma}_A^2}$$

Theoretically, the total genetic variance (among and within half-sib families) equals the total genetic variance in the reference population, i.e.,  $\hat{\sigma}_G^2 = \hat{\sigma}_A^2 + \hat{\sigma}_D^2$ . From this total variance  $(1/4)\hat{\sigma}_A^2$  is expressed *among* half-sib families.

The remainder,  $(3/4)\hat{\sigma}_A^2 + \hat{\sigma}_D^2$ , is expected to be present *within* half-sib families over the entire population of families.

Genetic variance among full-sib families is obtained from Table 2.3 as

$$\begin{aligned}\hat{\sigma}_{\text{FS}}^2 &= \sum F_i X_i^2 - \bar{X}_{\text{FS}} \\ &= pq [a + (q - p)d]^2 + p^2 q^2 d^2\end{aligned}$$

Since,

$$\hat{\sigma}_D^2 = 4p^2 q^2 d^2$$

then,

$$\hat{\sigma}_{\text{FS}}^2 = (\frac{1}{2})\hat{\sigma}_{\text{A}}^2 + (\frac{1}{4})\hat{\sigma}_{\text{D}}^2$$

Relative to the total genetic variance of a population it is seen that  $\hat{\sigma}_{\text{FS}}^2 = (\frac{1}{2})\hat{\sigma}_{\text{A}}^2 + (\frac{1}{4})\hat{\sigma}_{\text{D}}^2$ . This is the portion of the total genetic variance expressed *among* families. Thus the remainder of the total genetic variance,  $(\frac{1}{2})\hat{\sigma}_{\text{A}}^2 + (\frac{3}{4})\hat{\sigma}_{\text{D}}^2$ , is expected to be present *within* families over the whole population.

If epistasis is considered, the approximation to the real genetic variance is  $\hat{\sigma}_{\text{G}}^2 = \hat{\sigma}_{\text{A}}^2 + \hat{\sigma}_{\text{D}}^2 + \hat{\sigma}_{\text{AA}}^2 + \hat{\sigma}_{\text{DD}}^2 + \hat{\sigma}_{\text{AD}}^2 + \hat{\sigma}_{\text{AAA}}^2 + \hat{\sigma}_{\text{AAD}}^2 + \dots$ , where all components were defined previously.

In the same way,  $\hat{\sigma}_{\text{HS}}^2 = (\frac{1}{4})\hat{\sigma}_{\text{A}}^2 + (\frac{1}{16})\hat{\sigma}_{\text{AA}}^2 + (\frac{1}{64})\hat{\sigma}_{\text{AAA}}^2 + \dots$ , or simply  $\hat{\sigma}_{\text{HS}}^2 = (\frac{1}{4})\hat{\sigma}_{\text{A}}^2 + \dots$ , indicating a bias due to epistatic components of variance. The variance among full-sib families is  $\hat{\sigma}_{\text{FS}}^2 = (\frac{1}{2})\hat{\sigma}_{\text{A}}^2 + (\frac{1}{4})\hat{\sigma}_{\text{D}}^2 + (\frac{1}{4})\hat{\sigma}_{\text{AA}}^2 + (\frac{1}{16})\hat{\sigma}_{\text{DD}}^2 + \dots$ , or simply  $\hat{\sigma}_{\text{FS}}^2 = (\frac{1}{2})\hat{\sigma}_{\text{A}}^2 + (\frac{1}{4})\hat{\sigma}_{\text{D}}^2 + \dots$ , indicating a bias due to epistatic components.

The relation between the variance among families and the covariance between relatives within families are presented in Chapter 3.

### 2.8.5 Variance Among Inbred Families

Variance among  $S_1$  families from a non-inbred reference population is obtained from Table 2.4 as

$$\begin{aligned}\hat{\sigma}_{S_1}^2 &= \sum F_i X_i^2 - \bar{X}_{S_1}^2 \\ &= 2pq \left[ a + \left(\frac{1}{2}\right)(q - p)d \right]^2 + p^2 q^2 d^2\end{aligned}$$

A problem that arises with inbreeding is that genetic variance *among* inbred families is not linearly related to genetic components of variance of the reference population. From the above expression it can be seen that  $\hat{\sigma}_{S_1}^2$  can be translated into  $\hat{\sigma}_{\text{A}}^2$  and/or  $\hat{\sigma}_{\text{D}}^2$  only in the following situations:

$$\begin{aligned}\text{No dominance } (d_i = 0 \text{ for all loci): } & \hat{\sigma}_{S_1}^2 = \hat{\sigma}_{\text{A}}^2 \\ \text{Gene frequency } \frac{1}{2} \text{ for all loci: } & \hat{\sigma}_{S_1}^2 = \hat{\sigma}_{\text{A}}^2 + \left(\frac{1}{4}\right)\hat{\sigma}_{\text{D}}^2\end{aligned}$$

If the restriction of no dominance is imposed, it can also be demonstrated that the genetic variance among  $S_2$  families (after two generations of selfing) is  $\hat{\sigma}_{S_2}^2 = (\frac{3}{2})\hat{\sigma}_{\text{A}}^2$ . In the same way, if gene frequencies are assumed to be  $\frac{1}{2}$ , then  $\hat{\sigma}_{S_2}^2 = (\frac{3}{2})\hat{\sigma}_{\text{A}}^2 + (\frac{3}{16})\hat{\sigma}_{\text{D}}^2$ .

An alternative way to understand what was explained above is by emphasizing there are two types of variances: among and within progenies. The one among progenies does not generate additive values from heterozygote plants since the variation

among  $S_1$  plants within progenies is not included. In addition, the variance within parenthesis resembles to the variance among  $F_2$  individuals.

