Quantitative Genetics for Plant Breeding

Walter Suza (Editor); Kendall Lamkey (Editor); William Beavis; Katherine Espinosa; Mark Newell; and Anthony Assibi Mahama

Iowa State University Digital Press
Ames, Iowa

Quantitative Genetics for Plant Breeding Copyright © by Walter Suza (Editor); Kendall Lamkey (Editor); William Beavis; Katherine Espinosa; Mark Newell; and Anthony Assibi Mahama is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, except where otherwise noted.

You are free to copy, share, adapt, remix, transform, and build upon the material, so long as you follow the terms of the license.

How to cite this publication:

Suza, W., & Lamkey, K. (Eds.). (2023). Quantitative Genetics for Plant Breeding. Iowa State University Digital Press.

This is a publication of the
Iowa State University Digital Press
701 Morrill Rd, Ames, IA 50011
https://www.iastatedigitalpress.com
digipress@iastate.edu

Contents

About the PBEA Series

Background

The Plant Breeding E-Learning in Africa (PBEA) e-modules were originally developed as part of the Bill & Melinda Gates Foundation Contract No. 24576.

Building on Iowa State University's expertise with online plant breeding education, the PBEA e-modules were developed for use in curricula to train African students in the management of crop breeding programs for public, local, and international organizations. Collaborating with faculty at Makerere University in Uganda, University of KwaZulu-Natal in South Africa, and Kwame Nkrumah University of Science and Technology in Ghana, our team created several e-modules that hone essential capabilities with real-world challenges of cultivar development in Africa using Applied Learning Activities. Our collaboration embraces shared goals, sharing knowledge and building consensus. The pedagogical emphasis on application produces a coursework-intensive MSc program for Africa.

- **PBEA Project Director:** Walter Suza
- **Original Module Coordinator:** William Beavis

Collaborating Faculty and Experts in Africa: Richard Akromah, Stephen Amoah, Maxwell Asante, Ben Banful, John Derera, Richard Edema, Paul Gibson, Sadik Kassim, Rufaro Madakadze, Settumba Mukasa, Margaret Nabasirye, Daniel Nyadanu, Thomas Odong, Patrick Ongom, Joseph Sarkodie-Addo, Paul Shanahan, Husein Shimelis, Julia Sibiya, Pangirayi Tongoona, Phinehas Tukamuhabwa.

The authors of this textbook series adapted and built upon the PBEA modules to develop a series of textbooks covering individual topic areas. It is our hope that this project will facilitate wider dissemination and reuse of the PBEA modules' content.

Explore the Series

- Crop Genetics
- Quantitative Methods for Plant Breeding
- Molecular Plant Breeding
- Quantitative Genetics for Plant Breeding

- Crop Improvement
- Cultivar Development

Chapter 1: Gene Frequencies

William Beavis; Kendall Lamkey; and Anthony Assibi Mahama

The challenge of Quantitative Genetics is to connect traits measured on quantitative scales with genes that are inherited and evaluated as discrete units. This challenge was addressed through the development of theory between 1918 and 1947. The theory is now referred to as the **Modern Synthesis** and required another 50 years for technological innovations and experimental biologists to validate. Luminaries such as RA Fisher, Sewell Wright, JBS Haldane, and John Maynard Smith were able to develop the theory that is still widely applied without the benefit of high throughput 'omics' technologies. Indeed, modern synthesis was developed before the knowledge of the structure of DNA.

Population genetics characterizes how discrete units, i.e., alleles, change in breeding populations. Such characterization is the basis for understanding the structure of genomes and breeding populations. The forces of mutation (Fig. 1), migration, selection, and drift will alter the structure of breeding populations. Herein we will learn how to characterize population structure at one or two loci in diploid crop species. This will set the foundation for characterizing structure based on any number of loci and for polyploid crops that you may encounter in more advanced courses.

Learning Objectives

- Demonstrate the relevance of population genetics concepts to plant breeding populations.
- Demonstrate the relevance of a purely theoretical Ideal Population to plant breeding populations.
- Demonstrate understanding of the purpose of populations in Hardy-Weinberg Equilibrium
- Distinguish populations in Hardy-Weinberg Equilibrium from the Ideal Population.
- Describe the impact of mutation, selection, and drift on breeding populations.

Fig. 1 A red Darwin hybrid tulip "Appeldoorn" with a mutation resulting in half of one petal being yellow. Photo by LepoRello; Licensed under CC BY-SA 3.0 via Wikimedia Commons.

Allelic and Genotypic Variation

Ideal Population

In order to understand the genetic structure of a population, it is necessary to establish a standard reference population so that the breeding population can be characterized relative to the standard. For this purpose, an 'ideal' conceptual base population can be defined as infinitely large with the potential to extract finite sub-populations through sampling, such as depicted in the following figure and described in Falconer and Mackay (1996):

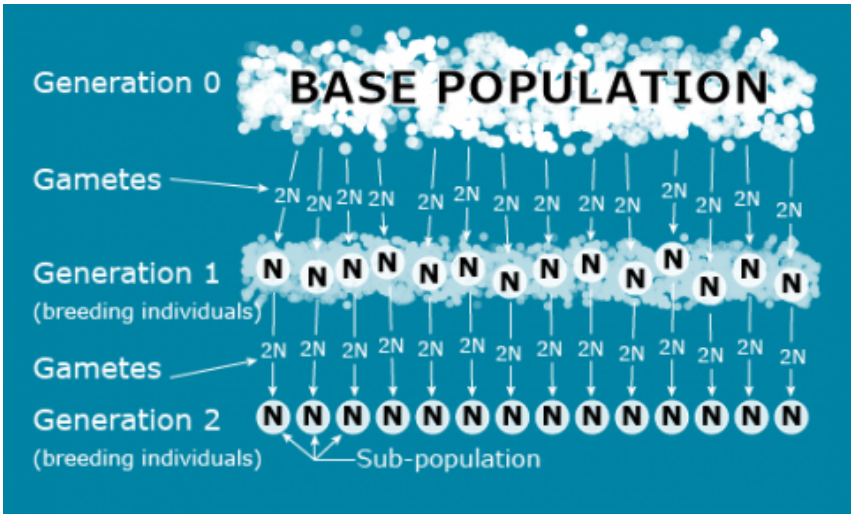


Fig. 2 Reference population. Adapted from Falconer and Mackay, 1996.

Note that the sub-populations depicted in Fig. 2 are based on a genetic sampling process that

is affected by the reproductive biology of the species. Unlike animal species, crop species can reproduce in a variety of ways:

- Sexual
  - Cross-Pollination
  - Self-Pollination
  - Mixtures of Self and Cross-pollination
- Asexual
  - Clonal
  - Doubled haploids
  - Apomixis

Assumptions

In the ideal population depicted in Fig. 2, the following assumptions are true:

1. The base population is infinite or at least too large to count.
2. There is no migration between sub-populations.
3. There is no breeding between overlapping generations.
4. The number of breeding individuals is the same in each subpopulation.
5. There is random mating within a subpopulation.
6. There is no Selection.
7. There is no Mutation.

Of course, in real populations, these assumptions are violated.

Allelic and Genotypic Frequencies

We first model a single locus with only two alleles in an ideal breeding population of diploid individuals. Define the following:

N = number of breeding individuals in a subpopulation (population size)
t = time usually measured in terms of generations
q = frequency of one of two alleles at a locus within a subpopulation
p = 1 – q = frequency of a second allele at a locus within a subpopulation
$\overline{p}$ = frequency of a second allele across the subpopulations (the mean of p)
$\mathbf{p_0}$ = frequency of a second allele in the base population

Due to the assumptions associated with an ideal reference population, $\overline{q}$ = q_0 at any stage or generation of the sampling process, so q_0 can be used interchangeably with $\overline{q}$.

The alleles, allele frequencies, genotypes, and genotypic frequencies can be represented in Table 1 and in Equations 1 and 2.

Table 1 Alleles, allele frequencies, genotypes, and genotypic frequencies.

	Alleles		Genotypes		
	A	a	AA	Aa	aa
Frequencies	p	q	P_{AA}	P_{Aa}	P_{aa}

$$p + q = 1.$$

Equation 1 Sum of allele frequencies.

$$P_{AA} + P_{Aa} + P_{aa} = 1.$$

Equation 2 Sum of genotype frequencies,

where:
p, q are as defined earlier,
P_{AA}, P_{Aa}, P_{aa} = frequencies of the three genotypes.

Variance of Allele Frequency

The relationship between allele frequencies and genotype frequencies can be expressed as in Equation 3.

$$p = P_{AA} + \frac{1}{2}P_{Aa} \;\; and \;\; q = P_{aa} + \frac{1}{2}P_{Aa},$$

Equation 3 Equation for determining allele frequency,

where:
$terms$ are as defined earlier.

A particular sub-population is a random sample of N individuals or 2N gametes (for a diploid) from the base population. Therefore, the expected gene frequency of a particular allele in the sub-populations is $q_{0,}$ and the variance of q is represented by Equation 4.

$$\sigma_q^2 = \frac{p_0 q_0}{2N}.$$

Equation 4 Equation for estimating the variance of an allele,

where:
σ_q^2 = the variance of an allele,
N = the number of individuals.

Since q_0 is a constant, the variance of the change in allele frequency ($q_1 - q_0$) is also:

$$\sigma_{\Delta q}^2 = \frac{p_0 q_0}{2N}.$$

Equation 5 Equation for estimating the variance of change in allele frequency,

where:
$\sigma_{\Delta q}^2$ = the change in allele frequency,
other terms are as defined before.

Frequency Estimators

In addition to the genetic sampling process depicted in Fig. 2, a statistical sampling process can be used to estimate frequencies, variances, and covariances of alleles and genotypes in a sub-population. If we sample **n** individuals from a population of size **N,** then notationally (Equation 6),

$$n = n_{AA} + n_{Aa} + n_{aa}, \text{ then } n_A = 2n_{AA} + n_{Aa}$$

Equation 6 Equation for determining the number, n, of individuals and number of A individuals in a sample from a population,

where:
n = the sample size,
other terms are as defined before.

Estimates of the frequency of the A allele and AA genotype in the sample are obtained using Equation 7.

$$\hat{p}_A = \frac{1}{2n} n_A \text{ and } \hat{p}_{AA} = \frac{1}{n} n_{AA}$$

Equation 7 Estimating the frequency of A allele and AA genotype in a sample,

where:
$\hat{p}_A$ = the estimate of the A allele frequency,
$\hat{p}_{AA}$ = the estimate of the AA genotype frequency.

Expected Number of Alleles

Recognizing that statistical sampling at a locus with two alleles in a diploid population is represented as a binomial random process, the expected number of A alleles and their frequency in a sample can be determined using Equation 8.

$$E(n_A) = 2nP_{AA} + nP_{Aa} \;\; and \;\; E(\hat{p}_A) = P_A$$

Equation 8 Estimating the expected number of a allele,

where:
$E(n_A)$ = the expectation of the A allele,
Other terms are as defined previously.

Thus, $\hat{p}_A$ is an unbiased estimator of the population parameter p_A.

Using the definition of variance, we can likewise find the $Var(n_A)$ and $Var\left(\hat{p}_A\right)$ using Equations 9 and 10. All terms have been defined previously.

$$Var(n_A) = 2n(p_A + P_{AA} - 2p_A^2)$$

Equation 9 Calculating the variance of *n* number of the *A* allele.

$$Var\left(\hat{p}_A\right) = \frac{1}{2n}\left(p_A + P_{AA} - 2p_A^2\right)$$

Equation 10 Calculating the variance of the estimated frequency of the A allele.

Note that p_A and P_{AA} are usually unknown, so we often substitute $\hat{p}_A$ and $\hat{P}_{AA}$ in the calculation of the $Var(\hat{p}_A)$. Also, note that $Var(\hat{p}_A)$ is not the variance of a Binomial distribution. If the population sampled is in Hardy-Weinberg Equilibrium (see below), the genetic sampling of alleles will be random so that $P_{AA} = p_A^2$ and $P_{Aa} = 2p_Ap_a$. The variance of the estimated frequency of the A allele can be obtained using Equation 11, which has the form of the variance from a binomial distribution.

$$Var\left(\hat{p}_A\right) = \frac{1}{2n} p_A \left(1 - p_A\right)$$

Equation 11 Alternative equation for calculating the variance of estimated frequency of an allele.

where:
terms are as defined previously.

Hardy-Weinberg Equilibrium

The proof of the Hardy Weinberg Equilibrium (HWE) requires the following assumptions (Falconer and Mackay, 1996):

1. Allele frequency in the parents is equal to the allele frequency in the gametes.
    - Assumes normal gene segregation.
    - Assumes equal fertility of parents.
2. Allele frequency in gametes is equal to the allele frequency in gametes forming zygotes.
    - Assumes equal fertilizing capacity of gametes.
    - Assumes a large population.
3. Allele frequency in gametes forming zygotes is equal to allele frequencies in zygotes.
4. Genotype frequency in zygotes is equal to genotype frequency in progeny.
    - Assumes random mating.
    - Assumes equal gene frequencies in male and female parents.
5. Genotype frequencies in progeny do not alter gene frequencies in progeny.
    - Assumes equal viability.

For a two-allele locus in a population in HWE, $P_{AA} = p^2, P_{Aa} = 2pq, and\ P_{aa} = q^2$. HWE at a given genetic locus is achieved in one generation of random mating. Genotype frequencies in the progeny depend only on the allele frequencies in the parents and not on the genotype frequencies of the parents.

Disequilibrium

As discussed, there are several processes that can force allelic and genotypic frequencies to deviate from HWE. Deviations from equilibrium are referred to as **disequilibrium** and are often denoted with a disequilibrium coefficient, D. In the two allele case, the genotypic frequencies can be represented as $P_{AA} = p_A^2 + D_A, P_{Aa} = 2p_Ap_a - 2D_A, and\ P_{aa} = p_a^2 + D_A$.

Thus, the disequilibrium coefficient can be estimated using Equation 12.

$$\widehat{D}_A = \widehat{P}_{AA} - p_A^2.$$

Equation 12 Equation for estimating D.

where:
$E(\widehat{D}_A)$ = the estimate of the disequilibrium coefficient of the A allele,
Other terms are as defined previously.

Note that the expectation of $\widehat{D}_A$ can be obtained from Equation 13. All terms are defined earlier.

$$E(\widehat{D}_A) = D_A - \frac{1}{2n}[p_A(1-p_A) + D_A].$$

Equation 13 Equation for estimating D.

The, $\widehat{D}_A$ is biased. Although the estimate of D_A is biased, as the sample size, n, becomes large, the bias becomes small. Thus, emphasizing the need for large sample sizes in drawing inferences about disequilibrium from Hardy-Weinberg.

Variance

The $Var(\widehat{D}_A)$ can likewise be derived as (Equation 14:

$$\cong \frac{1}{n}[p_A^2(1-p_A)^2 + (1-2p_A)^2 D_A - D_A^2].$$

Equation 14 Equation for estimating the variance of $Var(\widehat{D}_A)$.

If n is large, $E(\widehat{D}_A) \cong D_A$, and $\widehat{D}_A \sim N[E(\widehat{D}_A), Var(\widehat{D}_A)]$, a normal distribution. So, a standard normal variate, Z can be constructed as:

$$Z = \frac{\widehat{D}_A - E(\widehat{D}_A)}{\sqrt{Var(\widehat{D}_A)}}.$$

Equation 15 Equation for estimating the standard normal variate, Z.

Goodness of Fit

Alternatively, because $Z^2 = \chi^2$, therefore Equation 16 allows the calculation of chi-square.

$$\chi^2_A = \frac{nD^2_A}{p^2_A(1-p_A)},$$

Equation 16 Equation for estimating chi-square.

This form enables the direct use of genotypic counts, n_{AA}, n_{Aa}, n_{aa}, as shown in Table 2.

Table 2 Representations of observed and expected genotypic counts and differences between the counts.

n/a	Genotypes		
n/a	AA	Aa	aa
Observed (O)	nAA	n_{Aa}	n_{aa}
Expected (E)	np^2_A	$2p_A(1-p_A)$	$n(1-p_A)^2$
O-E	$n\hat{D}_A$	$-2n\hat{D}_A$	$n\hat{D}_A$

The Goodness of Fit Statistic

Assessing the fit of observed data to expectation can be accomplished by using Equation 17.

$$\chi^2_A = \frac{(0-E)}{E}$$

Equation 17 Formula for calculating chi-square goodness of fit statistic.

where:
O = the observed data,
E = expected data.

Non-Random Mating

Two methods of non-random mating that are important in plant breeding are assortative mating and disassortative mating.

Assortative mating occurs when similar phenotypes mate more frequently than they would by chance. One example would be the tendency to mate early x early maturing plants and late x late maturing plants. The effect of assortative mating is to increase the frequency of homozygotes and decrease the frequency of heterozygotes in a population relative to what would be expected in a randomly mating population. Assortative mating effectively divides the population into two or more groups where matings are more frequent within groups than between groups.

Disassortative mating occurs when unlike or dissimilar phenotypes mate more frequently than would be expected under random mating. Its consequences are, in general, opposite those of assortative mating in that disassortative mating leads to an excess of heterozygotes and a deficiency of homozygotes relative to random mating. Disassortative mating can also lead to the maintenance of rare alleles in a population.

Factors Affecting Allele Frequency

The factors affecting changes in allele frequency can be divided into two categories: **systematic processes**, which are predictable in both magnitude and direction, and **dispersive processes**, which are predictable in magnitude but not direction. The three systematic processes are migration, mutation, and selection. Dispersive processes are a result of sampling in small populations.

Migration

Assume a population has a frequency of **m** new immigrants each generation, with **1-m** being the frequency of natives. Let $\mathbf{q_m}$ be the frequency of a gene in the immigrant population and $\mathbf{q_0}$ the frequency of the same gene in the native population. Then the frequency in the mixed population will be:

$$q_1 = mq_m + (1 - m)q_0 = m(q_m - q_0) + q_0.$$

Equation 18 Formula for calculating the frequency of an allele in a mixed population.

where:

q_1 = the frequency of the allele,
Other terms are as defined.

The change in gene frequency brought about by migration is the difference between the allele frequency before and after migration.

$$\Delta q = q_1 - q_0 = m(q_m - q_0).$$

Equation 19 Formula for calculating the change in gene frequency.

Thus the change in gene frequency from migration is dependent on the rate of migration and the difference in allele frequency between the native and immigrant populations.

Mutations

Mutations are the source of all genetic variation. Loci with only one allelic variant in a breeding population have no effect on phenotypic variability. While all allelic variants originated from a mutational event, we tend to group mutational events into two classes: rare mutations and recurrent mutations, where the mutation occurs repeatedly.

Rare Mutations

By definition, a rare mutation only occurs very infrequently in a population. Therefore, the mutant allele is carried only in a heterozygous condition and, since mutations are usually recessive, will not have an observable phenotype. Rare mutations will usually be lost, although theory indicates rare mutations can increase in frequency if they have a selective advantage.

Fate of a Single Mutation

Consider a population of only AA individuals. Suppose that one A allele in the population mutates to a. Then there would only be one Aa individual in a population of AA individuals. So, the Aa individual must mate with an AA individual, i.e., $AA \times Aa \rightarrow 1AA : 1Aa$
This mating has the following outcomes Li (1976; pp 388):

1. **No offspring are produced,** in which case the mutation is lost.
2. **One offspring is produced:** the probability of that offspring being AA is $\frac{1}{2}$, so the probability of losing the mutation is $\frac{1}{2}$.

3. **Two offspring are produced:** Aa can mate with more than one of the AA individuals in the population, thus if Aa mates with two AA individuals, the probability of both offspring being AA is $\frac{1}{4}$, so the probability of losing the mutation is $\frac{1}{4}$.