If the genetic variance for an  $F_2$  generation is

$$\begin{aligned}\hat{\sigma}_{F_2}^2 &= \sum g_i^2 - (\bar{X})^2 \\ &= \left[ \left( \frac{1}{4} \right) a^2 + \left( \frac{1}{2} \right) d^2 - \left( \frac{1}{4} \right) a^2 \right] - \left[ \left( \frac{1}{2} \right) d \right]^2 \\ &= \left( \frac{1}{2} \right) a^2 + \left( \frac{1}{2} \right) d^2 - \left( \frac{1}{4} \right) d^2 \\ &= \left( \frac{1}{2} \right) a^2 + \left( \frac{1}{4} \right) d^2\end{aligned}$$

$\hat{\sigma}_{F_2}^2 = \hat{\sigma}_A^2 + \hat{\sigma}_D^2$	Variance among individuals (individual plant basis)
--	---

The variance among  $S_1$  families follows:

$$\begin{aligned}\hat{\sigma}^2 \bar{s}_1 &= \left[ \left( \frac{1}{4} \right) a^2 + \left( \frac{1}{2} \right) (0 + \left( \frac{1}{4} \right) d^2 + 0) + \left( \frac{1}{4} \right) a^2 \right] - \left[ \left( \frac{1}{4} \right) d \right]^2 \\ &= \left[ \left( \frac{1}{2} \right) a^2 + \left( \frac{1}{2} \right) \left( \frac{1}{4} \right) d^2 \right] - \left( \frac{1}{16} \right) d^2 \\ &= \left( \frac{1}{2} \right) a^2 + \left( \frac{1}{16} \right) d^2\end{aligned}$$

$\hat{\sigma}^2 \bar{s}_1 = \hat{\sigma}_A^2 + \left( \frac{1}{4} \right) \hat{\sigma}_D^2$
---

While the variance within  $S_1$  families is

$$\begin{aligned}\hat{\sigma}^2 s_{1w} &= [0 + \left( \frac{1}{2} \right) \left( \frac{1}{4} \right) a^2 + \left( \frac{1}{4} \right) d^2 + 0] \\ &= \left( \frac{1}{4} \right) a^2 + \left( \frac{1}{8} \right) d^2\end{aligned}$$

$\hat{\sigma}^2 s_{1w} = \left( \frac{1}{2} \right) \hat{\sigma}_A^2 + \left( \frac{1}{2} \right) \hat{\sigma}_D^2$
---

More details can be found in Section 4.12.

### 2.8.5.1 Distribution of Genetic Variances Among and Within Lines

If we continue selfing from the  $F_3(S_1)$  generation to the  $F_8(S_6)$  generation of inbreeding the distribution of variances among and within lines changes depending on the inbreeding coefficient  $[F = 1 - (\frac{1}{2})^n]$  being 'n' the number of generations of continuous selfing (Table 2.13).

Assuming  $p = q = \frac{1}{2}$  then at the  $S_6$  level of inbreeding ( $F \sim 1$ )  $\hat{\sigma}_G^2 = 2 \hat{\sigma}_A^2$  among lines which have important breeding implications. As the inbreeding coefficient approaches unity the additive genetic variance among lines approaches twice the additive genetic variance of the reference population without inbreeding. Therefore,

**Table 2.13** Distribution of variances among and within lines under continuous selfing assuming  $p = q = 0.5$  and  $F = 1 - (1/2)^n$

Generation	$F$	Among lines		Within lines		Total	
		$\hat{\sigma}_A^2$	$\hat{\sigma}_D^2$	$\hat{\sigma}_A^2$	$\hat{\sigma}_D^2$	$\hat{\sigma}_A^2$	$\hat{\sigma}_D^2$
S <sub>1</sub>	$1/2$	1	$1/4$	$1/2$	$1/2$	$3/2$	$3/4$
S <sub>2</sub>	$3/4$	$3/2$	$3/16$	$1/4$	$1/4$	$7/4$	$7/16$
S <sub>3</sub>	$7/8$	$7/4$	$7/64$	$1/8$	$1/8$	$15/8$	$15/64$
S <sub>4</sub>	$15/16$	$15/8$	$15/256$	$1/16$	$1/16$	$31/16$	$31/256$
S <sub>5</sub>	$31/32$	$31/16$	$31/1024$	$1/32$	$1/32$	$63/32$	$63/1024$
S <sub>6</sub>	$63/64$	$63/32$	$63/4096$	$1/64$	$1/64$	$127/64$	$127/4096$
$\infty$	1	2	0	0	0	2	0

differences among lines increases and breeders will be more effective in identifying the best lines than to identify the best parents among plants that are not inbred.

A summary of the distribution of  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$  among and within lines for successive generations of inbreeding is given in Table 2.13. Two important features are that (1) the genetic variance among inbred families increases and within inbred families decreases with increased inbreeding and (2) the total genetic variance doubles from  $F = 0$  to  $F = 1$ , and all the genetic variance at  $F = 1$  is additive. General limitations when inbreeding is involved are discussed in Chapter 3.

### 2.8.6 Variance in Inbred Populations

The effect of inbreeding is to increase total genetic variance in the population, and such an increase depends on the level of inbreeding. Being  $u_s$  an alternative symbol for the mean of the inbred population the total genetic variance is calculated from Table 2.5 as follows:

$$\begin{aligned}\hat{\sigma}_{G_s}^2 &= (p^2 + Fpq)a^2 + 2pq(1 - F)d^2 + (q^2 + Fpq)a^2 - u_s^2 \\ &= 2pq(1 + F) \left[ a + \frac{1-F}{1+F}(q - p)d \right]^2 + 4pq \frac{1-F}{1+F}(p + Fq)(q + Fp)d^2\end{aligned}$$

which equals the genetic variance of a non-inbred population when  $F = 0$ . The additive genetic variance is

$$\hat{\sigma}_{A_s}^2 = 2pq(1 + F) \left[ a + \frac{1 - F}{1 + F}(q - p)d \right]^2$$

and the dominance variance is again calculated as  $\hat{\sigma}_{G_s}^2 - \hat{\sigma}_{A_s}^2$ :

$$\hat{\sigma}_{D_s}^2 = 4pq \frac{1 - F}{1 + F}(p + Fq)(q + Fp)d^2$$

When  $F = 0$ ,  $\hat{\sigma}_{A_S}^2$  and  $\hat{\sigma}_{D_S}^2 = \hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$ , respectively. For  $F > 0$ ,  $\hat{\sigma}_{A_S}^2$  and  $\hat{\sigma}_{D_S}^2$  cannot be expressed by  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$  nor be translated from one generation of inbreeding to another. If gene frequency is assumed to be  $1/2$ , then  $\hat{\sigma}_{A_S}^2 = (1 + F)\hat{\sigma}_A^2$  and  $\hat{\sigma}_{D_S}^2 = (1 - F^2)\hat{\sigma}_D^2$ . The first expression is also valid when only additive effects are considered and  $\hat{\sigma}_{G_S}^2 = (1 + F)\hat{\sigma}_A^2$ .