If k is the number of offspring from the above mating, then the probability of losing the mutation among the first generation of progeny is $(\frac{1}{2})^k$.

Probability of Loss

The probability of losing the gene in the second generation can be calculated by making the following assumptions:

- The number of offspring per mating is distributed as a Poisson process (which means that they follow a stochastic distribution in which events occur continuously and independently of one another).
- With the average number of offspring per mating = 2.
- New mutations are selectively neutral.

With these assumptions, the probabilities of extinction are as in Table 3:

Table 3 Probability of extinction in different generations.

Generation	Probability of Loss
1	0.37
7	0.79
15	0.89
31	0.94
63	0.97
120	0.98

Recurrent Mutations

Let the mutation frequencies be:

$$\text{Mutation Rate } \underset{p_o}{A} \overset{u}{\underset{v}{\leftrightarrows}} \underset{q_o}{a}$$

Then the change in gene frequency in one generation at equilibrium is determined using Equation 20, where $pu = qv$, and $q = \dfrac{u}{v+u}$.

$$\Delta q = up_0 - vq_0.$$

Equation 20 Formula for calculating the change in gene frequency due to mutation.

where:
u = the rate of mutation of A to a allele,
v = the rate of mutation of a to A allele
Other terms are as defined.

Conclusions

- Mutations alone produce very slow changes in allele frequency.
- Since reverse mutations are generally rare, the general absence of mutations in a population is due to selection.

Selection

Selection is one of the primary forces that will alter allele frequencies in populations. **Selection** is essentially the differential reproduction of genotypes. In population genetics, this concept is referred to as **fitness** and is measured by the reproductive contribution of an individual (or genotype) to the next generation. Individuals that have more progeny are more fit than those who have less progeny because they contribute more of their genes to the population.

The change in allele frequency following selection is more complicated than for mutation and migration because selection is based on phenotype. Thus, calculating the change in allele frequency from selection requires knowledge of genotypes and the degree of dominance with respect to fitness. Selection affects only the gene loci that affect the phenotype under selection—rather than all loci in the entire genome—but it also would affect any genes that are linked to the genes under selection.

Effects of Selection

Change in allele frequency: The strength of selection is expressed as a coefficient of selection, s, which is the proportionate reduction in gametic output of a genotype compared to a standard genotype, usually the most favored. Fitness (relative fitness) is the proportionate contribution of offspring.

Partial selection against a completely recessive allele: To see how the change in allele frequency following selection is calculated, consider the case of selection against a recessive allele:

Table 4 Initial genotypic frequencies, coefficient of selection, fitness, and gametic contribution by genotypes.

n/a	Genotypes			
n/a	**AA**	**Aa**	**aa**	**Total**
Initial Frequencies	p^2	2pq	q^2	1
Coefficient of Selection	0	0	s	n/a
Fitness	1	1	1 – s	n/a
Gametic Contribution	p^2	2pq	$q^2(1 - s)$	$1 - sq^2$

Frequency Equations

The frequency of allele a after selection is estimated using Equation 21:

$$q_1 = \frac{q^2(1-s)+pq}{1-sq^2} = \frac{q-sq^2}{1-sq^2}.$$

Equation 21 Formula for calculating the frequency of an allele following selection,

where:
s = the selection differential (represented as a deviation from the population mean),
Other terms are as defined.

The change in allele frequency is then represented as in Equation:

$$\Delta q = q_1 - q = \frac{q-sq^2}{1-sq^2} - q = \frac{sq^2(1-q)}{1-sq^2}.$$

Equation 22 Formula for calculating the change in allele frequency due to selection.

where:
terms are as defined.

In general, you can show that the number of generations, t, required to reduce a recessive from a frequency of q_0 to a frequency of q_t, assuming complete elimination of the recessive, i.e., $s = 1$ is (Equation 23):

$$t = \frac{1}{q_t} - \frac{1}{q_0}$$

Equation 23 Formula for calculating the number of generations required.

Small Population Size

Unlike the three systematic forces that are predictable in both amount and direction, changes due to small population size are predictable only in amount and are random in direction.

The effects of small population size can be understood from two different perspectives. It can be considered a sampling process, and it can be considered from the point of view of inbreeding. The inbreeding perspective is more interesting, but looking at it from a sampling perspective lets us understand how the process works.

Consequences of small population size

1. Random genetic drift: random changes in allele frequency within a subpopulation
2. Differentiation between subpopulations
3. Uniformity within subpopulations
4. Increased homozygosity

Example 1: Let q = 0.5 and N = 50, then

$$\sigma_q^2 = \frac{(0.5)(0.5)}{100} = 0.0025 \text{ , and } \quad \sigma_q = 0.05$$

Example 2: Let q = 0.5 and N = 4, then

$$\sigma_q^2 = \frac{(0.5)(0.5)}{8} = 0.03125 \text{ , and } \quad \sigma_q = 0.1768$$

Inbreeding and Small Populations

Inbreeding is the mating together of individuals that are related by ancestry. The degree of relationship among individuals in a population is determined by the size of the population. This can be seen by examining the number of ancestors that a single individual has:

Table 5 Number of ancestors of an individual in relation to the number of generations.

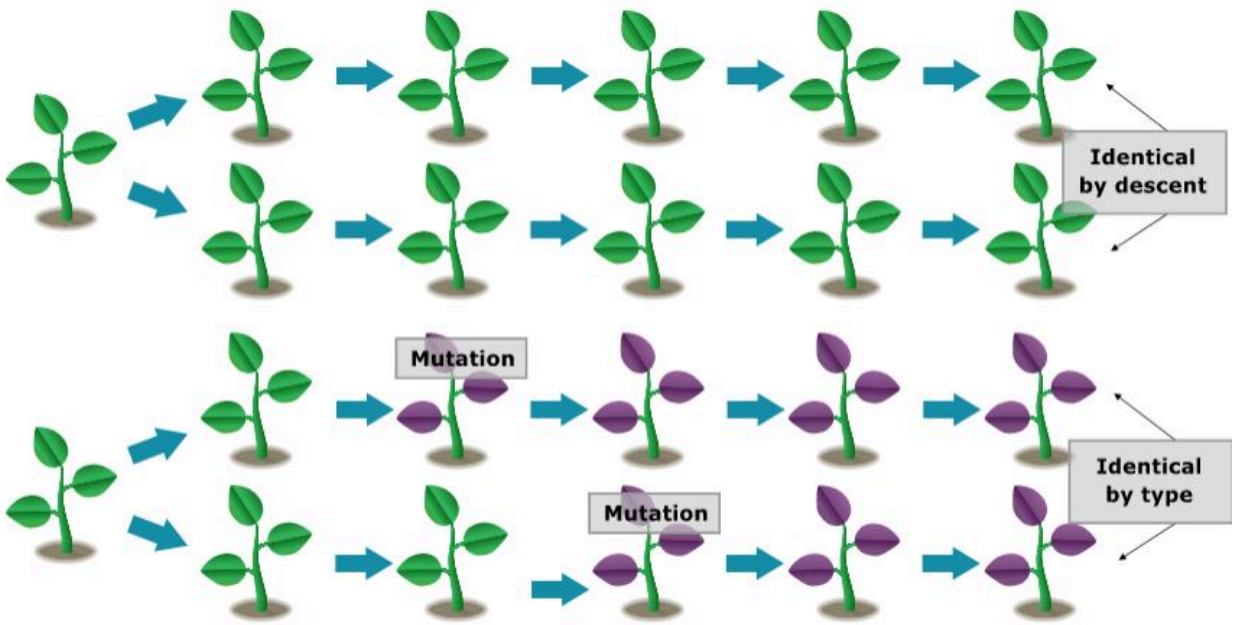
Generation	Ancestors
0	1
1	2
2	4
3	8
4	16
5	32
6	64
10	1,024
50	1,125,899,906,842,620
100	1,267,650,600,228,230,000,000,000,000,000
t	2^t

Just 50 generations ago, note that a single individual would have more ancestors than the number of people that have existed or could exist on earth.

Therefore, in small populations, individuals are necessarily related to one another. Pairs mating at random in a small population are more closely related than pairs mating together in a large population. Small population size has the effect of forcing relatives to mate even under random mating; thus, with small population sizes inbreeding is inevitable.

Identical by Descent and Identical by State

In finite populations, there are two sorts of homozygotes: Those that arose as a consequence of the replication of a single ancestral gene — these genes are said to be identical **by descent** (Bernardo, 1996). If the two genes have the same function but did not arise from the replication of a single ancestral gene, they are said to be alike **in state**. It is the production of homozygotes that are identical by descent that gives rise to inbreeding in a small population.



Coefficient of Inbreeding

The probability that two genes are identical by descent is called the **coefficient of inbreeding** and will be the measure of the relationship between mating pairs.

The coefficient of inbreeding (**F**) refers to the individual and expresses the degree of relationship between an individual's parents. The coefficient of inbreeding is always expressed relative to a specified base population. The reference population is assumed to be non-inbred (F=0).

Consider a base population consisting of N individuals, each shedding equal numbers of gametes uniting at random. Because the base is non-inbred, each individual in this population carries genes that are non-identical. The only way a homozygote that carries genes that are identical by descent can arise is by the mating of a male and female gamete from the same individual that carries a replication of the same gene. Because there are 2N gametes, the probability that two mating gametes are identical by descent is $\frac{1}{2N}$.

Equation of Coefficient of Inbreeding

In the second generation, there are two ways genes are identical by descent can be joined:

1. by a new replication of the same ancestral gene; and
2. by the previous replication that occurred in generation 1.

The probability of a new replication event is $\frac{1}{2N}$. The remaining proportion of zygotes, $1 - \frac{1}{2N}$, carry genes that are independent in origin from generation 1 but may have been identical in their origin in generation 0. The probability that the genes are identical by descent from generation 1 is the inbreeding coefficient of generation 1 is $F_1 = \frac{1}{2N}$.

Therefore, the probability of identical homozygotes in generation 2 is represent in Equation 23:

$$F_2 = \frac{1}{2N} + (1 - \frac{1}{2N})F_1,$$

Equation 23 Formula for calculating the probability of identical homozygotes.

where:
F_1 *and* F_2 = the inbreeding coefficients of generations 1 and progeny generation (PG),
N = the population size.

The same arguments apply to future generations, so we can write the recurrence equation as (Equation 24):

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}.$$

Equation 24 Formula for calculating the probability of identical homozygotes in future generation.

where:
F_t = the inbreeding coefficients of generation t,
N = the population size.

Inbreeding Coefficient

The inbreeding of any generation is composed of two components: new inbreeding, which arises from self-fertilization, and the old, which was already there.

Note that inbreeding is cumulative and that the absence of inbreeding in generation t does not change the fact that a population has inbreeding from prior generations.

Through a series of algebraic steps, we can write the inbreeding coefficient as a function of the number of generations removed from the reference populations (Equation 25):

$$F_1 = 1 - (1 - \Delta F)^t,$$

Equation 25 Formula for calculating the inbreeding coefficient.

where:

$\Delta F = \frac{1}{2N}$ is the change in inbreeding coefficients.

Dispersion

To relate inbreeding back to population size, we can rewrite the variance of the change in allele frequency $\sigma^2_{\Delta q} = \frac{p_0 q_0}{2N}$ as (Equation 26):

$$\sigma^2_{\Delta q} = p_0 q_0 \Delta F,$$

Equation 26 Alternative formula for calculating the variance of the change in allele frequency.

And also represent the variance of the allele frequency as in Equation 27,

$$\sigma^2_q = p_0 q_0 F_t.$$

Equation 27 Alternative formula for calculating the variance of the allele frequency.

Δ expresses the rate of dispersion and F expresses the amount of dispersion.

Changes in Frequencies

The genotype frequencies in a population can then be expressed as:

Table 6 Contribution of inbreeding coefficient F on genotypic frequencies for a two-allele locus case (Falconer and Mackay, 1996, p62).

n/a	n/a	n/a	Origin	
n/a	**Original Frequencies**	**Change due to inbreeding**	**Independent**	**Identical**
AA	p_0^2	$+p_0 q_0 F$	$= p_0^2(1 - F)$	$+p_0 F$
Aa	$2p_0 q_0$	$-2p_0 q_0 F$	$= 2p_0 q_0(1 - F)$	
Aa	q_0^2	$+p_0 q_0 F$	$= q_0^2(1 - F)$	$+q_0 F$

The algebra summarizes what is expected to happen "asymptotically". In any given breeding population, the results will vary due to sampling.

References

Bernardo, R. 1996. Best Linear Unbiased Prediction of Maize Single-Cross Performance Given Erroneous Inbred Relationships. Crop Sci. 36:862-866.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Pearson, Burnt Mill, England.

How to cite this chapter: Beavis, W., K. Lamkey, and A. A. Mahama. 2023. Gene Frequencies. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 2: Linkage

William Beavis and Anthony Assibi Mahama

Plant breeding populations, by definition, employ methods that force populations into states of disequilibrium. Plant breeders do not mate infinite (or even large) numbers of parents; thus, drift has a major impact on population disequilibrium. They select the parents that will be used in mating; thus, selection, linkage, and pleiotropy affect the population structure. New lines from external breeding projects are often introduced to the breeding nurseries; thus, migration affects the structure of plant breeding populations. After the passage of the Plant Variety Protection Act, plant breeders working in the commercial sector began to keep breeding records for purposes of protecting intellectual property. An unintended consequence has been the application of linear mixed models to produce predictors of performance, originally developed by animal breeders. These methods are predicated on the use of coefficients of relationship among cultivars with known performance and progeny with unknown or limited information on performance.

Herein we introduce gametic and linkage disequilibrium as measures of deviation (disequilibrium) from Hardy-Weinberg Equilibrium. In other words, the estimation of these population parameters is based on a reference population, and the reference population must be defined, or else the calculated values have no meaning.

Learning Objectives

- Demonstrate understanding that linkage and linkage disequilibrium are properties of populations, not individuals.
- Distinguish gamete from linkage disequilibrium.
- Demonstrate ability to estimate recombination and disequilibrium statistics.

Disequilibrium

The motivation is to 'map' genetic loci based on how they are most likely to be inherited relative to each other. If alleles at two loci are on the same chromosome in close proximity to each other, then they will be inherited together more often than not. It was recognized in the 1920s (Sax, 1923) that markers could have value for selecting phenotypes that are difficult to assay, but 60 years passed before the theory could be evaluated on a genome-wide scale. Linkage represents a mechanism that results in Disequilibrium among alleles at more than a single locus on the same chromosome. It is also possible that Disequilibrium among alleles at more than a single locus

can result from mechanisms other than linkage, e.g., selection and drift. Unfortunately, the term "linkage disequilibrium" has been applied to all forms of multi-locus disequilibrium. Herein we try to use the term "linkage disequilibrium" only for cases where we know alleles are on the same chromosome and "gametic disequilibrium" for situations when we do not know whether the loci are on the same chromosome.

Disequilibrium Example

Consider parent 1 with genotype A1A1B1B1C1C1D1D1 and parent 2 with A2A2B2B2C2C2D2D2. Loci A, B, and C are on a homologous chromosome, and D is on a separate chromosome (Fig. 1).

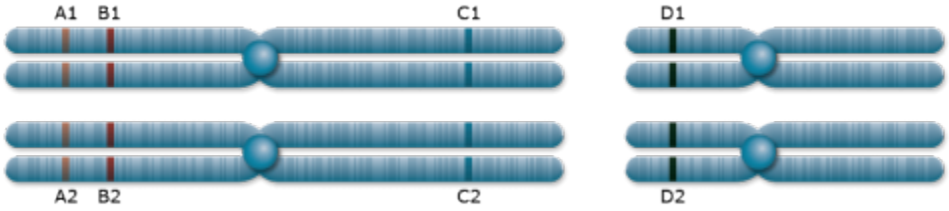


Fig. 1 A, B, C, and D loci on two pairs of homologous chromosomes.

The genotype of the F1 generation resulting from the cross between parent 1 and 2 will be A1A2 B1B2 C1C2 D1D2. Loci A and D are located on different chromosomes and will segregate independently according to the random segregation of chromosomes into gametes. For two different alleles at each locus, four possible combinations can occur, each with a chance of 25%. A and C are unlinked on the same chromosome. They are so far away from each other that recombination occurs between them in 50% of the meioses. The frequencies of all gametes involving alleles at the A and C locus (A1C1, A1C2, A2C1, A2C2) is 0.25, just as it is for the alleles for the A and D loci and the B and D loci. Since locus A and C assort independently, the frequency of double homozygous dominant and double homozygous recessive genotypes (A1A1C1C1, A2A2C2C2) is 0.25×0.25, and the frequency of double heterozygous genotypes (A1A2C1C2) is 0.5 x 0.5.

Loci A and B are linked because they are located in close proximity on the same chromosome resulting in recombination frequencies that are less than 0.5, e.g., 0.1. The difference between the expectation for unlinked loci and the estimated recombination frequency can be used to classify linkage, i.e., the likelihood of two loci being inherited together. To estimate recombination frequencies, non-parental gametes can be counted and divided by the total number of gametes.

Gametic Disequilibrium

Disequilibrium can be created by self-pollination, crossing relatives within a breeding population, mutation, drift, selection, and migration. For example, consider alleles at loci A and D. Let us assume that each contributes to phenotypic variability in flower initiation in an additive manner. Let us also assume selection for earlier flowering (conferred by the A1 and D2 alleles). The impact will be a negative covariance between the alleles at loci A and D, which reduces the genetic variances and creates disequilibrium between those loci. Even though A1 and D2 alleles are physically independent, they become correlated by selection which results in $D_{A1,\,D2} > 0$. This is also referred to as the **Bulmer effect**.

Although individual loci achieve HWE after one generation of random mating, genotype frequencies at two or more loci do not achieve equilibrium jointly after one generation of random mating.

To illustrate this point, consider two populations, one consisting entirely of AABB genotypes and the other consisting entirely of aabb genotypes. Assume they are mixed equally and allowed to mate randomly. The first generation would consist of the three genotypes AABB, AaBb, and aabb in the proportions $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$. However, for two loci with two alleles, nine genotypes are possible.

(For n alleles at each locus and k loci, there are $\left(\frac{n(n+1)}{2}\right)^k$ possible genotypes).

Continued random mating would produce the missing genotypes, but they would not appear at the equilibrium frequencies immediately.

Disequilibrium Table

Consider Table 1 below:

Table 1 Alleles and gametic types, their actual and equilibrium frequencies, and the difference between them.

Alleles	A	a	B	b
Allele Frequencies	p_A	$1 - p_A$	p_B	$1 - p_B$
n/a	n/a	n/a	n/a	n/a
Gametic Types	AB	Ab	aB	ab
Frequencies at Equilibrium	p_{AB}	$p_A(1 - pB)$	$(1 - p_A)p_{AB}$	$(1 - p_A)(1 - p_B)$
Actual Frequencies	R	S	T	U
Difference from Equilibrium	$+D_{AB}$	$-D_{AB}$	$-D_{AB}$	$+D_{AB}$

A coupling heterozygote would be $\frac{AB}{ab}$ and occur with frequency $2RU$, and the repulsion heterozygote would be $\frac{Ab}{aB}$ occurring with frequency $2ST$. If the frequency of these two genotypes is equal, the population is in equilibrium, and Equation 1 can be used to estimate the disequilibrium coefficient, D:

$$D = RU - ST = 0.$$

Equation 1 Formula for estimating D.

where:
R, S, T, U = the actual gamete frequencies.

It can be shown that after t generations of random mating, the disequilibrium is given by Equation 2:

$$D_t = D_0(1 - c)^t.$$

Equation 2 Formula for estimating D after t generations of mating.

where:
D_0, D_t = the disequilibrium in the 0 and t generations, respectively,
c = the recombination frequency, equals $\frac{1}{2}$ for independently segregating loci.