When half-sib and full-sib families are drawn from an inbred population, the families are themselves non-inbreds and the variance among families is expressed by

$$\hat{\sigma}_F^2 = \theta_1 \hat{\sigma}_A^2 + \theta_2 \hat{\sigma}_D^2 + \theta_1^2 \hat{\sigma}_{AA}^2 + \theta_2^2 \hat{\sigma}_{DD}^2 + \theta_1 \theta_2 \hat{\sigma}_{AD}^2 + \theta_1^3 \hat{\sigma}_{AAA}^2 + \dots$$

where  $\theta_1 = (1 + F)/4$  and  $\theta_2 = 0$  for half-sib families and  $\theta_1 = (1 + F)/2$  and  $\theta_2 = (1 + F)^2/4$  for full-sib families, according to adaptation from Cockerham (1963). In Chapter 3 these covariance terms are expressed in terms of covariance between relatives.

Under continuous selfing, assuming only additive effects, the total genetic variance is partitioned into among lines and within lines as follows:

Among lines	$2F\hat{\sigma}_G^2$
Within lines	$(1 - F)\hat{\sigma}_G^2$
Total	$(1 + F)\hat{\sigma}_G^2$

where  $\hat{\sigma}_G^2$  is the total genetic variance in a random mating non-inbred population. Thus when inbreeding is complete ( $F = 1$ ), the genetic variance among lines is twice the genetic variance of the reference population (non-inbred) and no genetic variation is expected within lines. At complete inbreeding ( $F = 1$ ), the additive model is completely valid (since there is no dominance) and is a good approximation for  $F$  slightly less than 1 (highly inbred lines).

### 2.8.7 Variance in a Cross Between Two Populations

The first cross between two distinct populations is not in equilibrium, and its total genetic variance is not linearly related to any of the parent populations. However, total genetic variance can be partitioned into additive and dominance components as follows:

$$\begin{aligned}\hat{\sigma}_{A_{(12)}}^2 &= pq[a + (s - r)d]^2 + rs[a + (q - p)d]^2 = 1/2(\hat{\sigma}_{A_{12}}^2 + \hat{\sigma}_{A_{21}}^2) \\ \hat{\sigma}_{D_{12}}^2 &= 4p(1 - p)r(1 - r)d^2 = 4pqrsd^2\end{aligned}$$

according to the notation used in Table 2.6 (Compton et al., 1965).

The two components may be called *homologues* of additive and dominance variances as defined for one population, i.e.,  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$ . In fact, they equal  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$  when  $p = r$ ; i.e., both populations have exactly the same gene frequency. In the above notation we used  $\hat{\sigma}_{A(12)}^2$  to denote the total additive genetic variance and either  $\hat{\sigma}_{A_{12}}^2$  or  $\hat{\sigma}_{A_{21}}^2$  to denote the subcomponents when either  $P_1$  or  $P_2$ , respectively, is used as the female parent.

When the crossed population is in a half-sib family structure, the variance among families is obtained from Table 2.7 as follows:

$$\begin{aligned}\hat{\sigma}_{HS_{12}}^2 &= p^2(ra + sd)^2 + \cdots + q^2(rd - sa)^2 - u_{HS_{12}}^2 = (\frac{1}{2})pq[a + (s - r)d]^2 \\ &= (\frac{1}{4})\hat{\sigma}_{A_{12}}^2\end{aligned}$$

and  $u_{HS}$  represents the mean of the half-sib population.

If population  $P_2$  is used as the female parent

$$\hat{\sigma}_{HS_{21}}^2 = (\frac{1}{2})rs[a + (q - p)d]^2 = (\frac{1}{4})\hat{\sigma}_{A_{21}}^2$$

If both types of families are drawn, their mean variance is  $(\frac{1}{4})\hat{\sigma}_{A_{12}}^2 + (\frac{1}{4})\hat{\sigma}_{A_{21}}^2$ ; for  $p = r$  this equals  $(\frac{1}{2})[(\frac{1}{4})\hat{\sigma}_A^2 + (\frac{1}{4})\hat{\sigma}_A^2] = (\frac{1}{4})\hat{\sigma}_A^2$ , which is the variance among half-sib families for one population.

In the same way, genetic variance among full-sib families is obtained from Table 2.8 as follows:

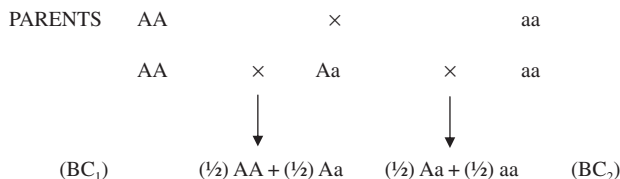
$$\begin{aligned}\hat{\sigma}_{FS(12)}^2 &= p^2r^2a^2 + \cdots + q^2s^2a^2 - u_{FS(12)}^2 \\ &= (\frac{1}{2})pq[a + (1 - 2r)d]^2 + (\frac{1}{2})rs[a + (1 - 2p)d]^2 + pqrds^2 \\ &= (\frac{1}{2})\hat{\sigma}_{A_{12}}^2 + (\frac{1}{4})\hat{\sigma}_{D_{12}}^2\end{aligned}$$

and  $u_{FS}$  represents the mean of the full-sib population.

Note that this result will be the same whatever parental population is used as female parent. When  $p = r$ , then  $\hat{\sigma}_{FS}^2 = (\frac{1}{2})\hat{\sigma}_A^2 + (\frac{1}{4})\hat{\sigma}_D^2$  the variance among full-sib families as previously demonstrated for one population.

## 2.9 Means and Variances in Backcross Populations

The same principles apply to backcross populations (see Section 4.12 for more details). Assuming gene frequencies in equilibrium we can describe these populations as follows (two alleles, A and a):





Then,

$$\begin{aligned}
 \bar{X}_{BC_1} &= (1/2)a + (1/2)d \\
 \hat{\sigma}_{BC_1}^2 &= (1/2)a^2 + (1/2)d^2 - ((1/2)a + (1/2)d)^2 \\
 &= (1/2)a^2 + (1/2)d^2 - [(1/4)a^2 + 2(1/2)(1/2)ad + (1/4)d^2] \\
 &= (1/4)a^2 + (1/4)d^2 - (1/2)ad \\
 \boxed{\hat{\sigma}_{BC_1}^2 = (1/2)\hat{\sigma}_A^2 + \hat{\sigma}_D^2 - (1/2)\widehat{CovAD}} &\quad \text{if } p = q = 1/2
 \end{aligned}$$

And then the second backcross population (Note this refers to backcross 1 with a different parent and it is not the result of two backcrosses):

$$\begin{aligned}
 \bar{X}_{BC_2} &= (1/2)d - (1/2)a \\
 \hat{\sigma}_{BC_2}^2 &= (1/2)d^2 + (1/2)a^2 - ((1/2)d - (1/2)a)^2 \\
 &= (1/2)d^2 + (1/2)a^2 - [(1/2)a^2 - 2(1/2)(1/2)ad + (1/2)d^2] \\
 &= (1/4)a^2 + (1/4)d^2 + (1/2)ad \\
 \boxed{\hat{\sigma}_{BC_2}^2 = (1/2)\hat{\sigma}_A^2 + \hat{\sigma}_D^2 + (1/2)\widehat{CovAD}} &\quad \text{if } p = q = 1/2
 \end{aligned}$$

Therefore, the sum of  $BC_1 + BC_2 = \hat{\sigma}_A^2 + 2\hat{\sigma}_D^2$  allows us to eliminate the covariance term. A similar procedure can be done for estimating the variance among and within selfed backcross generations (see Chapter 4).

## 2.10 Heritability, Genetic Gain, and Usefulness Concepts

Heritability is the degree of correspondence between the phenotype and the breeding value of an individual for a particular trait. The best way to determine the breeding value of a plant is to grow and examine its progeny. If the correlation between the breeding value and the phenotype is high then heritability is high (e.g., qualitative, less complex inherited traits). If dominance, epistatic, and environmental effects are large then heritability is low (e.g., quantitative traits). Also, the results of the estimation depend strictly on the population you are working with.

Estimates can be narrow or broad sense depending on which *genetic variance* is considered. Also, we can determine the heritability on an *individual plant basis* or on a *progeny mean basis* depending on the *generation* used.

Narrow-sense heritability can be defined as the ratio of the additive genetic variance to the phenotypic variance:

$$\boxed{\hat{h}^2 = \hat{\sigma}_A^2 / \hat{\sigma}_P^2} \quad \text{Narrow-sense heritability}$$

Warner (1952) developed a clear estimate of  $\hat{\sigma}_A^2$  by applying simple algebra on the generations studied above (for more details see Section 4.12). He showed the following:

$$\begin{aligned} \text{If } 2\hat{\sigma}_{F_2}^2 &= 2\hat{\sigma}_A^2 + 2\hat{\sigma}_D^2 \\ \frac{\hat{\sigma}_{BC_1}^2 + \hat{\sigma}_{BC_2}^2}{2} &= \frac{\hat{\sigma}_A^2 + 2\hat{\sigma}_D^2}{2} \\ &= \hat{\sigma}_A^2 \end{aligned}$$

Therefore,

$$\hat{h}^2 = \hat{\sigma}_A^2 / \hat{\sigma}_p^2$$

is a narrow-sense heritability estimate on an individual plant basis.

In maize, heritability estimates for grain yield vary from less than 0.1 (10%) when it is based on individual plants grown in one location (e.g., mass selection) to >0.8 (80%) when it is based on inbred progenies grown across locations with full-sibs and half-sibs having intermediate values.

There are several methods to estimate heritability on an individual plant basis. Different estimates can be obtained using the same populations:

**(a) Burton (1951)**

$$\boxed{\hat{h}^2 = \hat{\sigma}_{F_2}^2 - \hat{\sigma}_{F_1}^2 / \hat{\sigma}_{F_2}^2} \quad \text{broad sense}$$

**(b) Warner (1952)**

$$\boxed{\hat{h}^2 = [2\hat{\sigma}_{F_2}^2 - (\hat{\sigma}_{BC_1}^2 + \hat{\sigma}_{BC_2}^2)] / \hat{\sigma}_{F_2}^2} \quad \text{narrow sense}$$

**(c) Mahmud and Kramer (1951)**

$$\boxed{\hat{h}^2 = [\hat{\sigma}_{F_2}^2 - (\hat{\sigma}_{P_1}^2 \times \hat{\sigma}_{P_2}^2)] / \hat{\sigma}_{F_2}^2} \quad \text{broad sense}$$

**(d) Weber and Moorthy (1952)**

$$\boxed{\hat{h}^2 = \left[ \hat{\sigma}_{F_2}^2 - \left( \hat{\sigma}_{P_1}^2 \times \hat{\sigma}_{P_2}^2 \times \hat{\sigma}_{F_1}^2 \right)^{1/3} \right] / \hat{\sigma}_{F_2}^2} \quad \text{broad sense}$$

Going back to our generations:

$$\hat{h}_{F_2}^2 = \frac{\hat{\sigma}_{F_2}^2 - \hat{\sigma}_{We}^2}{\hat{\sigma}_{F_2}^2}$$

Usually less than 0.1  
Heritability in broad sense/individual plant basis  
It is common to use only one location  
 $\hat{\sigma}_{we}^2$  environmental effects (can be estimated)

$$\hat{h}_{F_3}^2 = \frac{\hat{\sigma}_{F_3}^2}{\hat{\sigma}_e^2/r + \hat{\sigma}_{F_3}^2}$$

Can be 0.75  
Heritability in broad sense  
It is common to use several observations per line  
Based on  $S_1$  progeny means

$$\hat{h}_{F_4}^2 = \frac{\hat{\sigma}_{F_4}^2}{\hat{\sigma}_e^2/r + \hat{\sigma}_{F_4}^2}$$

Can be 0.87 (50% more additive variance)  
Heritability in broad sense  
It is common to use several observations per line  
Based on  $S_2$  progeny means

If progenies (g) are evaluated in replicated (r) trials in different environments (e), heritability in the broad sense based on progeny means is:

$$\hat{h}^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_e^2/re + \hat{\sigma}_{ge}^2/e + \hat{\sigma}_g^2}$$

Heritability in broad sense  
Progeny mean basis

As expected heritability contributes to *genetic gain* ( $\Delta G$ ) as follows:

$$\Delta G = \hat{h}^2 S \quad \text{and } S = \text{selection differential} = (\bar{X}_s - \bar{X})$$

being  $\bar{X}_s$  = Mean of progeny selected and  $\bar{X}$  = Mean of the overall population

Table 2.14 shows an example of an intra-population recurrent selection program on NDSAB genetically broad-based population improved by 12 cycles of modified ear-to-row selection plus two cycles of full-sib recurrent selection. A heritability index was utilized for selecting the top progenies based on grain yield (YIELD), grain moisture at harvest (MOIST), test weight (TWT), and stalk lodging resistance (SL). TRT refers to treatment number and PI to a performance index including grain yield and grain moisture at harvest.