Dissipation of Disequilibrium

The dissipation of disequilibrium relative to generation 0 is given in the figure below:

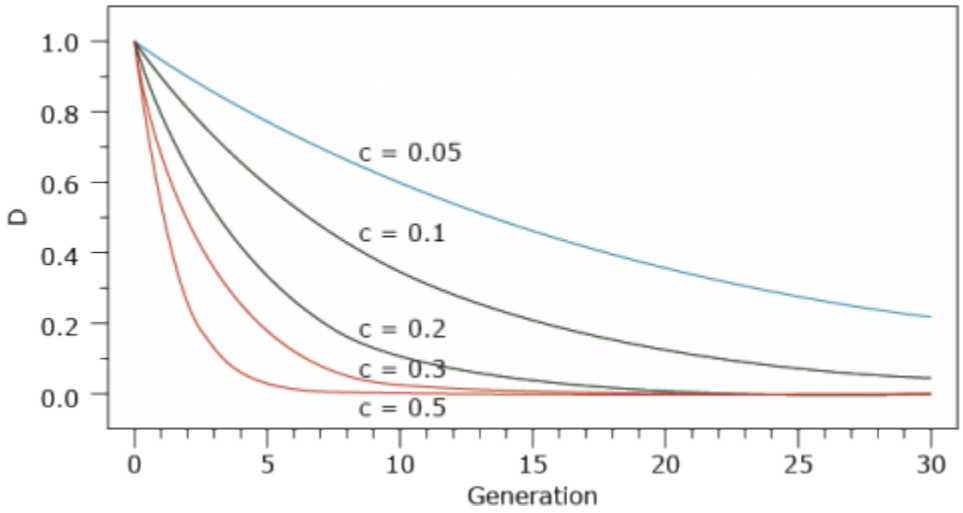


Fig. 2 Dissipation of disequilibrium.

Deviations from independence at multiple loci are often referred to as linkage disequilibrium, even if linkage is not the cause. Unless two loci are known to reside on the same chromosome the term Gametic Disequilibrium is a less ambiguous term to describe disequilibrium among loci.

Estimation and Testing

Disequilibrium at the A and B loci is a comparison of gametic frequency, p_{AB}, with the product of allele frequencies, $p_A p_B$; and is estimated with Equation 3,

$$\hat{D}_{AB} = \hat{p}_{AB} - \hat{p}_A \hat{p}_B$$

Equation 3 Formula for estimating disequilibrium at two loci.

where:
$\hat{D}_{AB}$ = the disequilibrium at loci A and B,
other terms are as defined previously.

The expectation of the estimated disequilibrium between two loci is calculated using Equation 4.

$$E(\hat{D}_{AB}) = \frac{2n-1}{2n} D_{AB}.$$

Equation 4 Formula for obtaining the expectation of estimated disequilibrium between two loci.

where:
terms are as defined previously.

The variance of the estimated disequilibrium is calculated using Equation 5.

$$Var(\hat{D}_{AB}) = \frac{1}{2n}[p_A(1-p_A)p_B(1-p_B) + (1-2p_A)(1-2p_B)D_{AB} + D_{AB}^2].$$

Equation 5 Formula for obtaining the variance of estimated disequilibrium between two loci.

where:
terms are as defined previously.

Note the similarities to $\hat{D}_A$. Thus, the distribution of estimated disequilibrium between two loci approaches a normal distribution (Equation 6).

$$\hat{D}_{AB} \ \tilde{N}[E(D_{AB}),\ Var(\hat{D}_{AB})]$$

Equation 6 Normal distribution equation of $\hat{D}_{AB}$.

where:
terms are as defined previously.

The Z statistic can be obtained using Equation 7;

$$Z = \frac{D_{AB} - E(D_{AB})}{\sqrt{Var(D_{AB})}}.$$

Equation 7 Formula for calculating Z statistic for.

where:
terms are as defined previously.

Chi-Square Statistic

Again, a chi-square statistic for the hypothesis of no disequilibrium can be calculated using Equation 8 and Table 2.

$$Z^2 = \chi^2_{AB} = \frac{2nD^2_{AB}}{p_A(1-p_A)p_B(1-p_B)}.$$

Equation 8 Formula for calculating chi-square statistic,

where:

terms are as defined previously.

Table 2 Arrangement of gametic types, their observed and expected counts for calculating chi-square statistic.

Gamete	**AB**	**AB**	**AB**	**AB**
Observed	$n_A B$	$n_A b$	$n_a B$	$n_A B$
Expected	$2n\hat{p}_A\hat{p}_B$	$2n\hat{p}_A\hat{p}_b$	$2n\hat{p}_a\hat{p}_B$	$2n\hat{p}_a\hat{p}_b$

References

Sax, K. 1923. The association of size differences with seed-coat pattern, and pigmentation in Phaseolus vulgaris. Genetics, 8, 552–560.

How to cite this chapter: Beavis, W. and A. A. Mahama 2023. Linkage. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 3: Resemblance Between Relatives

William Beavis; Kendall Lamkey; and Anthony Assibi Mahama

Plant breeding populations, by definition, employ methods that force populations into states of disequilibrium. Plant breeders do not mate infinite (or even large) numbers of parents; thus, drift has a major impact on population disequilibrium. They select the parents that will be used in mating; thus, selection linkage and pleiotropy affect the population structure. New lines from external breeding projects are often introduced to the breeding nurseries, thus migration affects the structure of plant breeding populations. After the passage of the Plant Variety Protection Act, plant breeders working in the commercial sector began to keep breeding records for purposes of protecting intellectual property. An unintended consequence has been the application of mixed linear models to produce predictors of performance, originally developed by animal breeders. These methods are predicated on the use of coefficients of relationship among cultivars with known performance and progeny with unknown or limited information on performance.

Fig. 1 Plant breeding specimens in a lab at Makerere University in Uganda. Photo by Iowa State University.

Learning Objectives

- Utilize population genetic concepts as a foundation to understand coefficients of inbreeding, parentage, and relationship.
- Calculate coefficients of parentage and inbreeding.

Background

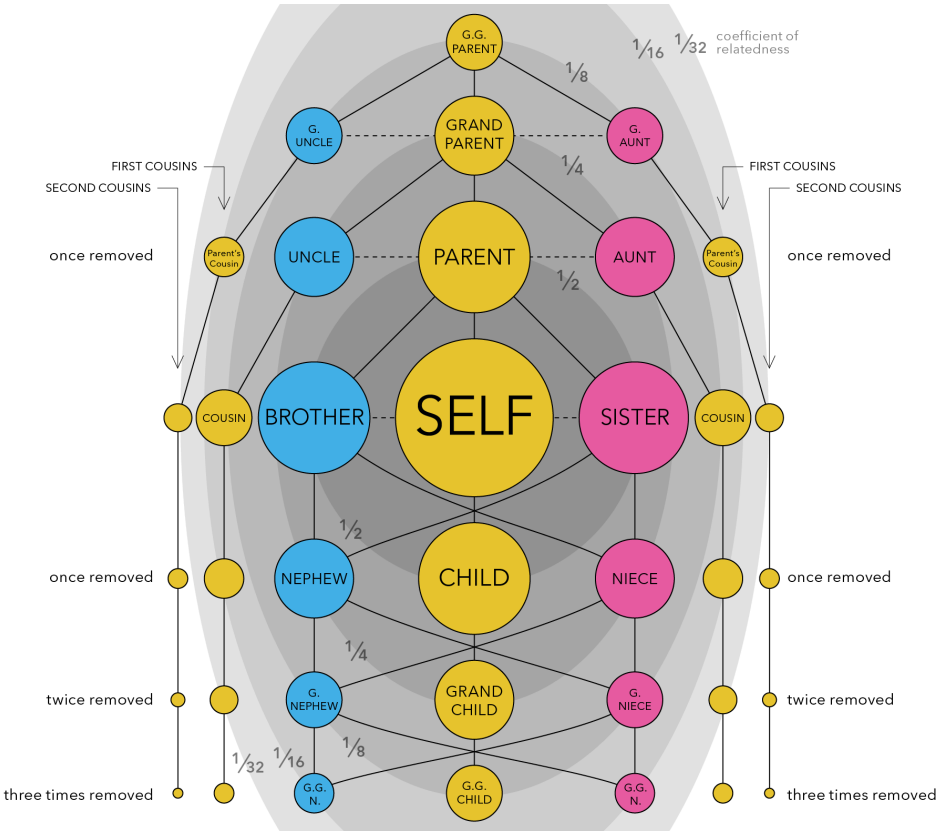


Fig. 2 Fractional relationships in the calculation of coefficients of relationships and inbreeding. CC BY-SA 4.0

The calculation of coefficients of relationships and inbreeding were originally developed as path coefficients by Sewall Wright and identity by descent by Gustave Malécot. The calculations were simplified by Emik and Terrill (1949) and extended to all possible measures of identity by Cockerham (1971). Example relatedness in humans is shown in Fig. 2.

Herein we introduce inbreeding and parentage as deviations (disequilibrium) from Hardy Weinberg Equilibrium. In other words, the calculations of all of these measures are based on a reference population and the reference population must be defined or else the calculated values have no meaning.

Coefficient of Inbreeding

Let us consider a random mating diploid population consisting of N individuals: Because there

are $2N$ gametes, the probability that two mating gametes are identical by descent is $\frac{1}{2N}$. Therefore, $F_1 = \frac{1}{2N}$. The remaining proportion of zygotes $1 - \frac{1}{2N}$ carry genes that are independent in origin from generation 1. Therefore, the probability of identical homozygotes in generation 2 is represented by Equation 1:

$$F_2 = \frac{1}{2N} + (1 - \frac{1}{2N})F_1,$$

Equation 1 Formula for calculating the probability of identical homozygotes in generation 2.

where:
F_1, F_2 = the inbreeding coefficients of generations 1 and 2,
N = the number of individuals in the population.

where F_1 and F_2 are the inbreeding coefficients of generations 1 and 2. The same arguments apply to future generations, so we can write the recurrence equation as in Equation 2:

$$F_t = \frac{1}{2N} + (1 - \frac{1}{2N})F_{t-1}.$$

Equation 2 Formula, the recurrence equation, for calculating the probability of identical homozygotes in generation t,

where:
F_t = the inbreeding coefficients of generations t,
N = the number of individuals in the population.

The inbreeding of any generation is composed of two components: new inbreeding, which arises from self-fertilization, and the “old” that was already there.

Note that inbreeding is cumulative and that the absence of inbreeding in generation t does not change the fact that a population may be inbred relative to prior generations.

General Principle

Rather than considering a random mating population, let’s consider a population that is experiencing a systematic inbreeding process. In this case, F refers to the proportionate reduction in heterozygosity (relative to a population that is in HWE) through inbreeding processes. For example, let us consider self-pollination. Begin with an F_1 from a cross of two

homozygous lines. We can self the F_1 to get an F_2. How about if we random mate F_1? Does this create a population in HWE? What is the reference population?

If F is the proportionate decrease in heterozygosity due to an inbreeding process, then with self-pollination F can be easily calculated for any generation of selfing as shown in Equation 3:

$$F_n = 1 - (\frac{1}{2})^{n-2}.$$

Equation 3 Formula for calculating the proportionate reduction in heterozygosity.

where:
n = the generation of interest.

Impact on Disequilibrium

The impact on deviations, i.e., disequilibrium, relative to HWE can be summarized as:

	HWE Frequencies	**Change due to inbreeding.**
AA	p_0^2	$+p_0 q_0 F$
Aa	$2p_0 q_0$	$-2p_0 q_0 F$
Aa	q_0^2	$+p_0 q_0 F$

Alternatively, we can think of the coefficient of inbreeding as the probability of identity by descent. In this case, the coefficient of inbreeding is the probability that two alleles at a locus in an individual are IBD. For two individuals X_{ab} and Y_{cd}, the relationship is represented in Equation 4.

$$F_x = P(a \equiv b), \text{ and } F_y = P(c \equiv d)$$

Equation 4 Formula for calculating the coefficient of inbreeding (equivalent to IBD).

where:
F_x = the coefficient of inbreeding of **x**,
F_y = the coefficient of inbreeding of **y**,
$P(a \equiv b), P(c \equiv d)$ = probability that **a** and **b** and **c** and **d** are IBD.

Coefficient of Parentage

What if two homozygous parents of an F_1 used to create an F_2 population are related? Let us think about the relationship, parentage, and co-ancestry between two individual people, dogs, corn plants, soybean plants, etc. Refer to these individuals as X and Y. Also, let us use a shorthand for a quantitative measure of this relationship. This relationship is also known as the coefficient of parentage and is defined as the probability that a random gene from an individual X is identical by descent (IBD) with a random allele at the same locus from an individual Y. That is, for **Xab and Ycd** , the probability of identity by descent is presented in Equation 5.

$$r_{xy} = \frac{1}{4}[P(a \equiv c) + P(a \equiv d) + P(b \equiv c) + P(b \equiv d)].$$

Equation 5 Formula for determining IBD of genes in individuals X and Y.

where:
r_{xy} = the probability that alleles in X and Y are identical by descent,
other terms = are as defined previously.

Historically, this measure has been denoted $\Theta_{X,Y}$ or $f_{X,Y}$. The inbreeding coefficient of the progeny is the coefficient of parentage of the parents.

Calculations

$\Theta_{X,Y}$ = 1 means that X and Y have the same identical alleles by descent across all loci. What is another name for this condition? (twins). $\Theta_{X,Y}$ = 0 means what? Is it possible that you and I have no alleles that are identical by descent?

There is a relationship between F_n and $\Theta_{X,Y}$. In an individual, if two alleles at a single locus are identical by descent, then this is a special case of $\Theta_{X,Y}$, where X and Y are the same individual, i.e., $F_X = \Theta_{X,X}$. To return to the original question "What if two homozygous parents of an F_1 used to create an F_2 population are related?" The relationship is represented by Equation 6.

$$F_n = 1 - (\frac{1}{2})^{n-2}\{1 - \theta_{X,Y}\}.$$

Equation 6 Formular for calculating the coefficient of inbreeding in generation n.

where:
$\theta_{X,Y}$ = the coefficient of parentage of the parents.

Inbreeding Coefficient

Consider the following pedigree:

$$(X_{ab})|Y_{cd}) \implies Z \quad \text{i.e., Z is a progeny of the mating between X and Y.}$$

Individual Z has the following probabilities of containing the various alleles (Equation 7):

$$\frac{1}{4}ac + \frac{1}{4}ad + \frac{1}{4}bc + \frac{1}{4}bd,$$

Equation 7 Formula for the probability of individual Z containing various combinations of alleles.

where:
$F_z = r_{xy}$ and are as defined previously.

Example Calculations

What is the probability that the two mating gametes A1A2 x A3A4 at locus A are identical by descent in the F_1? Assume here that the two parents are not related; that is,
$A_1A_2 \times A_3A_4 \rightarrow F_1$.

$$r_{xy} = \frac{1}{4}[P(a \equiv c) + P(a \equiv d) + P(b \equiv c) + P(b \equiv d)].$$

$$r_{F_1} = \frac{1}{4}[\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}] = \frac{1}{2} = 0.5.$$

The probability that the gametes are identical by descent in the F_1 = 0.5

Self Pollination

The relationship expressed in Equation 7 can be applied in determining the coefficient of inbreeding following self-pollination as in Equation 8. That is,$X_{ab} \times X_{ab} \rightarrow Z$.

$$F_z = r_{xx} = \frac{1}{4}[P(a \equiv a) + P(a \equiv b) + P(b \equiv a) + P(b \equiv b)],$$

$$P(a \equiv a) = P(b \equiv b) = 1,$$

$$P(a \equiv b) = P(b \equiv a) = F_x,$$

$$F_z = r_{xx} = \frac{1}{4}[2 + 2F_x],$$

$$= \frac{1}{2}[1 + F_x].$$

Equation 8 Alternative formula for calculating Fz.

where:
F_x = the coefficient of inbreeding of individual X.

Panmictic Index

Panmictic Index, **P**, is the probability that two alleles at a locus are not IBD and is related to F as, P = 1 – F. We can use the known equations and relationships in the earlier section and with substitution of P for the n^{th} generation we can calculate P using Equation 9.

$$1 - P_z = \frac{1}{2}[1 + 1 - P_x],$$

$$P_z = \frac{1}{2}P_x,$$

$$P_1 = \frac{1}{2}P_0,$$

$$P_2 = \frac{1}{2}P_1 = \frac{1}{2} \cdot \frac{1}{2}P_0 = (\frac{1}{2})^2 P_0,$$

$$P_n = (\frac{1}{2})^n P_0.$$

Equation 9 Formula for calculating the panmictic index.

where:
P_n = the panmictic index in generation n,

P_z = the panmictic index of alleles in z.

P_0 = the panmictic index in generation 0,

P_x = the panmictic index of alleles in X.

For diploids this is also the percent of heterozygotes at a locus

Full-Sib Mating (1)

Fig.3 and schematic of Full-sib mating design, from generation n to generation n+ 2

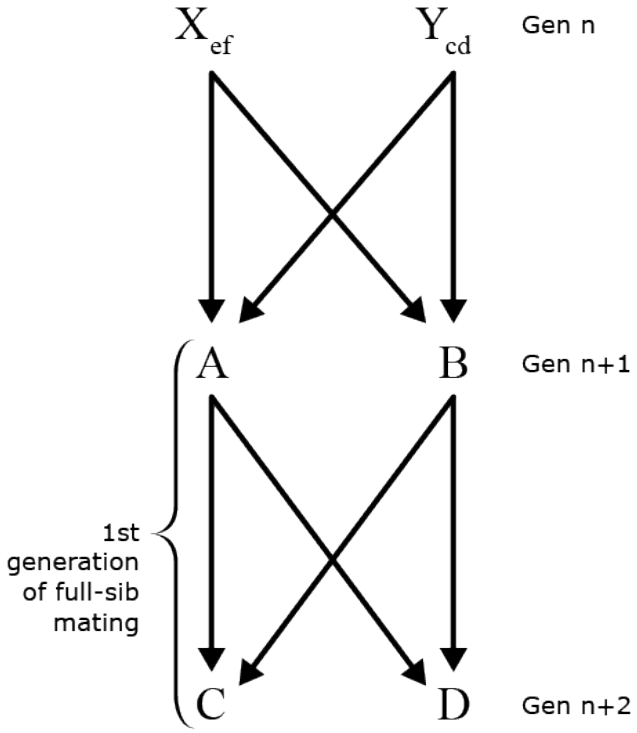


Fig. 3 Full-sib mating design.

Probability that A & B both receive ***e*** from X =

$$(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$$

Probability that A & B both receive ***f*** from X =

$$(\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$$

Probability that ***e*** and ***f*** are not IDB =

$$1 - F_x$$

Probability that A & B contain an identical allele from X by chance (given that ***e*** and ***f*** are not IBD) =

$$(\frac{1}{4}+\frac{1}{4})(1-F_x)$$

Equation 10 Formula for calculating probability that an allele in A and B form parent X is IBD knowing that alleles e and f in X are not IBD.

where:
F_x = the coefficient of inbreeding of X, i.e., the probability that e and f are IBD.

Full-Sib Mating (2)

Probability that ***e*** and ***f*** are identical =

$$F_x$$

Probability A & B receive an identical allele from X (given that ***e*** and ***f*** are IBD) = 1

Total probability that A & B receive an identical allele from X, and from Y can be determined using Equation 11.

$$F_x+\frac{1}{2}(1-F_x)=\frac{1}{2}(1+F_x)=r_{xx},\ and\ F_y+\frac{1}{2}(1-F_y)=\frac{1}{2}(1+F_y)=r_{yy}$$

Equation 11 Formula for calculating the probability that progenies A and B received identical alleles from parents X and Y,

where:
r_{xx}, r_{yy} = the probability of identical alleles from either parent,
F_x = the coefficient of inbreeding of **x**,
F_y = the coefficient of inbreeding of **y**.

Full-Sib Mating (3)

Let us consider full-sib mating where the relationship between genes in offspring from the two parents can be represented by Equation 12.

$$r_{AB}=\frac{1}{4}[r_{xx}+2r_{xy}+r_{yy}]$$

Equation 12 Formula for calculating the probability that a gene from parent X to progeny A, and from parent Y to progeny B are IBD.

where:
terms are as described below.

Probability that a gene from X to A and one from Y to B are IBD is r_{xy}.
Probability that a gene from Y to A and one from X to B are IBD is r_{xy}.

Full-Sib Mating (4)

The relationship between X and Y, rxy, could be zero if the reference population from which X and Y are sampled is considered to be random mating and large. Note that if the population is random mating but is not large, then the relationship coefficient may not be zero. The relationship can be represented as in Equation 13.

$$F_{xy} = F_y = F_n, \; and \; r_{xy} = r_n$$

Equation 13 Formula for calculating the relationship between two parents.

where:
terms have been defined previously.

In the form represented in Equation 13, notice that the relationship between X and Y is equal to the average relationship in the n^{th} generation of random mating and the series of equations in Equation 14 shows the relationships in progression.

$$r_{n+1} = \frac{1}{4}[r_{xx} + r_{yy} + 2r_{xy}]$$

$$= \frac{1}{4}[\frac{1}{2}(1 + F_n) + \frac{1}{2}(1 + F_n) + 2r_n]$$

$$r_{n+1} = F_{n+2}$$

$$r_n = F_{n+1}$$

$$F_{n+2} = r_{n+1} = \frac{1}{4}[\frac{1}{2}(1 + F_n) + \frac{1}{2}(1 + F_n) + 2F_{n+1}]$$

$$= \frac{1}{4}[1 + F_n + 2F_{n+1}]$$

$$P_{n+2} = \frac{1}{2}P_{n+1} + \frac{1}{4}P_n$$

$$F_n = 1 - (\frac{1}{2})^n (1 - F_0)$$

Equation 14 Formula for calculating the relationship between two parents in different generations of random mating.

where:

terms are as defined previously.

Self-Pollination

Assume original population is non-inbred (by definition = F_2). Using Equation 9 the relationship between the panmictic index and the coefficient of relationship can be calculated for different generations of self pollination assuming the original population is non-inbred (by definition = F2) (Table 2). Where:

$$P_n = (\frac{1}{2})^n P_0, \ and \ F = \frac{1}{2}[1 + F_x].$$

Table 2 Relationship between panmictic index (P) and coefficient of relationship (F) with self-poolination and F_2 as original population.

Generation	P	F
0	1.00000000	0.00000000
1	0.50000000	0.50000000
2	0.25000000	0.75000000
3	0.12500000	0.87500000
4	0.06250000	0.93750000
5	0.03125000	0.98437500
6	0.01562500	0.98437500
7	0.00781250	0.99218750
8	0.00390625	0.99609375
9	0.00195313	0.99804688
10	0.00097656	0.99902344
∞	0.00000000	1.00000000

Full-Sibing

In a similar manner, the relationship between the panmictic index and the coefficient of relationship can be calculated for different generations for full-sib mating situation as shown in Table 3, where:

$$P_{n+2} = \frac{1}{2}P_{n+1} + \frac{1}{4}P_n$$

Table 3 Relationship between panmictic index (P) and coefficient of relationship (F) with full-sib mating and F_2 as original population.

Generation	P	F
0	1.00000000	0.00000000
1	1.00000000	0.00000000
2	0.75000000	0.25000000
3	0.62500000	0.37500000
4	0.50000000	0.50000000
5	0.40625000	0.59375000
6	0.32812500	0.73437500
7	0.26562500	0.78515625
8	0.21484375	0.78515625
9	0.17382813	0.82617188
10	0.14062500	0.85937500
∞	0.00000000	1.00000000

References

Emik, l. O., and C. E. Terrill. 1949. Systematic procedures for calculating inbreeding coefficients. J. Heredity, 40 (2): 51–55.

Cockerham, C. C. 1971. Higher order probability functions of identity of alleles by descent. Genetics 69:235–246.

How to cite this chapter: Beavis, W., K. Lamkey, and A. A. Mahama. 2023. Resemblance Between Relatives. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 4: Measures of Similarity

William Beavis; Mark Newell; and Anthony Assibi Mahama

In an ideal reference breeding population, there is no structure consisting of sub-populations or aggregates of relatives organized into families and tribes. Plant Breeding populations, on the other hand, are organized into sub-populations. Perhaps the best-known example is represented by the heterotic germplasm pools in maize, e.g., Stiff Stalks, Non-Stiff Stalks, Lancasters, and Iodents. In cytoplasmic male sterile hybrid systems such as sorghum, the restoration pattern can be the primary divider of germplasm with additional subdivisions based on morphological characteristics and geographic origins, e.g., Kaoliang, Durra, and Feterita. Alternatively, coefficients of relationship and inbreeding among members of a breeding population can be used to represent the structure of the breeding population. Also, with the emergence of high throughput molecular marker technologies, it is possible to represent relationships among members of a breeding population using identity in state to produce a realized kinship matrix.

Learning Objectives

- Utilize coefficients of inbreeding and parentage to construct the numerator relationship matrix
- Utilize molecular marker information to construct a realized kinship matrix

Population Structure Based on Pedigree Information

Animal breeders were the first to utilize relationships among individuals for the purpose of providing Best Linear Unbiased Predictions in linear mixed models. The "A" matrix in the linear mixed model equation, also known as the Numerator Relationships Matrix (NRM) was originally used by Henderson to capture information from relatives to predict breeding values of animals. In essence, the A-matrix provides information on the proportion of alleles that are identical by descent between all pairs of individuals in a breeding population.

Specifically, the numerator relationships are equal to twice the coefficient of coancestry between any pair of individuals. In other words, $A_{x,y} = 2\Theta_{x,y}$. Thus, if we know the pedigrees of all members of a breeding population, we can construct an A-matrix using a recursive tabular method.

Recursive Tabular Method

A recursive tabular method for constructing the A-Matrix is described below:

1. Order members of a pedigree chronologically, i.e., list parents before offspring. Assume that founder lines are not inbred and are not related to each other.
2. Transpose the list and use this to represent columns for the A matrix.
3. Beginning with the cell represented by $A_{1,1}$ compute $\Theta_{1,1}$.
4. Move to cell $A_{1,2}$ and compute $\Theta_{1,2}$. This will be the same value that can be used for cell $A_{2,1}$
5. Move to cell $A_{2,2}$ and compute $\Theta_{2,2}$.
6. Move to cell $A_{1,3}$ and compute $\Theta_{1,3}$. This will be the same value for $A_{3,1}$
7. Move to cell $A_{2,3}$ and compute $\Theta_{2,3}$. This will be the same value for $A_{3,2}$
8. Move to cell $A_{3,3}$ and compute $\Theta_{3,3}$
9. Repeat until all elements of the A matrix are completed.

Population Structure Based on Markers

The Realized Kinship Matrix

Consider two cultivars scored for 1400 SNPs. We can ask whether this pair of cultivars has the same or different alleles at each locus. Intuitively, if they had the same allele at all 1400 loci, we would say that there are no detectable allelic differences between the two genotypes, i.e., that they are identical in state or that their similarity index = 1.0. Alternatively, if none of the alleles are the same at all 1400 loci, then we would say that the genotypes have no alleles in common, i.e., that their similarity index is zero. In practice, the two genotypes will exhibit a measure of similarity somewhere between these extremes.

Quantitative Measure for Similarity

Let us take this intuition and develop a quantitative measure for similarity. If the two cultivars (x and y) have the same pair of alleles at a locus, score the locus = 2; if one of the alleles is the same, score the locus = 1; otherwise, the score = 0. If we sum these up across all loci, the maximum score would be 2800. If we divide the summed score by 2800, we would obtain a proportion measure (designated $S_{x,y}$) to quantify the similarity between the pair of lines. This concept can be represented algebraically as:

$$S_{x,y} = \frac{1}{2n} \sum_{i=1}^{n} X_i Y_i$$

Equation 1 Formula for calculating the similarity between pairs of lines.

where:
n = the number of loci,
X, Y = the two cultivars.

Such a similarity measure could be converted into an "intuitive genetic distance" measure by subtracting $S_{x,y}$ from 1.

Measures of Distance

Our intuitive genetic distance would make sense if:

1. There are only two alleles per locus.
2. Our interpretation of the result does not include inferences about identity by descent, and
3. There is no LD among the SNP loci.

However, most populations are more complex, requiring more nuanced measures of genetic distance. Population geneticists tend to use three distance measures depending upon the inference about the population structure they are trying to understand. These are:

- **Nei's Distance** assumes all loci have the same neutral rate of mutation, mutations are in equilibrium with genetic drift, and the effective population size is stable. The interpretation is a measure of the average number of changes per locus and that differences are due to mutation and genetic drift.
- **Cavalli-Sforza's Distance** assumes differences are due to genetic drift between populations with no mutation and interprets the genetic distance as an Euclidean Distance metric.
- **Reynolds Distance** is applied to small populations; thus, it assumes differences are due to genetic drift and is based on knowledge about coancestry, i.e., identity by descent for alleles that are the same.

Application of Distance and Similarity Measures

There are a large number of additional distance and similarity measures that can be applied to molecular marker scores, including Euclidean, Mahalanobis, Manhattan, Chebyshev, and

Goldstein. Also, Bayesian Statistical approaches can be used to identify structure in the population (Pritchard et al, 2000) without resorting to the calculation of distance metrics. The choice of an appropriate method depends upon the type of molecular marker data and the research question. A thorough presentation of distance measures is beyond the scope of this course, but there are graduate courses on multivariate statistics in which issues associated with each of the distance metrics can be explored.

For now, let us assume that we decided to use our $S_{x,y}$ to represent differences between all pairs ($x_i y_j$) of breeding lines. Next, suppose we extend the example from two lines to 1800 lines scored for 1400 SNPs. In this case, there are $\frac{n(n-1)}{2} = \frac{1,800(1,800-1)}{2}$ = 1,619,100 estimates of pairwise distances among the lines.

Clearly, any attempt to find patterns in a data matrix consisting of all pairwise measures of similarity or distance will take considerable effort. Yet, these patterns in the data are essential to quantifying the structure in a breeding population because the structure will affect inferences about genetic effects. It is the need to find patterns in such large data sets that motivated the application of multivariate statistical methods such as principal components and cluster analyses in plant breeding populations.

Principal Component Analysis

The primary purpose for applying principal component analysis (PCA) to genetic distance matrices is to summarize, i.e., reduce dimensionality so that the underlying population structure can be visualized.

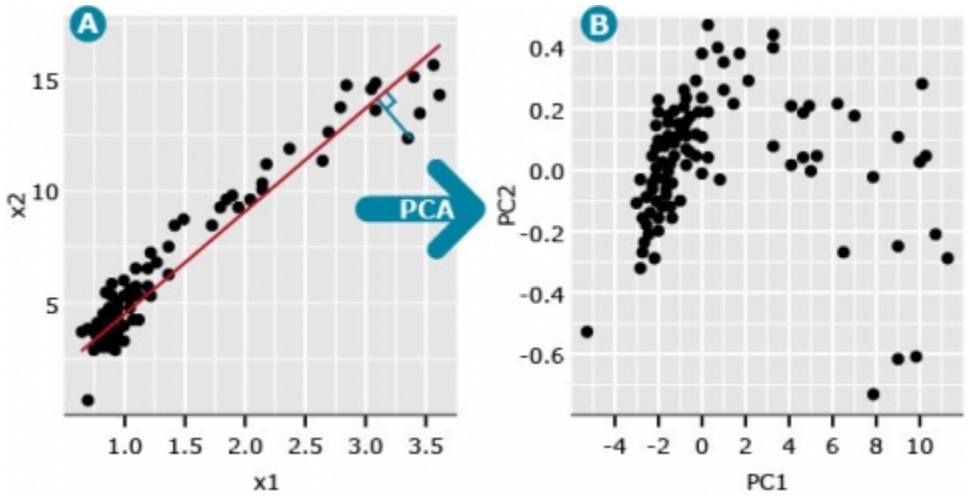


Fig. 1 Effect of principal component analysis.

Conceptual Interpretation

Imagine we have two variables, denoted x1 and x2, where x1 represents the distance scores between cultivar 1 and all other cultivars, and x2 represents the distance scores between cultivar 2 and all other cultivars. If we plot the x1 and x2 pairs of data, we might generate a plot such as seen in Fig. 1A. We could add distance data for a third cultivar and represent the data with a 3-dimensional plot. We could obtain data for as many cultivars as we might have interest in, but the ability to plot these in multi-dimensional space is not possible.

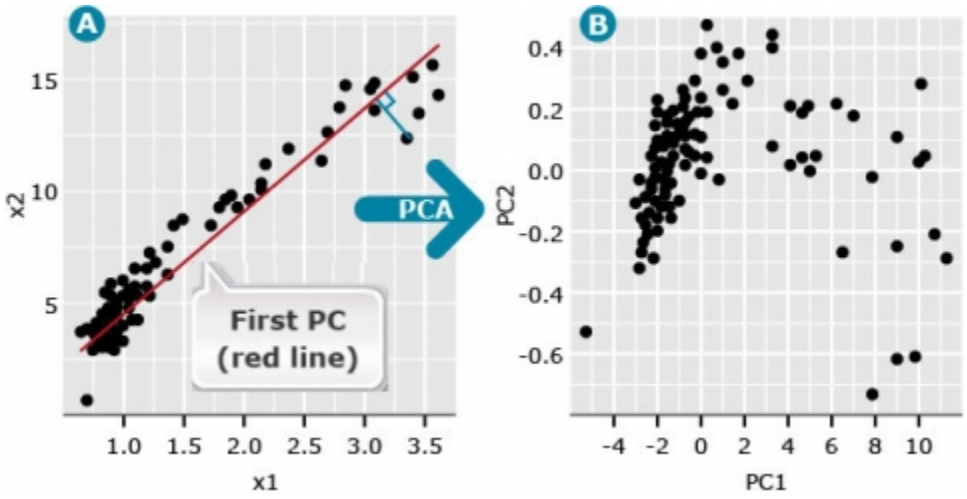


Fig. 2 Effect of principal component analysis with the first PC.

We refer to the first principal component (PC), also known as the first eigenvector, as a line (red) that minimizes the perpendicular distances (blue line) between the red line and the data points (Fig. 2A).

Principal Component Analysis – Interpretation

The second PC follows the same definition except that it represents a line through the data that minimizes the distance between a second line that is orthogonal (at a right angle) to PC1. The second PC minimizes the distance between the data and the second line. Since the second PC is orthogonal to the first, the distance among the data points represented by each PC is maximized. Thus we can plot data points represented by the first two principal components (Fig. 3B). By plotting the PCs instead of the raw data, we often find hidden structures in the data (compare Fig. 3A vs. 3B).

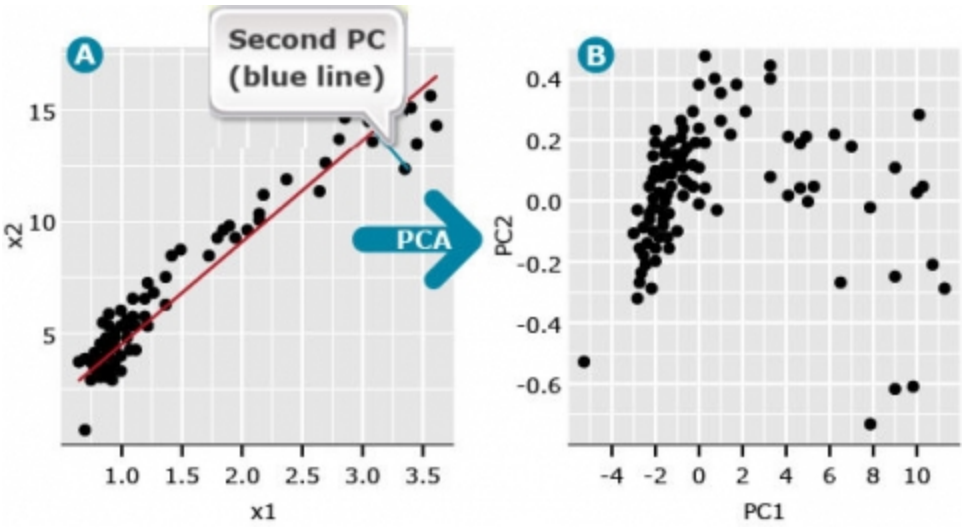


Fig. 3 Effect of principal component analysis with second PC.

Subsequent PCs represent lines that are orthogonal to all previous PCs and minimize the distance between each PC and data points that maximize the variability among the orthogonal PCs. This means that each PC is uncorrelated to all other PCs.

A useful measure in PCA is the eigenvalue associated with each eigenvector (PC). The first eigenvalue is the proportion of maximum variability among the multidimensional data that is explained by the first PC. For the data depicted in Fig. 3B, the first eigenvalue is 0.997, and the second eigenvalue is equal to 0.003. Since the first PC is the vector (or line) that is plotted in the direction of maximum variability among data points, the first eigenvalue is always the largest, and each consecutive eigenvalue accounts for less variability than the prior PCs.

PCA Example

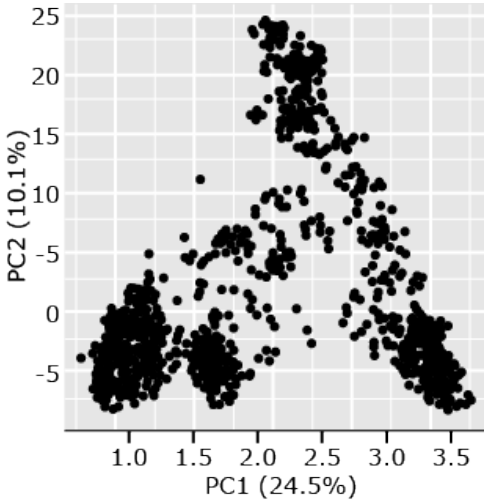


Fig. 4 Four distinct clusters produced by PCA.

Let us consider an example from a set of 1816 barley lines scored for 1416 SNPs (Hamblin et al. 2010). In this analysis, there were $\binom{1816}{2} \Rightarrow \frac{n \times (n-1)}{2}) = \frac{1816 \times (1815)}{2}) = 1,648,020$ estimates of pairwise distances based on 1416 SNP scores for each of the barley lines. Eigenvalues for PC1 and PC2 accounted for 24.5% and 10.1% of the variability among pairwise genotypic distances. By plotting PC1 versus PC2 (Fig. 4), we observe four distinct clusters. Subsequent analyses of the lines represented by each point in the clusters revealed that the members of each cluster are from 2-row, 6-row, spring, or winter barley types. From a breeding perspective, we can see that most breeding for barley occurs within types rather than between types. The population structure is a result of breeding processes of selection, drift, and non-random mating.

Cluster Analysis

Similar to PCA, the purpose of applying cluster analysis to matrices of pairwise distance measures among a set of genotypes is to segregate the observations into distinct clusters. There are many types of cluster analyses, and a primary distinction is between supervised and non-supervised clustering. K-means is one of the supervised methods that have been widely adopted by plant population geneticists. The clustering method is supervised in the sense that K represents a pre-determined number of clusters. Designating the number of clusters is usually

based on prior knowledge about groups of lines that are being clustered. For example, it might make sense to designate the four clusters of barley lines based on known breeding history in which different barley agronomic types are not inter-mated. K-means represents an iterative procedure with the following steps:

1. An initial set number of K means (seed points) are determined (also called initialization); these are the initial means for each of the K clusters.
2. Each genotype is then assigned to the nearest cluster based on its pairwise distances to all other genotypes within and among clusters.
3. Means for each cluster are then re-calculated, and genotypes are re-assigned to the nearest cluster.
4. Steps ii and iii are then repeated until no more changes occur.