Breeders want to identify populations with high mean and large genetic variance. A function was created to combine information on the mean performance and genetic variance of a population and is called *usefulness criterion* ( $U$ ):

$$U = \bar{X} + \Delta G$$

**Table 2.14** Recurrent selection program including 200 full-sib progenies grown across three ND locations arranged in an augmented unreplicated experimental design. Only the top 16 full-sib progenies that were included to form cycle 14 are included

Pedigree	TRT	Yield	PI	MSTR	SL	TWT	Index
NDSAB(MER-FS)C13 SYN1-142	145	81.3	118.5	24.5	4.5	53.5	7.42
NDSAB(MER-FS)C13 SYN1-3	12	70.5	120.2	21.3	2.7	53.6	6.85
NDSAB(MER-FS)C13 SYN1-106	54	74.0	125.3	21.0	12.5	56.3	4.73
NDSAB(MER-FS)C13 SYN1-76	95	71.9	109.4	23.6	6.9	52.4	4.54
NDSAB(MER-FS)C13 SYN1-112	43	66.5	107.2	22.1	6.2	55.4	4.06
NDSAB(MER-FS)C13 SYN1-73	80	63.9	113.8	20.0	8.1	56.8	3.99
NDSAB(MER-FS)C13 SYN1-119	55	64.3	102.5	22.3	4.3	57.1	3.97
NDSAB(MER-FS)C13 SYN1-4	3	67.2	108.2	21.8	9.1	55.1	3.52
NDSAB(MER-FS)C13 SYN1-17	18	68.0	114.9	20.8	11.5	57.7	3.52
NDSAB(MER-FS)C13 SYN1-24	129	74.0	121.7	21.7	15.1	52.8	3.50
NDSAB(MER-FS)C13 SYN1-14	5	55.4	100.8	19.7	2.9	58.1	3.47
NDSAB(MER-FS)C13 SYN1-97	64	67.4	99.7	24.0	5.6	52.6	3.46
NDSAB(MER-FS)C13 SYN1-61	84	59.0	98.9	21.6	3.1	56.7	3.35
NDSAB(MER-FS)C13 SYN1-36	122	79.2	112.1	25.3	13.8	51.1	3.32
NDSAB(MER-FS)C13 SYN1-94	58	68.4	107.0	23.0	9.3	54.5	3.07
NDSAB(MER-FS)C13 SYN1-92	70	61.7	94.3	23.4	3.0	53.4	3.04
$\bar{X}$		57.9	91.0	23.0	15.1	53.9	-1.78
$\bar{X}_s$		68.3	109.7	22.2	7.4	54.8	4.11
$S$		10.37	18.65	0.75	7.69	0.88	5.89

$$\text{Index} = [(0.30 \times \text{Yield}) + (0.69 \times \text{TWT}) - (0.58 \times \text{MSTR}) - (0.33 \times \text{SL})]$$

## 2.11 Generation Mean Analysis

We will now introduce our first genetic analysis with development of progenies. This is a type of genetic analysis that can be useful for preliminary studies. Several models have been developed for the analysis of generation means. It is important to note that these models estimate the relative importance of genetic effects based upon the means of different generations and not based on their variances. Genetic variances are determined from the summation of squared effects for each locus. This type of genetic analysis does not involve development of progenies that have a family structure of sib-ships as in Chapter 4, but it includes genetic populations (or generations) that are similar to those characterized by the special case of  $p = q = 0.5$  (e.g.,  $F_2$  populations). Instead of estimating genetic variation within generations, we will concern ourselves with relative genetic effects estimated from the means of different generations. Mather (1949) presented several generation comparisons to test for additiveness of genetic effects for estimation of  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$ . If the scale of measurement deviated from additivity, he suggested a transformation to make the effects additive. The generation models were extended to include estimation of epistatic effects. Several models have been developed for analysis of generation means (Anderson and Kempthorne, 1954; Hayman, 1958, 1960; Van der Veen, 1959; Gardner and Eberhart, 1966). Because of similarity in nomenclature we

will use the Hayman (1958, 1960) model to illustrate the type of genetic information obtained from generation mean analyses. As an example, consider the generations produced from the cross of two inbred lines (e.g., special case of  $p = q = 0.5$ ). For estimation of genetic effects we will use the means of each generation rather than develop progenies within the segregating generations.

Hayman (1958) defined his base population as the  $F_2$  population resulting from a cross of two inbred lines. If they differ by any number of unlinked loci, expectations of parents and their descendant generations in terms of genetic effects relative to the  $F_2$  generation are as follows:

$$\begin{aligned}
 P_1 &= m + a - (\tfrac{1}{2})d + aa - ad + (\tfrac{1}{4})dd \\
 P_1 &= m - a - (\tfrac{1}{2})d + aa + ad + (\tfrac{1}{4})dd \\
 F_1 &= m + (\tfrac{1}{2})d + (\tfrac{1}{4})dd \\
 F_2 &= m \\
 F_3 &= m - (\tfrac{1}{4})d + (\tfrac{1}{16})dd \\
 F_4 &= m - (\tfrac{3}{8})d + (\tfrac{9}{64})dd \\
 F_5 &= m - (\tfrac{7}{16})d + (\tfrac{49}{256})dd \\
 F_6 &= m - (\tfrac{15}{32})d + (\tfrac{225}{1024})dd \\
 BC_1 &= m + (\tfrac{1}{2})a + (\tfrac{1}{4})aa \\
 BC_2 &= m - (\tfrac{1}{2})a + (\tfrac{1}{4})aa \\
 BC_1^2 &= m + (\tfrac{3}{4})a + (\tfrac{9}{16})aa \\
 BC_2^2 &= m - (\tfrac{3}{4})a + (\tfrac{9}{16})aa \\
 BS_1 &= m + (\tfrac{1}{2})a - (\tfrac{1}{4})d + (\tfrac{1}{4})aa - (\tfrac{1}{4})ad + (\tfrac{1}{16})dd \\
 BS_2 &= m - (\tfrac{1}{2})a - (\tfrac{1}{4})d + (\tfrac{1}{4})aa - (\tfrac{1}{4})ad + (\tfrac{1}{16})dd
 \end{aligned}$$