Cluster Analysis Example

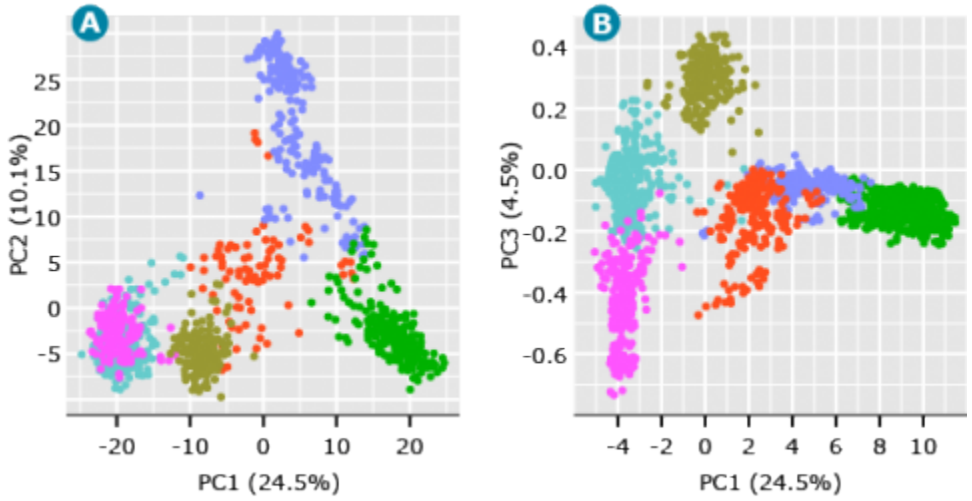


Fig. 5 PCA-produced k-means.

For the barley data, since the inter-mating rule is not absolute, i.e., some agronomic types are occasionally inter-mated, it could be informative to designate K = 6 (Fig. 5). Note that a plot of PC1 vs PC3 (Fig. 5B) demonstrates the value of plotting PCs beyond the first two. While the third PC accounts for only 4.5% of the variability among genotypes, the third PC helps to distinguish what appears to be members of the same cluster in Fig. 5A.

Hierarchical Clustering

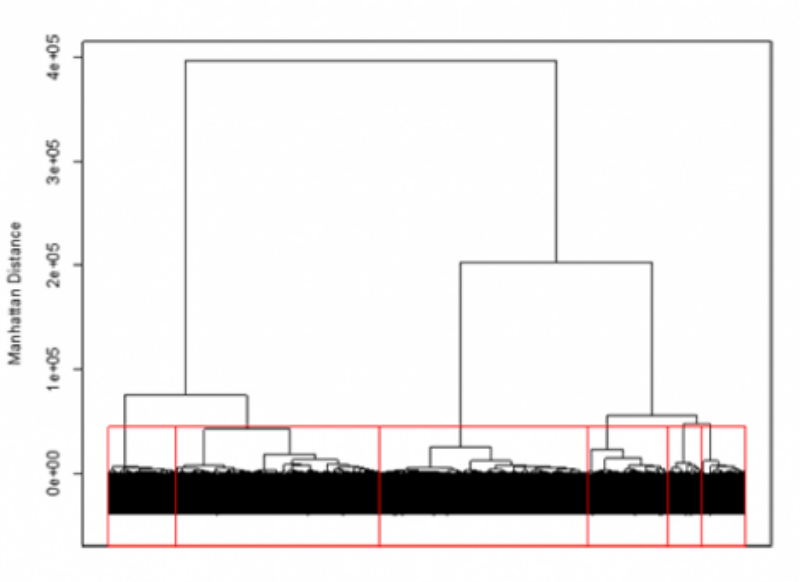


Fig. 6 Dendogram observations data

An unsupervised approach to clustering genotypic distance data is hierarchical clustering. This approach sequentially lumps or splits observations to make clusters. Applying the hierarchical approach to the barley data set, we can visualize the results using a dendrogram (Fig. 6). In the dendrogram, observations are arrayed along the x-axis, and the y-axis refers to the average genetic distance between breakpoints. For example, the horizontal line at 4e+05 indicates that there are two major groups with a distance between them of 4e+05. The user determines the height (distance along the y-axis) at which a horizontal line is drawn, and the number of clusters is chosen; this is drawn below in red for 6 clusters. The user may determine this by using the PC plots, cluster dendrogram, and any prior information that is known about the germplasm.

Hierarchical clustering can be implemented in many different ways. For genotypic data, the most common method is Ward's, which attempts to minimize the variance within clusters and maximize the variance between clusters. Similar to K-means clustering, we can look at the PC plots to explore the results for hierarchical clustering to see how the lines were assigned to clusters.

How to cite this chapter: Beavis, W., M. Newell, and A. A. Mahama. 2023. Measures of Similarity. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 5: Gene Effects

William Beavis; Kendall Lamkey; Katherine Espinosa; and Anthony Assibi Mahama

In 1918, RA Fisher provided the first major contribution to the Modern Synthesis by proposing a model that reconciled the inheritance of discrete characteristics (Mendel) and continuous, or quantitative, characteristics (Darwin) in breeding populations. Herein, the same theoretical foundations are introduced.

For the beginning student, it will seem that the primary purpose of theory and modeling is to provide interpretations of observational and experimental results. Without this theoretical foundation, there would be no genetic understanding of the results from plant breeding experiments.

However, there is a more important practical justification for learning theoretical models: Theory provides predictions. Predictions are the basis for generating testable hypotheses. Also, with a theoretical model, it is possible to simulate many different breeding strategies. These can be compared, and the most promising can be used to design and implement the most effective and efficient breeding strategies. Thus, the theory provides a rational basis for designing plant breeding programs.

Learning Objectives

- Model Genotypic and Phenotypic Values of individuals in crop breeding populations.
- Integrate genotypic effect models with allele frequency models at single and multiple loci.
- Distinguish and estimate genetic effects, effects of allele substitutions, and Breeding Values at single and multiple loci.
- Distinguish and estimate dominance and epistatic deviations from additive effect models.
- Integrate concepts to applied breeding programs with data sets consisting of genotypic (marker) information with phenotypic information for QTL analyses.

Linear Models for Phenotypic Values

Single Locus

The **phenotypic value** of an individual, or group of individuals, is observed when a character or

trait is measured. For example, if a corn plant was measured and found to be 275 cm tall, then that would be its phenotypic value for height.

To draw inferences about the genetic properties of a trait, we model phenotypic values using linear components. The most common model consists of a part due to genetics and a part due to non-genetic effects such as the environment. This is usually written as:

$$P = G + E,$$

Equation 1 Linear model for evaluating phenotype.

where:
P is the phenotypic value,
G is the genotypic value, and
E represents the non-genetic factors.

If we assume that $\sum E = 0$, then $\sum P = \sum G$, and $\bar{P} = \bar{G}$.

Population Mean

The mean phenotypic value of a population is equal to the mean genotypic value when the non-genetic (environmental) deviations sum to zero.

To calculate the expected genotypic properties of a population for a single locus, we assign arbitrary genotypic values to each locus.

Consider a single locus with two alleles A and a.

Coded genotypic value of one homozygote $AA = +a$.

Coded genotypic value of the other homozygote $aa = -a$

Coded genotypic value of a heterozygote $Aa = d$.

We can arbitrarily designate the A allele as the allele that increases the genotypic value. The genotypic value of the heterozygotes (*d*) depends on the level of dominance:

Degree of Dominance

1. No dominance when $d = 0$ (Fig. 1).

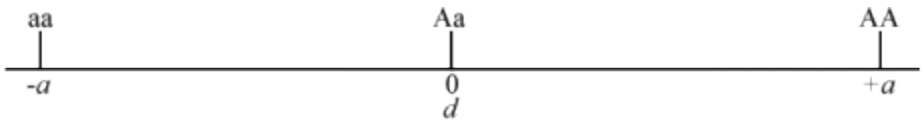


Fig. 1 Line diagrammatic representation of no dominance.

2. If A is dominant or partially dominant relative to the a allele, then d is positive towards the AA genotype, as shown in Fig. 2.

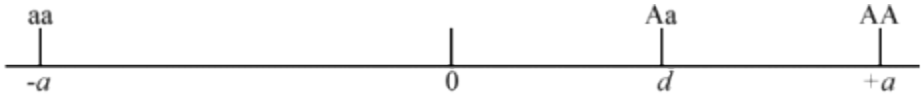


Fig. 2 Line diagrammatic representation of dominance or partial dominance of *A* over *a* allele.

3. If a is dominant or partially dominant to the A allele, then d is negative (Fig. 3).

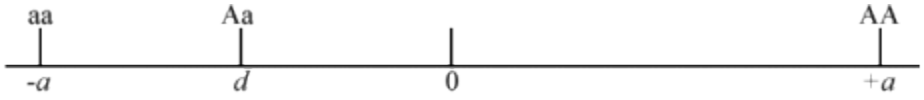


Fig. 3 Line diagrammatic representation of dominance or partial dominance of *a* over *A* allele.

4. If dominance is complete: d = +a or -a

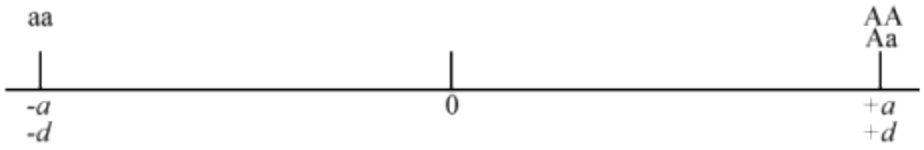


Fig. 4 Line diagrammatic representation of complete dominance of *A* over *a* allele or *a* over *A* allele.

5. If there is overdominance: d is greater than +a or less than -a

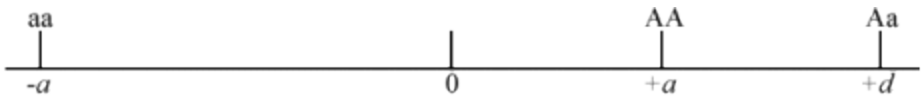


Fig. 5 Line diagrammatic representation of overdominance.

Allele Frequencies and Population Mean (Table 1)

Table 1 Influence of allele frequencies and dominance deviation on the average value of the trait in the population.

n/a	Genotype			
n/a	**AA**	**Aa**	**aa**	**Total**
Frequency	p^2	$2pq$	q^2	1
Genotypic Value	Y_{AA}	Y_{Aa}	Y_{aa}	n/a
Coded GV	$+a$	d	$-a$	n/a
Freq. x Coded GV	$+p^2a$	$2pqd$	$-q^2a$	$= a(p-q)+2pqd$

Table 2 An example of a genotype that controls the number of flowers and the expected population value for number of flowers.

n/a	Genotype			n/a
n/a	**AA**	**Aa**	**aa**	**Total**
Frequency	0.64	0.32	0.04	1
Genotypic Value	16	12	0	n/a
Coded GV	8	4	-8	n/a
Freq. x Coded GV	5.8	1.28	-0.32	14.08

Note: Coded Genotypic Values are obtained by subtracting the mid-parent value i.e., the midpoint between the genotypic values of the two homozygotes (Table 2).

Population mean $= \bar{Y} = \mu +$ Expected value of g_{ij}, and
$mh =$ mid-homozygote value.

The population mean is estimated using Equation 2.

$$\bar{Y} = mh + a(p-q) + 2pqd.$$

Equation 2 Formula for calculating population mean.

where:
p = frequency of ***A*** allele,
q = frequency of ***a*** allele,
a = coded genotypic values of **AA**, **aa** genotypes,
d = coded genotypic value of **Aa** genotype.

Additive Gene Action

Applying Equation 2 and the data in Table 2, the population mean is calculated as shown below.

$$a = Y_{11} - [\frac{Y_{11} + Y_{22}}{2}]$$

$$a = 16 - [\frac{16 + 0}{2}]$$

$$d = Y_{12} - [\frac{Y_{12} + Y_{22}}{2}]$$

$$d = 12 - [\frac{16 + 0}{2}] = 4$$

$$\bar{Y} = 8 + 0.64 \times 8 + 0.32 \times 4 + 0.04 \times (-8) = 14.08$$

$\bar{Y} = a(p - q) + 2pqd$ is both the mean genotypic value and the mean phenotypic value of the population with respect to the trait.

Notice that **if *d* = 0**, the heterozygote genotype has no impact on the population mean, and we say that **completely additive gene action exists**.

Two Loci

Next, consider the contributions of alleles at more than one locus and find the joint effect on the mean (Table 3).

Consider two single loci:

- Genotypic value of AA is a_A
- Genotypic value of BB is a_B

Consider multiple loci:

- Genotypic value of $AABB$ is $a_A + a_B$.
- Total genotypic value is $G_T = G_A + G_B$.
- The mid-homozygote genotypic value is the average of double homozygotes (A1A1B1B1 and A2A2B2B2).

Table 3 Joint effects of coded genotypic values and frequencies of alleles at two loci.

Two-locus genotypic values and frequencies		A locus genotype		
		AA	**Aa**	**aa**
B locus genotype	**Coded Genotypic Value/Freq.**	a_A p_A^2	d_A $2p_Aq_A$	$-a_A$ q_A^2
BB	a_B p_B^2	$G_T = \mu + a_A + a_B$ $p_A^2 p_B^2$	$G_T = \mu + d_A + a_B$ $2p_Aq_Ap_B^2$	$G_T = \mu + a_A + a_B$ $q_A^2 p_B^2$
Bb	d_B $2p_Bq_B$	$G_T = \mu + a_A + d_B$ $p_A^2 2p_Bq_B$	$G_T = \mu + d_A + d_B$ $2p_Aq_A(2p_Bq_B)$	$G_T = \mu + a_A + d_B$ $q_A^2 2p_Bq_B$
bb	$-a_B$ q_B^2	$G_T = \mu + a_A - a_B$ $p_A^2 q_B^2$	$G_T = \mu + d_A - a_B$ $2p_Aq_Aq_B^2$	$G_T = \mu - a_A - a_B$ $q_A^2 q_B^2$

Population Mean

Population mean, $\bar{Y}$ = μ + Expected values of G_A and G_B

G_A and G_B are weighted averages based on allele frequencies and coded genotypic values. The population mean is then represented by Equation 3.

$$\bar{Y} = \mu + E(G_A + G_B) = E(G_A + G_B),\ or\ \bar{Y} = \mu + \{a_A(p_A - q_A) + 2p_Aq_Ad_A\} + \{a_B(p_B - q_B) + 2p_Bq_Bd_B\}$$

Equation 3 Alternative formula for calculating population mean.

where:
G_A = weighted average of ***A*** allele,
G_B = weighted average of ***a*** allele,

$E(G_A + G_B)$ = expectation of sum of the two values,
other terms are as defined previously.

A numerical example is given in Table 4.

Table 4 A numeric example involving dominance and allele frequencies that are not equal at two loci.

Two-locus genotypic values and frequencies		A locus genotype		
		AA	Aa	aa
B locus genotype	Boded Genotypic Value/Freq.	8 0.64	4 0.32	-8 0.04
BB	4 0.04	12 0.0256	8 0.0128	-4 0.0016
Bb	2 0.32	10 0.2048	6 0.1024	-6 0.0128
bb	-4 0.64	4 0.4096	0 0.2048	-12 0.0256
Average at locus A		6.24	2.24	-9.76
Average at locus B		10.08	8.08	2.08

$$mh = \frac{12 + (-12) + 4 + -4}{4} = 0$$

$$\bar{Y} = \mu + E(G_A + G_B) = E(G_A + G_B)$$
$$\bar{Y} = \mu + \{a_A(p_A - q_A) + 2p_A q_A d_A\} + \{a_B(p_B - q_B) + 2p_B q_B d_B\}$$
$$\bar{Y} = 0 + \{8(0.8 - 0.2) + 2 * 0.8 * 0.2 * 4\} + \{4(0.2 - 0.8) + 2 * 0.8 * 0.2 * 2\}$$
$$\bar{Y} = 0 + 6.08 - 1.76 = 4.32$$

Extension To More Than 2 Loci

We can extend the concept discussed above for two-locus case to more than two loci and calculate the population mean in a similar manner; where

- $G_T = \sum G_i$
- midpoint is the average of the most extreme multi-locus-homozygotes
- Population mean = $\bar{Y}$, is represented by Equation 4

$$\bar{Y} = \mu + E(\sum G_i) = \mu + \sum E(G_i) = \mu + \sum\{a_i(p_i - q_i) + 2p_i q_i d_i\} = \sum a_i(p_i - q_i) + 2\sum p_i q_i d_i$$

Equation 4 Formula for calculating population mean involving more than two loci.

where:

$E(\sum G_i)$ = expectation of the sum of all **G** values,

other terms are as defined previously.

Average Genetic (Allelic) Effects

Individuals chosen as parents transmit only a sample consisting of ½ of its alleles. With selection, we are concerned with the transmission of value from parent to offspring. This cannot be determined based on genotypic value alone. Parents pass on their genes or alleles, NOT their genotypes, to the next generation. Genotypes are created anew in each generation. One result is that some aspects of the value of a particular genotype are unpredictable. Yet, selection theory can work only with the predictable aspects of the union of two gametes. Therefore, we introduce the *average effect of a gene (allele)* to represent this concept.

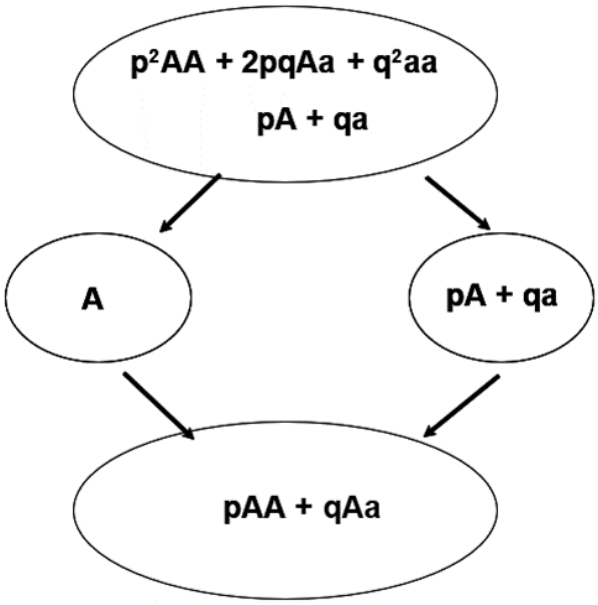


Fig. 6 Average effects of alleles on populations.

Key Concept: Although genotypes determine genotypic values, alleles and not genotypes are inherited by progeny.

Average effect of a gene (allele): The mean deviation from the population mean of individuals who received the gene (allele) from one parent is the average effect of the gene (allele). The effect of the other gene (allele) received from the remaining parent is represented as a random allele from the population … for this concept.

Formula of Average Effect of an Allele

Conceptually, let a number of gametes carrying the A allele unite at random with gametes from the population; then, the mean of the genotypes deviates from the population mean by an amount that is the average effect of the A gene. This represents the average allele effect and is the average deviation from the population mean of individuals who received a specific allele from one parent and the other allele at random from the population.

Table 5 Average effect of an allele.