or in general the observed mean  $= m + \alpha a + \beta d + \alpha^2 aa + 2\alpha\beta ad + \beta^2 dd$ , where  $\alpha$  and  $\beta$  are the coefficients of  $a$  and  $d$ . Because the  $F_2$  mean is  $(\tfrac{1}{2})d$  and the  $F_1$  mean is equal to  $d$ , the  $F_1$  mean relative to the  $F_2$  mean has an added increment of  $(\tfrac{1}{2})d$ . Terms  $a$  and  $d$  are the same as those illustrated in the special case of  $p = q = 0.5$ , where  $a$  indicates additive effects and  $d$  indicates dominance effects. Hayman (1958) used lowercase letters to indicate summation over all loci by which the two inbred lines differ. Thus  $a$  measures the pooled additive effects;  $d$  the pooled dominance effects; and  $aa$ ,  $ad$ , and  $dd$  the pooled digenic epistatic effects.

The different generations listed can be produced rather easily for cross- and self-pollinated species. Hand pollinations are necessary for all generations in maize, but selfing generations can be obtained naturally in self-pollinated species. Bults of progenies of each generation are evaluated in replicated experiments repeated over environments, and generation means can be determined for traits under study. In growing different generations, one should be cognizant of two important considerations in order to have valid estimates of the generation means:

- (1) Sufficient sampling of segregating generations is necessary to have a representative sample of genotypes. In parental and  $F_1$  generations no sampling is involved, but  $F_2$ ,  $F_3$ ,  $F_4$ , . . . , and backcross generations will be segregating and sample size has to be considered.

- (2) In maize it is necessary to consider the level of inbreeding of each generation, and it becomes necessary to have sufficient border rows in experimental plots to minimize competition effects of adjacent plots.

Several different possibilities exist for the type and number of generations that can be included in a generation mean experiment. If the two parents and the  $F_1$ ,  $F_2$ , and  $F_3$  generations are evaluated, we have five means for comparison. Expectations of each generation can be determined and a least-squares analysis made to estimate  $m$ ,  $a$ , and  $d$  with a fair degree of precision. For this simple experiment we can also make a goodness-of-fit test (observed means compared with predicted means) to determine the sufficiency of the model for  $m$ ,  $a$ , and  $d$  to explain the differences among the generation means.

Letting  $m$  = general mean,  $a$  = sum of signed additive effects, and  $d$  = signed dominance effects, we have the following expressions with  $F_2$  as the base population:

$$\begin{aligned} P_1 &= m + a & F_1 &= m + (\frac{1}{2})d \\ P_2 &= m - a & F_2 &= m \\ & & F_3 &= m - (\frac{1}{4})d \end{aligned}$$

By use of the technique suggested by Mather (1949), the five equations can be reduced to the following normal equations:

$$\begin{aligned} 5m + (\frac{1}{4})d &= Q_1 (P_1 + P_2 + F_1 + F_2 + F_3) \\ 2a &= Q_2 (P_1 - P_2) \\ (\frac{1}{4})m + (\frac{5}{16})d &= Q_3 (F_1 + F_2) \end{aligned}$$

In matrix form the set of equations is equal to

$$\begin{bmatrix} 5 & 0 & \frac{1}{4} \\ 0 & 2 & 0 \\ \frac{1}{4} & 0 & \frac{5}{16} \end{bmatrix} \begin{bmatrix} m \\ a \\ d \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \\ Q_3 \end{bmatrix}$$

Solving for parameters  $m$ ,  $a$ , and  $d$ , we get the following estimates:

$$\begin{aligned} \hat{m} &= (\frac{5}{24}) Q_1 - (\frac{1}{6}) Q_3 \\ \hat{a} &= (\frac{1}{2}) Q_2 \\ \hat{d} &= -(\frac{1}{6}) Q_1 + (\frac{10}{3}) Q_3 \end{aligned}$$

The estimates of  $m$ ,  $a$ , and  $d$  can be inserted in the predicted values and can be compared with the observed values for each generation. If the squares of deviations of the expected from the observed are significant, the three estimated parameters

are not sufficient to explain differences among the generation means; this is a goodness-of-fit test for epistasis and/or linkage to determine if the three parameters included are sufficient or if more are needed. The best procedure would be to sequentially fit the successive models starting with the mean and add one term with each successive fit. Tests of residual mean squares can be made for each model to determine how much of the total variation among generations is explained by different parameters in the model. High-speed computers facilitate such computations, and a weighed least-squares analysis can be done rather easily.

The similarity of genetic populations included for this simple generation mean experiment and genetic populations used for estimation of genetic variances for the special case of  $p = q = 0.5$  is obvious. If necessary measurements have been made for the different genetic populations, we can also obtain the following sets of equations:

Variance among $F_2$ individuals	$= \hat{\sigma}_A^2 + \hat{\sigma}_D^2 + E_1$
Variance among $F_3$ progeny means	$= \hat{\sigma}_A^2 + (1/4)\hat{\sigma}_D^2 + E_2$
Variance within $F_3$ progenies	$= (1/2)\hat{\sigma}_A^2 + (1/2)\hat{\sigma}_D^2 + E_1$
Covariance between $F_2$ individuals and $F_3$ progeny means	$= \hat{\sigma}_A^2 + (1/2)\hat{\sigma}_D^2$
Variance among parents and $F_1$ individuals	$= E_1$
Experimental error	$= E_2$