Alleles	Genotypes, coded genotypic values and frequencies			Mean value of genotypes produced	Population mean to be deduced	Average effect of the allele (α_i)
n/a	AA a	Aa d	aa $-a$	next gen	this gen	(next gen) — (this gen)
A	p	q		$pa + qd$	$[a(p-q) + 2dpq]$	$q[a + d(q-p)]$
a		p	q	$-qa + pd$	$[a(p-q) + 2dpq]$	$-p[a + d(q-p)]$

The average effect of the *A* allele (or the *a* allele) from a single locus is designated as α_A (or α_a) and calculated for data presented in Table 5 as:

$$\alpha_A = q[a + d(q-p)] = 0.2[8 + (0.2 - 0.8)4] = 1.12$$
$$\alpha_a = -p[a + (q-p)d] = -0.8[8 + (0.2 - 0.8)4] = -4.48$$

Allele Substitution Effect

The average effect of an allele substitution, often designated as α, is the difference between the average effects of each allele (Equation 5).

$$\alpha = \alpha_A - \alpha_a = a + d(q-p)$$

Equation 5 Formula for calculating the average effect of allele substitution.

where:
α_A, α_a = the average effects of allele **A** and **a**, respectively,
other terms are as defined previously.

For example, from Table 5, α_1 and α_2 represent average allele effects of A and a, respectively.

Average effects of each allele can be calculated as:

$$\alpha_A = q\alpha = 0.2(5.6) = 1.12$$
$$\alpha_a = -p\alpha = (-0.8)(5.6) = -4.48$$

Thus, the average effect of allele substitution =

$$\alpha = \alpha_A - \alpha_a = 1.12 - (-4.48) = 5.6$$

Note the average genetic effect is:

- Dependent on genotypic value,
- Dependent on gene frequencies,
- A property of the population as well as the genes concerned.

Breeding Value

Breeding value is a concept that is based on the following:

- The average value of a parent is judged by its progeny.
- Alleles carried by an individual and transmitted to its offspring can be inferred from the progeny,
- Which represents the sum of the average effects of all alleles an individual carries.

Let us use the average effect of alleles to rewrite the Equation 1 as:

$$P = G + E = \alpha_i + \alpha_j + \delta_{ij} + E$$

Equation 6 Alternative formula for calculating phenotypic value based on average effect of alleles.

where:

α_i , α_j = the average effects of allele **i, j** in a diploid individual **i, j**, respectively,
δ_{ij} is the dominance deviation.

Breeding value is the value of an individual judged by the average value of its progeny. The breeding value of an individual is equal to the sum of the average effects of the alleles it carries. The summation is over pairs of alleles at a locus and over all loci (Table 5). It is defined as twice the expected deviation of the individual's progeny mean from the population mean when the individual is mated at random to other individuals from the same population (Tables 6 and 7).

Mean Breeding Value in Random Population

The mean breeding value in a random mating population is zero.

Table 6 Relationship of breeding values to genotypes.

Genotype	Breeding value
AA	$2\alpha_A = 2q\alpha$
Aa	$\alpha_A + \alpha_a = (q-p)\alpha$
aa	$2\alpha_a = -2q\alpha$

Table 7 Theoretical example of calculations of breeding values.

Genotype	Breeding value
AA	$2\alpha_A = 2(0.2)(5.6) = 2.24$
Aa	$\alpha_A + \alpha_a = (0.2-0.8)5.6 = -3.36$
aa	$2\alpha_a = -2(0.8)(5.6) = -8.96$

Deviations for Average Genetic Effects

Dominance Deviation

For a single locus: the difference between the genotypic value and the breeding value of a particular genotype is known as the dominance deviation. It is associated with the Aa genotype. δ_{ij} represents the deviation of genotypic value (i.e., G_{ij}) from the regression-fitted genotypic value and is zero when dominance is absent ($d = 0$)

Consider Genotype AA. Recall that the Coded genotypic value of AA = $+a$, and the population

mean equals $= a(p - q) + 2dpq$.
If **a** is expressed as deviation relative to the population mean, then is can be calculated using Equation 7.

$$a - [a(p - q) + 2dpq] = a(1 - p + q) - 2dpq = 2qa - 2dpq = 2q(a - dp)$$

Equation 7 Alternative formula for calculating average effect of allele substitution.

where:
a = the average effect of allele substitution,
other terms are as defined previously.

Notice that if ***d*** is not **0**, and ***p*** is not equal to ***q***, then ***a*** is affected by ***d***. Also, recall that ***a*** can be expressed in terms of the average effect of an allele substitution (Equations 8 and 9), where terms are as defined previously,

$$a = \alpha - d(q - p)$$

Equation 8 Alternative formula for calculating average effect of allele substitution.

Thus

$$2q(a - dp) = 2q(\alpha - d(q - p) - dp) = 2q(\alpha - dq + dp - dp) = 2q(\alpha - dp)$$

Equation 9 Alternative formula for calculating average effect of allele substitution.

Using similar algebra, the dominance deviation is represented by Equation 10 as,

$$\delta_{ij} = 2q(\alpha - dp) - 2pq = -2q^2 d)$$

Equation 10 Alternative formula for calculating the dominance deviation.

Observations About Dominance

Notice that

- If there is no dominance, d is zero, and the dominance deviations are also zero.
- In the absence of dominance, breeding values and genotypic values are the same.
- Alleles involved with genotypes that show no dominance, i.e., d = 0, are sometimes called ‘additive genes’, or are said to ‘act additively’.

Breeding Values and Dominance Deviations

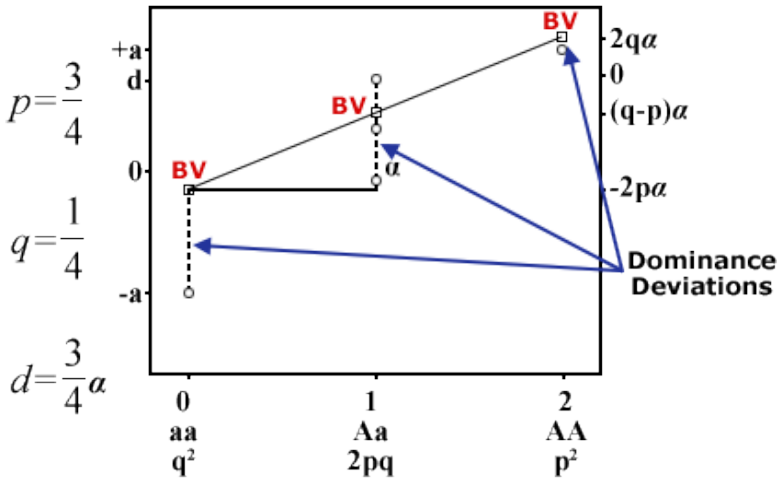


Fig. 7 Breeding values and dominance deviations.

As shown in Equation 6 the algebra of breeding values and dominance deviations provides the theoretical basis for subdividing the G component of the Phenotypic model,

$$P = G + E = \alpha_i + \alpha_j + \delta_{ij} + E$$

Thus, based on the algebra and substituting data from Table 7, dominance deviation is calculated as shwon below.

$$\delta_{AA} = G_{AA} - \mu - \alpha_A - \alpha_A = a - \mu - 2(pa + qd - \mu) = -2q^2 d$$

$$= -2(0.2)2(4) = -0.32$$

$$\delta_{Aa} = G_{Aa} - \mu - \alpha_A - \alpha_a = d - \mu - (pa + qd - \mu) - (pd - qa - \mu) = 2pqd$$

$$= 2(0.8)(0.2)(4) = 1.28$$

$$\delta_{aa} = G_{aa} - \mu - \alpha a = -a - \mu - 2(pd - qa - \mu) = -2p^2 d$$

$$= -2(0.8^2)(4) = -5.12$$

Epistasis

Epistasis exists when genotypes at two or more loci result in a genotypic value that is greater or less than the sum of the average genotypic effects at each of the individual loci. For example,

Table 8 Two-locus genotypic values that do not exhibit epistasis. The total genotypic value is the sum of the individual locus genotypic values.

Genotype at Locus A	Genotype at Locus B		
n/a	**BB**	**Bb**	**bb**
AA	22	18	6
Aa	20	16	4
aa	14	10	-2

Table 9 Two-locus genotypic values that exhibit epistasis. The total genotypic value is NOT equal to the sum of the genotypic values at the loci ($G_A + G_B$).

Genotype at Locus A	Genotype at Locus B		
n/a	**BB**	**Bb**	**bb**
AA	24	18	6
Aa	20	16	4
aa	14	10	-2

Graphical View of Epistasis

Epistasis between loci within an individual can be represented as the reaction norm of different genotypes at one locus, plotted against the genotypes at a second locus.

Figure 8A: Epistasis between loci occurs because the reaction norms of the locus B genotypes differ in slope.

Figure 8B: The reaction norms are parallel; thus, the effects of the two loci are independent, and no epistasis is present.

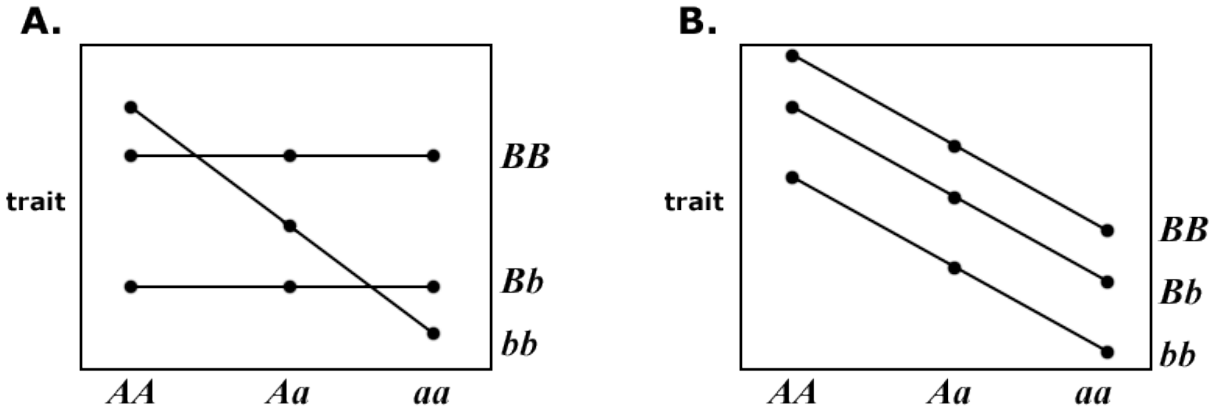


Fig. 8 Graphs of epistasis (A) and no epistasis (B).

Physiological and Statistical Dominance

Cheverud and Routman (1995) identified two concepts associated with the term epistasis: **physiological epistasis and statistical epistasis.**

The distinction between these two concepts is similar to that made between physiological and statistical dominance:

Physiological Dominance:

- Heterozygote is not midway between two homozygotes.
- Values of a and d are not dependent on allele frequencies.
- When $d \neq 0$, it reflects intralocus interaction is present.
- Least-squares solution of the unweighted regression of the number of genotypic values on the number of "a" alleles.
- Physiological dominance contributes to both additive and dominance values and variances.

Statistical Dominance Deviations:

- Deviations of single-locus from the additive combination of alleles contribute to the genotype.
- Depend on allele frequencies and will change with changes in allele frequencies.
- Least-squares solution of a weighted (weighted by genotypic frequencies) regression of genotypic value on the number of alleles.

Physiological Epistasis

In physiological epistasis (or mechanistic epistasis):

- Interaction effects occur "within" genotypes, where genes expressed within a single genome interact.
- Simply recognizes that certain genotypes at two or more loci interact in the production of a phenotype.
- If all possible genotypic classes are equally frequent in a population, the influence of genetic interactions on phenotypes will be directly observable.
- The contribution of physiological epistasis to populations is a function of the frequencies of interacting genotypes in a population.

Model for Physiological Epistasis

Let the phenotypic value of an individual be determined by the combination of the alleles present at two loci. This model is used to illustrate how physiologically based gene interactions map to components of genetic variation.

Consider the two loci, each with two alleles per locus; the two-locus (physiological) genotypic values, G_{ijkl}, are the average phenotype of individuals with the ij^{th} genotype at the first locus, and the kl^{th} genotype at the second locus. Notice that we are not given the genotypic values for the A locus nor for the B locus. So, we will determine the unweighted marginal means for each genotype.

Table 10 Example of Epistasis using Genotypic Values.

n/a	AA	Aa	aa	**Unweighted marginal mean**
BB	$G_{AABB} = 22$	$G_{AaBB} = 18$	$G_{aaBB} = 6$	$G_{..BB} = 15.33$
Bb	$G_{AABb} = 20$	$G_{AaBb} = 16$	$G_{aaBb} = 4$	$G_{..Bb} = 13.33$
bb	$G_{AAbb} = 14$	$G_{Aabb} = 10$	$G_{aabb} = -2$	$G_{..bb} = 7.33$
Unweighted marginal means	$G_{AA..} = 18.66$	$G_{Aa..} = 14.66$	$G_{aa..} = 2.66$	$G_{....} = 12$

Single Locus Genotype

The single-locus genotype is defined as the unweighted average across the genotypes at the second locus, for example, Equation 11.

$$G_{ij..} = \frac{(G_{ij11} + G_{ij12} + G_{ij22})}{3}, \; G_{..kl} = \frac{(G_{11kl} + G_{12kl} + G_{22kl})}{3}$$

Equation 11 Alternative formula for calculating the dominance deviation.

where:
$G_{ij..}, G_{..kl}$ = the single locus genotype value.

Thus, the value at locus A is,

$$G_{AA..} = \frac{(22 + 20 + 14)}{3} = 18.66$$

and at locus B is,

$$G_{..BB} = \frac{(22 + 18 + 6)}{3} = 15.33$$

Non-Epistatic Genotypic Value

Subscripts A and a or B and b refer to the two alleles at the interacting loci. The single locus values of a and d are computed as in Equation 12.

$$a_i = G_{11..} - \frac{(G_{11..} + G_{22..})}{2}, \; d_i = G_{12..} - \frac{(G_{11..} + G_{22..})}{2}$$

Equation 12 Formula for calculating the single locus values,

where:
a_i, d_i = the single locus values.

Thus, the value at locus A is,

$$a_A = G_{11..} - \frac{(G_{11..} + G_{22..})}{2} = 18.66 - \frac{18.66 + 2.66}{2} = 18.66 - 10.66 = 8$$

and

$$d_A = G_{12..} - \frac{(G_{11..} + G_{22..})}{2} = 14.66 - \frac{18.66 + 2.66}{2} = 14.66 - 10.66 = 4$$

Similarly, a_A and d_A can be calculated to be 4 and 2, respectively. Try it!

The non-epistatic genotypic value, ne_{ijkl}, is calculated using Equation 13, represented by,

$$ne_{ijkl} = G_{ij..} + G_{..kl} - G_{....}$$

Equation 13 Formula for calculating the non-epistatic genotypic value,

where:
all terms are as defined previously.

For **AABB** genotype,

$$ne_{AABB} = G_{AA..} + G_{..BB} - G_{....} = 18.66 + 15.33 - 12 = 21.99.$$

Epistatic Genotypic Value

Table 11 Non-epistatic values.

n/a	AA	Aa	aa
BB	21.99	17.99	5.99
Bb	19.99	15.99	3.99
bb	13.99	9.99	-2.01

The epistatic genotypic value is represented by Equation 14.

$$e_{ijkl} = G_{ijkl} - ne_{ijkl}$$

Equation 14 Formula for calculating the epistatic genotypic value.

Example from data in Table 11 is,

$$e_{ijkl} = 22 - 21.99 = 0.01.$$

A value of e_{ijkl} different from zero indicates that **Physiological Epistasis** is present. In this example, there is little evidence for epistasis for this cell. Is there evidence for epistasis in the other cells?

Statistical Epistasis

In statistical epistasis (or population epistasis):

- The term is used to refer to the amount of population variation in genotypic values associated with variation among loci.
- Notation that is often used includes V_i, or G_{AxB}, or I_{AxB}.
- The amount of statistical epistasis present in a population is a function of the frequencies of interacting multilocus genotypes and therefore is a function of population allele frequencies as it is for additive and dominance variance (V_A and V_D). It is calculated using Equation 15.

$$G_T = G_A + G_B + G_{A \times B}$$

Equation 15 Formula for calculating the statistical genotypic value.

where:
G_T = the total epistatic (statistical) value,
G_A = value at locus A,
G_B = value at locus B,
$G_B + G_{A \times B}$ = value due to A by B interaction.

Presence of epistasis between locus A and B changes the population mean, (Y), mid-homozygote value (μ), a (additive), and d (dominance) values.

Table 12 An example with Epistatic effects.

Two-locus genotypic values and frequencies		A locus genotype		
		AA	**Aa**	**aa**
B locus genotype	Coded Genotypic Value/ Freq	8 0.64	4 0.32	−8 0.04
BB	4 0.04	24 0.0256	18 0.0128	6 0.0016
Bb	2 0.32	20 0.2048	16 0.1024	4 0.0128
bb	−4 0.64	14 0.4096	10 0.2048	−2 0.0256
Average at locus A		16.32	12.24	0.24
Average at locus B		21.36	18.08	12.08

Epistasis Effects

The genotypic value of *AABB* has been increased from 22 to 24 due to epistatic effects.

There are changes in population mean, mid-homozygote values for *A* and *B* locus, and the average at *A* and *B* locus as shown in calculations below.

$$\text{Epistatic effects between } AABB = 2$$
$$G_T = \mu + G_A + G_B = 10 + 8 + 4 = 22$$
$$\text{With epistatic effects } G_T = \mu + G_A + G_B + G_{A\times B} = 10 + 8 + 4 + 2 = 24$$

$$\text{Changes in the mid-homozygote } (\mu) = \frac{24 + (-2)}{2} = 11$$

$$\text{Changes in the mean population } (Y) = 14.37$$
$$\mu_A = 8.28$$
$$\mu_B = 16.72$$
$$G_A : +a = 8.04; d = 3.96; -a = -8.04$$
$$G_B : +a = 4.64; d = 1.36; -a = -4.64$$

References

Cheverud, J. M. and E. J. Routman. 1995. Epistasis and its contribution to genetic variance components. Genetics, 139, 1455±1461.

How to cite this chapter: Beavis, W., K. Lamkey, K. Espinosa, and A. A. Mahama. 2023. Gene Effects. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding.* Iowa State University Digital Press.

Chapter 6: Components of Variance

William Beavis; Kendall Lamkey; Katherine Espinosa; and Anthony Assibi Mahama

This chapter explores sources of phenotypic variation (Fig. 1), whether genetic or environmental and how these contribute to the heritability of selected traits. It covers the derivation of variance components and covariance, the relationship among variance components, and the role of epistasis.

Learning Objectives

Type your learning objectives here.

- Learn to model components of genetic variances for purposes of estimating heritability.
- Be able to explain:
    - The impact of allele frequencies on genetic components of genotypic variability,
    - The reason estimates of components of genetic variability are limited to the population from which they are estimated,
    - The reason additive variance does not imply additive gene action and
    - How additive genetic variance can arise from genes with any degree of dominance or epistasis.

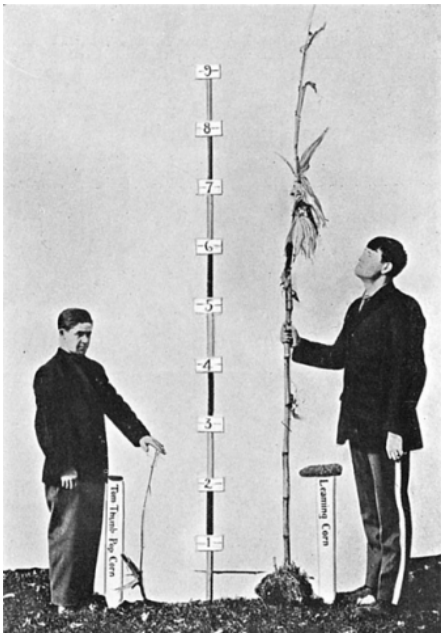


Fig. 1. Variation in height depicted in human and corn plants. From "Critique of the Theory of Evolution" (1915) by Thomas Hunt Morgan, available freely at Project Gutenberg. Licensed under Public Domain via Wikimedia Commons.