Six equations are available for estimation of two heritable and two non-heritable sources of variation. Direct estimates of  $E_1$  and  $E_2$  are available, but an unweighted (or preferably a weighted) least-squares analysis can be used to estimate the four parameters ( $\hat{\sigma}_A^2$ ,  $\hat{\sigma}_D^2$ ,  $E_1$ , and  $E_2$ ) from the six equations. If one estimates the genetic effects  $a$  and  $d$  (by use of generation mean analysis) and the genetic variances  $\hat{\sigma}_A^2$  and  $\hat{\sigma}_D^2$ , there probably will be little relation in the magnitude of the two sets of estimates. This should be expected because in the first instance we are estimating the sum of the signed genetic effects, whereas in the second instance we are estimating variances that are the squares of the genetic effects. For maize it seems that estimates of the  $d$  effects are usually greater, especially if the  $F_1$  generation is included. On the other hand, estimation of genetic variances in maize usually shows that estimates of  $\hat{\sigma}_A^2$  are similar to or greater than estimates of  $\hat{\sigma}_D^2$ . The expression of heterosis in  $F_1$  crosses of two inbred lines of maize probably has a much greater effect on the estimate of  $d$  for maize than for many other crop species.

Limitations of the generation mean analysis if the model includes epistatic effects were discussed by Hayman (1960). Briefly, if the residuals are not significantly different from zero after  $m$ ,  $a$ , and  $d$  are fitted, we have unique estimates for  $a$  and  $d$ . However, if it is necessary to include epistatic effects in the model, estimates of digenic epistatic effects are unique but estimates of  $a$  and  $d$  are confounded with some of the epistatic effects. Hence if epistatic effects are not present, estimates of  $a$  and  $d$  effects are meaningful and unbiased by linkage disequilibrium; if epistatic effects are present, estimates of  $a$  and  $d$  effects are biased by epistatic effects and linkage disequilibrium (if present). Estimation of digenic epistatic effects

is unbiased if linkage of interacting loci and higher order epistatic effects are absent. Because of the bias in the estimates of  $a$  and  $d$  effects, when a model that includes epistatic effects is used, the relative importance of  $a$  and  $d$  effects vs. epistatic effects cannot be directly assessed. Some indication of their relative importance may be gained by comparing residual sums of squares after fitting the three-parameter ( $m$ ,  $a$ , and  $d$ ) and the six-parameter ( $m$ ,  $a$ ,  $d$ ,  $aa$ ,  $ad$ , and  $dd$ ) models.

It seems that the primary function of generation mean analysis is to obtain some specific information about a specific pair of lines. How useful the information obtained from generation mean analysis is to the maize breeder is not obvious. For quantitative traits the estimates of genetic effects would be quite different for different pairs of lines, depending on the relative frequency of opposing and reinforcing effects for the specific pair of lines studied. The cancellation of opposing effects may confound interpretations, but a complete diallel of interested lines could be used to determine which have opposing and reinforcing effects. Generation mean analysis is amenable for use in self-pollinated species because limited hand pollinations are required to produce the different generations; hence generation mean analysis may provide some information on the relative importance of non-additive genetic effects for the justification of a hybrid breeding program. The relative importance of dominance effects could be determined by comparing different generations derived from the  $F_1$  generation, which would involve only the cross of the two parents to produce the  $F_1$  generation. For maize, however, controlled pollinations would be necessary for all generations.

Generation mean analysis has some advantages and disadvantages in comparison with mating designs used for estimation of genetic components of variance (see Chapter 4). Because we are working with means (first-order statistics) rather than variances (second-order statistics), the errors are inherently smaller. We can rather easily extend generation mean analysis to more complex models that include epistasis, but the main effects ( $a$  and  $d$ ) are not unique when epistatic effects are present. Generation mean analysis is equally applicable to cross- and self-pollinating species. Smaller experiments are required for generation mean analysis to obtain the same degree of precision. However, an estimate of heritability cannot be obtained and one cannot predict genetic advance because estimates of genetic variances are not available. Cancellation of effects may be a significant disadvantage because, say, dominance effects may be present but opposing at various loci in the two parents and cancel each other. Generation mean analysis does not reveal opposing effects, but this may be overcome to some extent by a balanced set of diallel crosses.

This discussion for generation mean analysis is restricted to the use of parents being either inbred or pure lines, i.e., relatively homozygous and homogeneous. Generation mean analysis has been extended to populations generated from parents that are not homozygous. Robinson and Cockerham (1961) presented an analysis for two varieties and  $n$  alleles at each locus. Gardner and Eberhart (1966) included  $n$  varieties with only two alleles at a locus. The analysis by Robinson and Cockerham (1961) is orthogonal in the partitioning of sums of squares, and tests can be made to determine the presence of non-additive effects. All the generation mean



analyses provide information on the relative importance of genetic effects, but the information in most instances may not be useful to applied breeders, particularly those that are conducting long-term selection programs.

In summary, if we consider the generations produced from the cross of two inbred lines, Hayman (1958) defines the  $F_2$  generation as his base or reference population. Therefore, if inbred lines differ by any number of unlinked loci, expectations of parents and their descendant generations in terms of genetic effects relative to the  $F_2$  generation are based on the following model:

$$Y_{ijkl} \text{ (observed mean)} = m + \alpha_i + \beta_j + \alpha_k^2 + 2\alpha\beta_{ij} + \beta_l^2 \quad \text{or}$$

$$Y = m + \alpha_a + \beta_d + \gamma_{aa}^2 + 2\alpha\beta_{ad} + \beta^2_{dd} \quad \text{following notation of Gamble (1962a, b)}$$

‘ $a$ ’ indicates pooled additive effects across loci

‘ $d$ ’ indicates pooled dominance effects across loci

‘ $aa$ ,’ ‘ $ad$ ,’ and ‘ $dd$ ’ indicate pooled digenic epistatic effects

Considering the means for each generation as follows:

Generation	Mean	Expectations
$P_1$	$a$	$m + a - (\frac{1}{2})d + aa - ad + (\frac{1}{4})dd$
$P_2$	$-a$	$m - a - (\frac{1}{2})d + aa + ad + (\frac{1}{4})dd$
$F_1$	$d$	$m + (\frac{1}{2})d + (\frac{1}{4})dd$
$F_2$	$(\frac{1}{2})d$	$m$
$BC_1$	$(\frac{1}{2})a + (\frac{1}{2})d$	$m + (\frac{1}{2})a + (\frac{1}{4})aa$
$BC_2$	$-(\frac{1}{2})a + (\frac{1}{2})d$	$m - (\frac{1}{2})a + (\frac{1}{4})aa$

More generations can be included. The more generations we include the better the estimates (e.g., less error, more precision) but more experiments might be needed.

Therefore, we can estimate gene effects when epistasis is present with information of generation means:

$$‘a’ = \overline{BC}_1 - \overline{BC}_2$$

$$‘d’ = \overline{F}_1 - 4\overline{F}_2 + 2\overline{BC}_1 + 2\overline{BC}_2 - (\frac{1}{2})\overline{P}_1 - \overline{P}_2$$

$$‘aa’ = 2\overline{BC}_1 + 2\overline{BC}_2 - 4\overline{F}_2$$

$$‘ad’ = (\frac{1}{2})\overline{P}_2 - (\frac{1}{2})\overline{P}_1 + \overline{BC}_1 - \overline{BC}_2$$

$$‘dd’ = \overline{P}_1 + \overline{P}_2 + 2\overline{F}_1 + 4\overline{F}_2 - 4\overline{BC}_1 - 4\overline{BC}_2$$

In this case we do not expect deviations since the number of unknown estimates equals the number of generations used in the model.

### **2.11.1 Assumptions for Analysis**

- 1) Two alleles per locus.
- 2) Most positive alleles in  $P_1$  and most negative alleles in  $P_2$ .
- 3) No linkage of interacting loci.
- 4) No trigenic or higher order epistasis.
- 5)  $F_2$  is the reference population.
- 6) Sufficient sampling of segregating generations (representative sample of genotypes).
- 7) No competition effects among generations at different levels of inbreeding.

### **2.11.2 Limitations of This Analysis**

- 1) Unlike variances, means can estimate neither heritability nor genetic gain for prediction.
- 2) Genetic effects are always summing or subtracting. Therefore, there is a cancellation of effects that are not detected.

### **2.11.3 Advantages of This Analysis**

- 1) Useful for preliminary studies, i.e., is there enough dominance to have good hybrids? Allows to study a new or unknown population.
- 2) Means are estimated with greater precision than variances.
- 3) Can be extended to more complex models.

## **References**

- Anderson, V. L., and O. Kempthorne. 1954. A model for the study of quantitative inheritance. *Genetics* 39:883–98.
- Bernardo, R. 2002. *Breeding for Quantitative Traits in Plants*. Stemma Press, Woodbury, MN.
- Cockerham, C. C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics* 39:859–82.
- Cockerham, C. C. 1963. Estimation of genetic variances. In *Statistical Genetics and Plant Breeding*, Vol. 982, W. D. Hanson and H. F. Robinson, (eds.), pp. 53–94. NAS-NRC. Washington, DC.
- Compton, W. A., C. O. Gardner, and J. H. Lonquist. 1965. Genetic variability in two open-pollinated varieties of corn (*Zea mays* L.) and their  $F_1$  progenies. *Crop Sci.* 5:505–8.
- Falconer, D. S. 1960. *Introduction to Quantitative Genetics*. The Ronald Press, New York, NY.
- Falconer, D. S., and T. F. C. Mackay. 1996. *Introduction to Quantitative Genetics*, 4th edn. Longman Group, Essex.
- Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* 52:399–433.
- Gamble, E. E. 1962a. Gene effects in corn (*Zea mays* L.). I. Selection and relative importance of gene effects for yield. *Can. J. Plant Sci.* 42:339–48.

- Gamble, E. E. 1962b. Gene effects in corn (*Zea mays* L.). II. Relative importance of gene effects for plant height and certain component attributes of yield. *Can. J. Plant Sci.* 42:349–58.
- Gardner, C. O., and S. A. Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. *Biometrics* 22:439–52.
- Geldermann, H. 1975. Investigations on inheritance in quantitative characters in animals by gene markers. I. Methods. *Theor. Appl. Genet.* 46:319–330.
- Hayman, B. I. 1958. The separation of epistatic from additive and dominance variation in generation means. *Heredity* 12:371–90.
- Hayman, B. I. 1960. The separation of epistatic from additive and dominance variation in generation means. II. *Genetica* 31:133–46.
- Kempthorne, O. 1954. The correlations between relatives in a random mating population. *Phil. R. Soc. Lond. B* 143:103–13.
- Kempthorne, O. 1957. *An Introduction to Genetic Statistics*. Wiley, New York, NY.
- Lonnquist, J. H. 1963. Gene action and corn yields. *Annu. Corn Sorghum Res. Conf. Proc.* 18:37–44.
- Lush, J. L. 1945. *Animal Breeding Plans*. Iowa State University, Press, Ames, IA.
- Mather, K. 1941. Variation and selection of polygenic characters. *J. Genetics* 41:159–193.
- Mather, K. 1949. *Biometrical Genetics*. Methuen, London.
- Robinson, H. F., and C. C. Cockerham. 1961. Heterosis and inbreeding depression in populations involving two open-pollinated varieties of maize. *Crop Sci.* 1:68–71.
- Snape, J., J. Simmonds, M. Leverington, L. Fish, E. Sayers, L. Alibert, S. Orford, M. Ciavarrella, and S. Griffiths. 2008. The yield dynamics of European winter wheat improvement revealed by large scale QTL analysis. In *Breeding 08: Conventional and Molecular Breeding of Field and Vegetable Crops*, p. 29. Novi Sad, Serbia.
- Sorrells, M. E. 2008. Association breeding strategies for improvement of self-pollinated crops. In *Breeding 08: Conventional and Molecular Breeding of Field and Vegetable Crops*, p. 20. Novi Sad, Serbia.
- Van der Veen, J. H. 1959. Tests of non-allelic interaction and linkage for quantitative characters in generations derived from two diploid pure lines. *Genetica* 30:201–32.
- Warner, J. N. 1952. A method for estimating heritability. *Agron J.* 44:427–30.

Quantitative Genetics in Maize Breeding

Hallauer, A.R.; Carena, M.J.; Miranda Filho, J.B.

2010, XVI, 664 p., Hardcover

ISBN: 978-1-4419-0765-3