Phenotypic Components of Variance

Recall that our working model for the phenotype includes genotypic and non-genotypic (environmental) sources of variability (Equation 1):

$$P = \mu + G + E$$

Equation 1 Working model of phenotypic, genotypic, and environmental effects,

where:
μ = overall mean,
G = genotypic effect,
V_E = environmental effect.

The source of phenotypic variability determines whether selection for the trait will result in a heritable response, i.e., will be passed on to the next generation.

For purposes of making decisions in plant breeding, if two populations have different phenotypic means, we want to know whether the differences are due to different environments, different genotypes, or some combination of both. If the differences are due to genotypic differences, then what proportion of the genotypic differences is heritable?

Fig. 2 Comparing phenotypic traits of plant populations at the University of KwaZulu-Natal in South Africa. Photo by Iowa State University.

Algebraic Description

Using simple algebra and our working model, we can show that the phenotypic variance $\mathbf{V_P}$ within a population is equal to the sum of the genotypic variance $\mathbf{V_G}$ and environmental variance V_E, assuming that $\mathbf{V_G}$ and $\mathbf{V_E}$ are independent (Equation 2);

$$V(P) = V(G) + V(E)$$

Equation 2 Working model of phenotypic, genotypic, and environmental variances,

If the genotypic values and environmental deviations are not independent, the **V**(**P**)=**V**(**G**+**E**), and **V**(**P**) can be increased by twice the covariance of **G** with **E** if they are not independent (Equation 3):

$$V(P) = V(G) + V(E) + 2Cov(G, E)$$

Equation 3 Composition of phenotypic variance,

where:

$Cov(G, E)$ = joint variation between G and E.

Genetic Components of Variance

Genetic components of variability can be divided into several subcategories, including additive variance, $\mathbf{V_A}$, dominance variance, $\mathbf{V_D}$, and epistatic variance, $\mathbf{V_I}$. Together, the values for each of these subcategories yield the total amount of genetic variation, $\mathbf{V_G}$, responsible for a particular phenotypic trait: $\mathbf{V_P} = V_\mathbf{G} + \mathbf{V_E}$.

Consider the ratio of $\mathbf{V_G}$ to $\mathbf{V_P}$. This was originally recognized by statistical geneticists (such as RA Fisher) as the **genotypic intra-class correlation**. To understand this, consider the evaluation of a line **i**, for a phenotype, **Y**. Next, imagine that you can evaluate line **i** repeatedly. Let us designate these repeated measurements as **j**. We can then designate these repeated measurements of the phenotype as $\mathbf{Y_{ij}}$. There is a Covariance among these repeated evaluations that we can represent as $Cov(Yij, Yij') = Var(Gi), for\ j \neq j')$.

Explanation of Formula

Thus the correlation among these repeated measures is (Equation 4)

$$\rho(Y_{ij}, Y_{ij'}) = \frac{Cov(Y_{ij}, Y_{i,j'})}{\sqrt{Var(Y_{ij})Var(Y_{ij'})}},$$

Equation 4 Correlation among repeated measures,

where:
ρ = correlation between $\mathbf{Y_{ij}}$ and $\mathbf{Y_{ij}}$',
$Cov(Y_{ij}, Y_{i,j'})$ = covariance between $\mathbf{Y_{ij}}$ and $\mathbf{Y_{ij}}$',
$Var(Y_{ij})$ & $Var(Y_{ij'})$ = variance of$\mathbf{Y_{ij}}$ and $\mathbf{Y_{ij}}$'.

Because $Var(Y_{i,j}) = V(G_i) + V(E_{ij})$, the correlation is represented by Equation 5.

$$\rho(Y_{ij}, Y_{ij'}) = \frac{V(G_i)}{V(G_i) + V(E_{ij})} = \frac{V(G_i)}{V(P_{ij})}$$

Equation 5 Correlation among repeated measures,

where:
ρ = correlation between $\mathbf{Y_{ij}}$ and $\mathbf{Y_{ij}}$',
$V(G_i$ = genotypic variance of genotype **i**,

$Var(E_{ij})$ = variance of environment **j** for genotype **i**,
$Var(P_{ij})$ = phenotypic variance of genotype **i** in environment **j**.

Heritability in the Broad Sense

Broad sense heritability is estimated by Equation 6.

$$H = \frac{V_G}{V_P}$$

Equation 6 Formula for estimating broad sense heritability,

where:
V_G = the total genetic variance,
V_P = the phenotypic variance.

JL Lush, an animal breeder, also referred to this intra-class correlation coefficient as **heritability in the broad sense** (1937). He wanted to distinguish the application of intra-class correlation to animals from the concept of repeatability. Repeatability as an engineering concept refers to the same measurement procedure conducted by a single observer, using a single measuring instrument, under the same conditions, at a single location, over a short period of time. As a result, plant and animal breeders tend to prefer the use of broad sense heritability for the genotypic intra-class correlation, although both plant and animal breeders routinely evaluate a single trait on individual genotypes repeatedly over time and space (locations and years).

Broad-Sense Components

The genetic variance can be recognized as consisting of several components (Equation 7):

$$V_G = V_A + V_D + V_I$$

Equation 7 Composition of total genotypic variance,

where:
V_G = total genotypic variance,
V_A = additive genetic variance, that is, the variance of breeding values, and refers to the deviation from the mean phenotype due to inheritance of a particular allele and this allele's relative effect on phenotype, i.e., relative to the mean phenotype of the population,
V_D =dominance variance due to interactions between alternative alleles at a specific locus,
V_I = epistatic variance due to interaction between alleles at different loci.

Heritability in the Narrow Sense

Heritability in the narrow sense was defined by JL Lush (1937) to represent the extent to which phenotypes are determined by the genes transmitted from their parents (Equation 8):

$$h^2 = \frac{V_A}{V_P}$$

Equation 8 Composition of total genotypic variance,

where:
h^2 = heritability in the narrow sense,
V_A = additive genetic variance, that is, the variance of breeding values, and refers to the deviation from the mean phenotype due to inheritance of a particular allele and this allele's relative effect on phenotype, i.e., relative to the mean phenotype of the population,
V_P = phenotypic variance.

So, we can now expand our model for the phenotypic variance to include several genetic variance components and environmental variance as in Equation 9.

$$V_P = V_A + V_D + V_I + V_E$$

Equation 9 Composition of total genotypic variance,

where:
V_E = environmental variance. Other variables are as described previously.

Table 1 Variance components and sources of variation.

Variance component	Symbol	Source of variation
Phenotypic	V_P	Phenotypic value
Genotypic	V_G	Genotypic Value
Additive	V_A	Breeding Value
Dominance	V_D	Dominance deviation
Interaction	V_I	Interaction deviation
Environmental	V_E	Non-genetic deviation

Deriving Variance Components

The genetic components of variance are influenced by the gene frequency and the assigned genotypic values$a$ and d. The information needed to derive V_A and V_D are:

Table 2 Derivation of additive and dominance variance components of genetic variance.

Genotypes	AA	Aa	aa
Frequencies	p^2	$2pq$	q^2
Coded GV	a	d	$-a$
Genotypic Value	$2q(a-pd)$ $2q(a-qd)$	$a(q-p)+d(1-2pq)$ $(q-p)a+2pqd$	$-2p(a+qd)$ $-2p(a+pd)$
Breeding Value	$2qa$	$(q-p)a$	$-2pa$
Dominance Deviation	$-2q^2d$	$2pqd$	$-2p^2d$

The variances are thus obtained by squaring the values in the table, multiplying by the frequency of the genotype concerned, and summing over the three genotypes (Equation 10).

$$
\begin{aligned}
V_A &= 4p^2q^2\alpha^2 + 2pq(q-p)^2\alpha^2 + 4p^2q^2\alpha^2 \\
&= 2pq(2pq + q^2 - 2pq + p^2 + 2pq)\alpha^2 \\
&= 2pq(p^2 + 2pq + q^2)\alpha^2 \\
&= 2pq\alpha^2 \\
&= 2pq[a + d(q-p)]^2 \\
V_D &= d^2(4q^4p^2 + 8p^3q^3 + 4p^4q^2) \\
&= 4p^2q^2d^2(q^2 + 2pq + p^2) \\
&= (2pqd)^2
\end{aligned}
$$

Equation 10 Derivation of additive and dominance variances,

where:
p = frequency of allele A,
q = frequency of allele a,
α = average effect of an allele,
a, d = coded genotypic values.

Covariance

If there is no dominance at the locus under consideration **d** = **0**, then: $V_A = 2pq\alpha^2$.

If there is complete dominance **d** = **a**, the additive variance becomes $V_A = 8pq^3\alpha^2$.

The total genetic variance is estimated with Equation 11.

$$V_G = V_A + V_D + 2COV_{AD}$$

Equation 11 Total genetic variance formula including covariance of additive and dominance variances,

where:

COV_{AD} is the covariance of breeding values with dominance deviations, which can be demonstrated to be zero. Thus substituting in Equation 12,

$$V_G = V_A + V_D = 2pq[a + d(q - p)]^2 + [2pqd]^2$$

Equation 12 Total genetic variance formula relating additive and dominance variances to allele frequencies and coded genotypic values,

where:

V_A, V_D, p, q, and, d are as described previously.

Component Relationships

The relationships among variance components, gene action, and allele frequencies for the two allele case can be graphically represented (Figs. 3, 4, 5).

Additive Gene Action

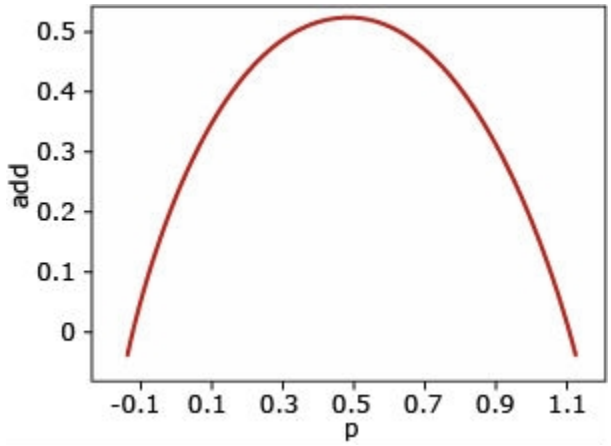


Fig. 3 Genetic variance due to additive gene action only and related to allele frequency changes.

Additive gene action: There is no dominance (a>0, d=0). In this case, the genetic variance is additive, and it is greatest when p=q=0.5.

Complete Dominance

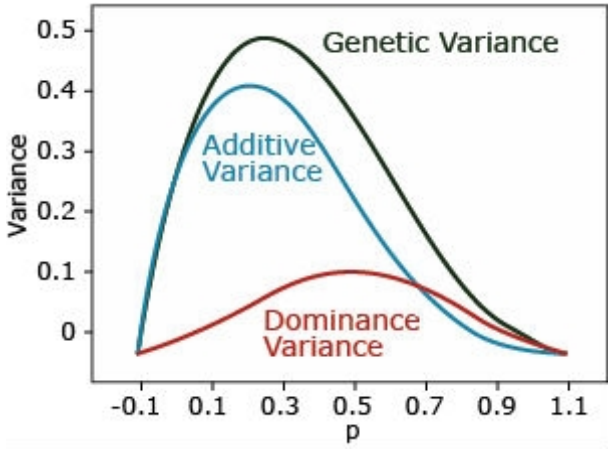


Fig. 4 Variance changes in response to allele frequency changes.

Complete dominance: ($a > 0, d = a$). The dominance variance is maximal when p=q=0.5. The additive is maximal when p=0.3.

Overdominance

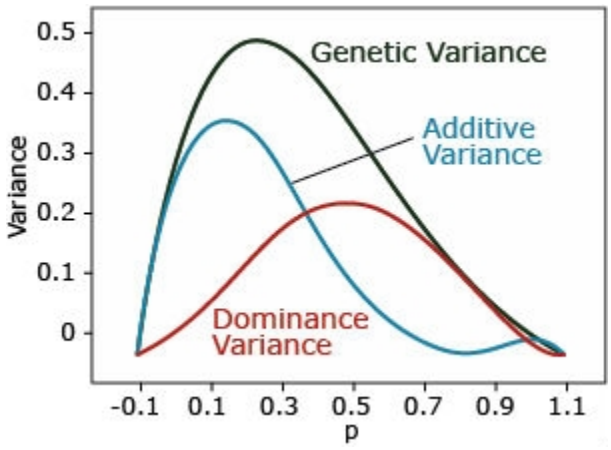


Fig. 5 Variance changes in response to allele frequency changes.

Overdominance: $(a = 0, d > 0)$. The dominance variance is the same as incomplete dominance.

Principles to Remember

Important principles to remember:

1. All the components of genetic variance are dependent on the gene frequencies.
2. Estimates of components of genetic variances are valid only for the population from which they are estimated.
3. The concept of additive variance does not carry with it the assumption of additive gene action; the existence of additive variance is not an indication that genes act additively.
4. Additive variance can arise from genes with any degree of dominance or epistasis.

Fig. 6 Examining phenotypic traits of maize fields at the University of KwaZulu-Natal, South Africa. Photo by Iowa State University.

Influence of Epistasis

Two Or More Loci: Influence of Epistasis on Components of Genetic Variance

When more than one locus is under consideration, then deviations due to interactions among loci give rise to additional variance components due to epistatic interactions, $\mathbf{V_i}$ (Equation 13).

$$V_I = V_{AA} + V_{AD} + V_{DD} + etc.$$

Equation 13 Components of epistatic interaction variance,

where:

V_{AA} = additive × additive variance is the interaction between two breeding value,

V_{AD} = additive × dominance variance is the interaction between the breeding value of one locus and the dominance deviation of the other,

V_{DD} = dominance × dominance variance is the interaction between the two dominance deviations.

Epistatic Model

A non-intuitive consequence of the epistatic models is that additive variance can arise from purely epistatic genetics. For example, let's consider the special case of an F_2 population with equal frequencies of two alleles at each of two independently segregating loci. Let's imagine that we know the genotypes at each of these functional loci and analyze the F_2 population using a regression approach for each of the loci and their interactions.

Table 3 Sources of genetic variability and associated df in an analysis of independently segregating loci in an F_2 population based on a regression approach of analysis.

Source of variance	**Df**
Locus A	**2**
Linear (Additive)	1
Quadratic (Dominance)	1
Locus B	**2**
Linear (Additive)	1
Quadratic (Dominance)	1
Epistasis	**4**
Linear A x Linear B (A * A)	1
Linear A x Quadratic B (A * D)	1
Quadratic A x Linear B (D * A)	1
Quadratic A x quadratic B (B * D)	1
Total	**8**

Example 1

Analysis of a phenotype in an F_2 population with equal frequencies of alleles at two functionally polymorphic loci, each contributing only additive coded genotypic values from the A locus and a B locus, i.e., $P = \mu + a_A + a_B$, where μ= 5, a_A = 3, and a_B =1.

Table 4 Parameters and their Coded Genotypic values for P (example 1).

Parameter	Value
a_A	3
d_A	0
a_B	1
d_B	0
μ	5

Table 5 Coded Genotypic values for two functional bi-allelic loci in an F_2 population derived from a cross of two inbred lines (example 1).

n/a	n/a	A_1A_1	A_1A_2	A_2A_2	**Mean**
n/a	n/a	1/4	1/2	1/4	n/a
B_1B_1	1/4	9	6	3	6
B_1B_2	1/2	8	5	2	5
B_2B_2	1/4	7	4	1	4
Mean	n/a	8	5	2	2.75

Table 6 Calculated variances components for the F_2 population described in example 1.

Variance component	Variance
$\delta^2_{A_A}$	4.5
$\delta^2_{A_B}$	0.5
$\delta^2_{D_A}$	0
$\delta^2_{D_B}$	0
δ^2_{AA}	0
δ^2_{AD}	0
δ^2_{DA}	0
δ^2_{DD}	0

Example 2

Analysis of a phenotype in an F_2 population with equal frequencies of alleles at two functionally polymorphic loci where only single epistatic interaction between the genotypes will produce an altered phenotype, $P = \mu + a_A + a_B + d_A + d_B + e_{AABB}$, where μ = 0, a_A = 0, a_B = 0, d_A = 0, d_B = 0, d_{AABB} = 50.

Table 7 Coded Genotypic values for two functional bi-allelic loci in an F_2 population derived from a cross of two inbred lines (example 2).

n/a	n/a	A_1A_1	A_1A_2	A_2A_2	**Mean**
n/a	n/a	1/4	1/2	1/4	n/a
B_1B_1	1/4	0	0	0	0
B_1B_2	1/2	0	0	0	0
B_2B_2	1/4	0	0	50	12.5
Mean	n/a	0	0	12.5	3.125

Table 8 Calculated variances components for the F_2 population described in example 2.

Population variances	Population	Percent
Total genetic	**146.484**	**100.0%**
Additive effects	39.063	26.7%
Dominance	19.531	13.3%
Epistasis	87.891	60.0%

References

Lush J. L. 1937. Animal Breeding Plans, Collegiate Press Inc, Ames, IA.

How to cite this chapter: Beavis, W., K. Lamkey, K. Espinosa, and A. A. Mahama. 2023. Components of Variance. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 7: Estimates of Variance

William Beavis; Kendall Lamkey; Katherine Espinosa; and Anthony Assibi Mahama

Estimating heritability is a fundamental concept of quantitative genetics. One method for obtaining estimates of heritability is the use of variance and covariance of a known collection of relatives from various types of progeny.

Learning Objectives

- Model components of genetic variances and covariances for purposes of estimating heritability, a fundamental concept of quantitative genetics.
- Explain why estimates of components of genetic variability are limited to the population from which they are estimated.
- Students will derive variance components and recognize the differences amongst components obtained from different progeny used for estimating heritability.
- Students will write out the correct linear models for the correct mean squares and expected mean squares in a ANOVA table, and correctly interpret the ANOVA and algebraically extract the correct values for estimating heritability.
- Leverage of the powerful algebraic equivalence of covariances within groups of relatives to variances among the same groups.

Covariance of Relatives

Recall that $Cov(Y_{ij},Y_{ij'}) = Var(G_i)$, for $j \neq j'$. In the context of genotypic sampling of relatives, this general relationship has a profound and powerful impact on interpretation of ANOVA. It means that the covariance among a sample of relatives can be used to estimate components of genetic variance associated with the genotypic effect.

Table 1 A general ANOVA table for any type of related progeny.

n/a			EMS	
Source	**df**	**MS**	**Variances**	**Covariances**
Reps	$r-1$	n/a	n/a	n/a
Progeny	$p-1$	MS_3	$\sigma_w^2 + k\sigma^2 + rk\sigma_p^2$	$\sigma_w^2 + k\sigma^2 + rk[\text{Cov(progeny)}]$
Error	$(r-1)(p-1)$	MS_2	$\sigma_w^2 + k\sigma^2$	$\sigma_w^2 + k\sigma^2$
Total	$rp-1$	n/a	n/a	n/a
Within Progeny	$rp(k-1)$	MS_1	σ_w^2	$\sigma_{we}^2 + [\sigma_T^2 - \text{Cov(progeny)}]$

Note that there are *p* progeny grown in *r* reps. Cov(progeny) refers to the covariance of the progeny, where the progeny can be full-sibs, half-sibs, S_1-progeny, S_2-progeny, testcross progeny, etc. The key is to know the progeny type and take advantage of the general rule that the variance among progeny is equal to the covariance of the progenies.

Note the use of σ^2_T instead of σ^2_G in the within progeny line of the ANOVA table. This is because σ^2_G is usually equal to $\sigma^2_A + \sigma^2_D$ the total variance in a non-inbred random mating population. If the population does not have a random mating structure, then the total variance will be something other than $\sigma^2_A + \sigma^2_D$. For example, the total genetic variance for an F_3 population is as in Equation 1.

$$\sigma_F^2 = \frac{3}{2}\sigma_A^2 + \frac{3}{2}\sigma_D^2$$

Equation 1 Formula for total genetic variance for and F_3 population.

where:

σ_F^2 = total genetic variance for F_3 population,

σ_A^2 = additive variance,

σ_D^2 = dominance variance.

Linear Models for Phenotypic Values

The covariance of relatives is simply that relatives tend to show more phenotypic similarities than with each other than with unrelated individuals. For example let X_{ij} represent an individual from the mating of parent *i* and parent *j*:

Table 2 Descriptions of relationships between individuals X_{ij} and $X_{i'j'}$.

Conditions	Description
$i = i', j = j'$	Full-sibs
$i = j', j = i'$	Reciprocal Full-sibs
$i = i', j \neq j'$	Maternal half-sibs
$i \neq i', j = j'$	Paternal half-sibs

Specifying covariance of relatives in terms of genetic variances has the following assumptions:

1. Regular diploid and solely Mendelian inheritance
2. No environmental correlations among relatives
3. No gametic disequilibrium
4. The relatives are not inbred
5. The relatives are considered to be random members of some non-inbred population

With these assumptions, we can specify the covariance of relatives as in Equation 2.

$$\text{Cov} = \alpha\sigma^2_A + \delta\sigma^2_D + \alpha^2\sigma^2_{AA} + \alpha\delta\sigma^2_{AD} + \delta^2\sigma^2_D + \alpha^3\sigma^2_{AAA} + \ldots$$

Equation 2 Formula for covariance of relatives,

where:
α = the coefficient of relative relationship,
σ^2_A = additive genetic variance,
δ = the dominance relationship coefficient,
σ^2_D = the dominance variance,
$\sigma^2_{AA}, \sigma^2_{AD}, \sigma^2_{AAA}$ = the epistatic variances.

Common Types of Relatives

Using the result of Equation 1 for some common types of relatives, it can be shown that:

Covariance of half-sibs with one common parent is represented by Equation 3.

$$Cov(HS) = \frac{(1 + F_A)}{4}\sigma^2_A + (\frac{(1 + F_A)}{4})^2\sigma^2_{AA} + \ldots$$

Equation 3 Formula for calculating covariance of half-sibs,

where:
F_A = inbreeding coefficient of parent A.

Covariance of full-sibs with parents A and B is estimated using Equation 4.

$$Cov(FS) = \frac{(2+F_A+F_B)}{4}\sigma^2_A + \frac{(1+F_A)(1+F_B)}{4}\sigma^2_D + (\frac{(2+F_A+F_B)}{4})^2\sigma^2_{AA} + (\frac{(2+F_A+F_B)}{4})(\frac{(1+F_A)(1+F_B)}{4})\sigma^2_{AD} + (\frac{(1+F_A)(1-F_B)}{4})^2\sigma^2_{DD} + (\frac{(2+F_A+F_B)}{4})^3\sigma^2_{AAA} + \dots$$

Equation 4 Formula for calculating covariance of full-sibs,

where:
F_A = inbreeding coefficient of parent A,
F_B = inbreeding coefficient of parent B,
$\sigma^2_{AA}, \sigma^2_{AD}, \sigma^2_{AAA}$ = the epistatic variances.

F_2 and F_3 Progenies

Table 3 F_3 progeny genotypes, frequencies, genotypic values and progeny mean values representation.

Genotype	Freq	GV	F3 Progeny			F3 Progeny Mean
			AA	Aa	aa	
AA	1/4	a	1	0	0	a
Aa	1/2	d	1/4	1/2	1/4	1/2d
aa	1/4	-a	0	0	1	-a

F_2 and F_3 Variances

Total genetic variance among F_2 individuals determined suing Eqaution 5:

$$\sigma^2_{F_2} = \sigma^2_A + \sigma^2_D = \frac{1}{2}a^2 + \frac{1}{4}d^2$$

Equation 5 Formula for calculating total genetic variance among F_2 individuals,

where:
a, d = genotypic values of **AA or aa** and **Aa** genotypes, respectively.

Total F_2 phenotypic variation:

$$\sigma^2_{F_2} = \sigma^2_A + \sigma^2_D + \sigma^2_{E1} = \frac{1}{2}a^2 + \frac{1}{4}d^2 + E1$$

Equation 6 Formula for calculating total phenotypic variance among F_2 individuals,

where:
$\sigma^2_{F_2}$ = total phenotypic variance among F2 individuals,
$E1$ = the non-genetic variation among F2 plants,
other terms are as described previously.

Recall that the F_2 is our reference population for interpretation of genetic results. To estimate the total genetic variation of an F_2, we need the parents and the F_1 (to estimate environmental effects)as well as the F_2 generation.

F_3 Variances

F_3 population mean is equal to $\frac{1}{4}d$

Variance among F_3 progeny means is determined using Equation 7.

$$\sigma^2_{\bar{F_3}} = [\frac{1}{4}a^2 + \frac{1}{2}(\frac{1}{2}d)^2 + \frac{1}{4}(-a^2)] - (\frac{1}{4}d)^2 = \frac{1}{2}a^2 + \frac{1}{16}d^2 = \sigma^2_A + \frac{1}{4}\sigma^2_D,$$

Equation 7 Formula for calculating variance among F_3 progeny means,

where:
$\sigma^2_{\bar{F_3}}$ = variance among F_3 progeny means,
other terms are as described previously.

Variance within F_3 progeny means is determined using Equation 8.

$$\bar{\sigma}^2_{F_3} = [[\frac{1}{4}a^2 + \frac{1}{2}d^2 + \frac{1}{4}(-a^2)] - (\frac{1}{2}d)^2] = \frac{1}{4}a^2 + \frac{1}{8}d^2 = \frac{1}{2}\sigma^2_A + \frac{1}{2}\sigma^2_D,$$

Equation 8 Formula for calculating total phenotypic variance within F_3 progeny means,

where:
$\bar{\sigma}^2_{F_3}$ = variance within F3 progeny means,
other components are as described previously.

Total variance among F_3 individuals is then estimated from Equation 9:

$$\sigma^2_{F_3} = \frac{3}{2}\sigma^2_A + \frac{3}{4}\sigma^2_D$$

Equation 9 Formula for calculating total variance among F_3 individuals,

where:
$\sigma^2_{F_3}$ = total variance among F3 individuals,
other terms are as described previously.

F_3 progenies can be grown in replicated trials, so a set of equations like the following could be written to estimate the variance in different generations (Equation 10).

$$\sigma^2_{F_2} = \sigma^2_A + \sigma^2_D + \sigma^2_{E1},$$

$$\sigma^2_{\bar{F}_3} = \sigma^2_A + \frac{1}{4}\sigma^2_D + \sigma^2_{E2},$$

$$\bar{\sigma}^2_{F_3} = \frac{1}{2}\sigma^2_A + \frac{1}{2}\sigma^2_D + \sigma^2_{E1},$$

$$and, \sigma^2_{E2} = \frac{\sigma^2}{r},$$

Equation 10 Formulae for calculating variances for F_2 and F_3,

where:
$E2$ = non-genetic variation among mean variance of F_3 progeny,
r = number of replications,
other terms are as described previously.

ANOVA for F_3 Progenies

ANOVA for F_3 progenies can be calculated from a replicated experiment.

Table 4 ANOVA for F_3 Progenies.

Source	df	MS	EMS
Reps	$r-1$	n/a	n/a
Progeny	$p-1$	M3	$\sigma_e^2 + r\sigma_{F_2}^2$
Error	$(r-1)(p-1)$	M2	σ^2
Total	$rp-1$	n/a	n/a
Within Progeny	$rp(k-1)$	M1	$\sigma_{F_a}^2 = \frac{1}{2}\sigma_A^2 + \frac{1}{2}\sigma_D^2 + \sigma_{E1}^2$

Then using Equation 11,

$$\sigma_{\bar{F}_3}^2 = \frac{M3 - M2}{r} = \sigma_A^2 + \frac{1}{4}\sigma_D^2$$

$$\sigma_{E2}^2 = \frac{M2}{r} + \frac{\sigma^2}{r}$$

Equation 11 Formulae for calculating variances using MS and EMS,

σ_{E1}^2 is estimated as environmental variance within P1 or P2 plots (the inbred lines).

Note that the phenotypic variance among F_3 families is determined with Equation 12:

$$\hat{\sigma}_p^2 = \frac{M3}{rk} = \frac{\sigma_w^2}{rk} + \frac{\sigma^2}{r} + \sigma_c^2,$$

Equation 12 Formula for calculating total phenotypic variance among F_3 families,

where:

$\hat{\sigma}_p^2$ = estimate of total phenotypic variance,

σ_w^2 = within progeny variance,

σ_c^2 = phenotypic variance among F_3 families = the genotypic variance,

r = number of replications,

k = individuals withing progeny type.

Estimate of Heritability

A type of heritability estimates on a progeny mean basis can be calculated as shown in Equation 13:

$$h^2 = \frac{\sigma_c^2}{\sigma_p^2} = \frac{\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2}{\frac{\sigma_w^2}{rk} + \frac{\sigma^2}{r} + \sigma_c^2},$$

Equation 13 Formulae for calculating heritability on progeny entry mean basis,

where:
Terms are as described previously.

Note that this estimate of heritability contains both additive and dominance variance. Recall that this is an estimate of intra-class correlation, thus it is a type of broad-sense heritability.

Limitations of this method (often referred to as Mather's methods)

1. Estimates apply only to specific parents.
2. Estimates for σ^2_{E1} may vary among generations.
3. Estimates for a particular set of F_2 plants can be obtained in only one environment.
4. Linkage will bias estimates.
5. Epistasis is assumed to be absent.

Bi-Parental Progenies

Bi-parental progenies are just crosses between individual plants; thus, genetically, they are full-sibs. For example, in a random mating maize population, you could cross two individual plants reciprocally and bulk the seed from the two ears. This would produce enough seed to plant FS progeny in 10-20 replications. We could then think about *n* plants and making *n* / 2 full-sib families. The covariance then be computed using Equation 14.

Table 5 ANOVA Table for bi-parental progenies

Source	df	MS	EMS
Reps	$r-1$	n/a	n/a
Among families	$\frac{n}{2}-1$	$\mathrm{M3}=\sigma_w^2+k\sigma^2+rk\sigma_c^2$	$\sigma_w^2+k\sigma^2+rk[Cov(FS)]$
Error	$(r-1)(\frac{n}{2}-1)$	$\mathrm{M2}=\sigma_w^2+k\sigma^2$	$\sigma_w^2+k\sigma^2$
Total	$r\frac{n}{2}-1$	n/a	n/a
Within families	$r\frac{n}{2}(k-1)$	$\mathrm{M1}=\ \sigma_w^2$	$\sigma_w^2+[\sigma_G^2-Cov(FS)]$

$$\sigma_c^2=Cov(FS)=\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_D^2;\sigma_c^2=\frac{M3-M2}{rk};\sigma_w^2=\sigma_w^2+\frac{1}{2}\sigma_A^2+\frac{3}{4}\sigma_D^2;\hat{\sigma}_w^2=M1$$

Equation 14 Extracting different variance components.

Summary

Table 6 Data from Cockerham, 1983.

Progeny Type	Cov(progeny)	Total Variance, σ_T^2
Half-sib	$\frac{1}{4}\sigma_A^2$	$\sigma_A^2+\sigma_D^2$
Full-sib	$\frac{1}{2}\sigma_A^2+\frac{1}{4}\sigma_D^2$	$\sigma_A^2+\sigma_D^2$
S1(F2:3)	$\sigma_A^2+\frac{1}{4}\sigma_D^2+D_1+\frac{1}{8}D_2$	$\frac{3}{2}\sigma_A^2+\frac{1}{2}\sigma_D^2+2D_1+\frac{1}{2}D_2+\frac{1}{4}H^*$
S2(F3:4)	$\frac{3}{2}\sigma_A^2+\frac{1}{8}\sigma_D^2+2.5D_1+\frac{9}{16}D^2+\frac{1}{16}H^*$	$\frac{7}{4}\sigma_A^2+\frac{1}{4}\sigma_D^2+3D_1+\frac{3}{4}D_2+\frac{3}{16}H^*$
Sn(F4:5)	$\frac{7}{4}\sigma_A^2+\frac{1}{16}\sigma_D^2+3.25D_1+\frac{25}{32}D_2+\frac{3}{64}H^*$	$\frac{15}{8}\sigma_A^2+\frac{1}{8}\sigma_D^2+3.5D_1+\frac{7}{8}D_2+\frac{7}{2\sigma_A^2+4D_1+D_264}$
S∞	$2\sigma_A^2+4D_1+D_2$	$2\sigma_A^2+4D_1+D_2$

Expected Mean Squares

The AOV tables cannot be interpreted without understanding the expected sources of variability represented by the Mean Squares. In the case of balanced field plot designs with only a few sources of variation, the expected mean squares are easily determined. If a particular design involves many sources of random and fixed factors, students have found the approach of Lorenzen and Anderson (1993, Design of Experiments: A No-Name Approach. p 71-72) to be useful.

1. Write the terms of the model with associated subscripts down the left side of the page. Across the top, write the single letter subscripts (i,j,k, etc.). Above each subscript, place either F or R if the factor associated with that transcript is fixed or random. Above that, place the number of levels associated with that subscript (I, J, K, etc.).
2. Enter a 1 in every slot where the subscript at the top is contained within brackets in the term at the left.
3. Enter a 0 in every slot where the subscript at the top is fixed and also contained in the term as the left. Enter a 1 in every slot where the subscript at the top is random and also contained in the terms at the left.
4. Fill in the remaining slots with the number of levels at the top of each column.
5. To compute the Expected Mean Squares (EMS) for a given term having df > 0, start at the bottom and work up. Only consider terms whose indices include all the indices in the term whose EMS you are deriving. Compute the coefficient of this term by covering the columns corresponding to the indices in the term whose EMS you are deriving and multiplying the values in the remaining columns. If there is a 0 column that is not covered, this term need not be written in the EMS. A factor is considered fixed and denoted with a Φ only if all of its indices are fixed. Otherwise, it is considered random and denoted by the appropriate σ^2 term.

Using the Algorithm

Notice that this algorithm can be used to compute EMS for all terms in the model, including those that have zero df. A term that has zero df has no expected mean squares. For this reason, we will not compute EMS for terms having zero df even though such terms are in the algorithm to make the EMS of the other terms come out right. Note that this simple algorithm for determining the EMS in an AOV assumes that the data are balanced, i.e., each of the sources of variability (model parameters) have data for all levels, i, j, and k.

Step 1

The phenotype Y for this typical field trial will be something like in Equation 15:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k},$$

Equation 15 Linear model for phenotype,

where:
Y_{ijk} = phenotypic measure of trait for the j^{th} genotype in the k^{th} block nested within the i^{th} environment,
μ = overall mean,
E_i = i^{th} environment,
$B(E)_{(i)k}$ = represents the k^{th} block nested within the i^{th} environment,
G_j = the j^{th} genotype,
GE_{ij} = interaction effect between the j^{th} genotype and the i^{th} environment,
$\varepsilon_{(ij)k}$ = the residual for genotype j in the k^{th} block nested within the i^{th} environment.

Notice that this algorithm can be used to compute EMS for all terms in the model, including those that have zero df. A term that has zero df has no expected mean squares. For this reason, we will not compute EMS for terms having zero df even though such terms are in the algorithm to make the EMS of the other terms come out right. Note that this simple algorithm for determining the EMS in an AOV assumes that the data are balanced, i.e., each of the sources of variability (model parameters) have data for all levels, i, j, and k.

To illustrate, let us consider a slightly more complex but typical RCBD design used by plant breeders to evaluate many genotypes grown in replicates at several environments for purposes of identifying and discarding poor-performing genotypes in a cultivar development project. **Equation 15** will appear for each of the steps below.

Write the terms of the model with associated subscripts down the left side of the page. Across the top, write the single letter subscripts (i,j,k, etc.). Above each subscript, place either F or R if the factor associated with that transcript is fixed or random. Above that, place the number of levels associated with that subscript (I, J, K, etc.).

Factors:

- Factor E – Fixed
- Factor G – Random
- Blocks – Random

Source	E F i	G R j	R R k	EMS
E_i				
$B(E)_{(i)k}$				
G_j				
GE_{ij}				
$\varepsilon_{(ij)k}$				

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k},$$

Step 2

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k},$$

Enter a 1 in every slot where the subscript at the top is contained within brackets in the term at the left.

Factors:

- Factor E – Fixed
- Factor G – Random
- Blocks – Random

Source	**E**	**G**	**R**	EMS
	F	**R**	**R**	
	i	**j**	**k**	
E_i	n/a	n/a	n/a	n/a
$B(E)_{(i)k}$	1	n/a	n/a	n/a
G_j	1	n/a	n/a	n/a
GE_{ij}	1	1	n/a	n/a
$\varepsilon_{(ij)k}$	1	1	1	n/a

Step 3

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k},$$

Enter a 0 in every slot where the subscript at the top is fixed and also contained in the term as the left. Enter a 1 in every slot where the subscript at the top is random and also contained in the terms at the left.

Factors:

- Factor E – Fixed
- Factor G – Random
- Blocks – Random

Source	E	G	R	EMS
	F	R	R	
	i	j	k	
E_i		n/a	n/a	n/a
$B(E)_{(i)k}$	0	1	n/a	n/a
G_j	1	n/a	n/a	n/a
GE_{ij}	1	1	n/a	n/a
$\varepsilon_{(ij)k}$	1	1	1	n/a

Step 4

The phenotype Y for this typical field trial will be something like this:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k},$$

Fill in the remaining slots with the number of levels at the top of each column.

Factors:

- Factor E – Fixed
- Factor G – Random
- Blocks – Random

Source	E	G	R	EMS
	F	R	R	
	i	j	k	
E_i	0	G	R	n/a
$B(E)_{(i)k}$	0	G	1	n/a
G_j	E	1	R	n/a
GE_{ij}	1	1	R	n/a
$\varepsilon_{(ij)k}$	1	1	1	n/a

Step 5

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k},$$

To compute the EMS for a given term having df > 0, start at the bottom and work up. Only consider terms whose indices include all the indices in the term whose EMS you are deriving. Compute the coefficient of this term by covering the columns corresponding to the indices in the term whose EMS you are deriving and multiplying the values in the remaining columns.

If there is a 0 column that is not covered, this term need not be written in the EMS. A factor is considered fixed and denoted with a Φ only if all of its indices are fixed. Otherwise it is considered random and denoted by the appropriate σ^2 term.

Factors:

- Factor E – Fixed
- Factor G – Random
- Blocks – Random

Source	**E**	**G**	**R**	**EMS**
	F	**R**	**R**	
	i	**j**	**k**	
E_i	0	G	R	$\sigma + G\sigma^2_{B(E)} + \sigma^2_E$
$B(E)_{(i)k}$	0	G	1	$\sigma^2 + G\sigma^2_{B(E)}$
G_j	E	1	R	$\sigma^2 + R\sigma^2_{GE} + RE\sigma^2_G$
GE_{ij}	1	1	R	$\sigma^2 + R\sigma^2_{GE}$
$\varepsilon_{(ij)k}$	1	1	1	σ^2

References

Cockerham, C.C. 1983. Covariances of relatives from self-fertilization. Crop Sci. 23: 1177-1180.

Lorenzen, T., and V. Anderson. 1993. Design of Experiments: A No-Name Approach. Routledge & CRC Press.

How to cite this chapter: Beavis, W., K. Lamkey, K. Espinosa, and A. A. Mahama. 2023. Estimates of Variance. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 8: Mating Designs

William Beavis; Kendall Lamkey; and Anthony Assibi Mahama

There are many mating designs developed for the purpose of estimating the magnitude of genetic variability in a reference population. This information is most often useful to the plant breeder who is developing a new breeding program in a new crop species or developing a novel germplasm resource for established crop species. For example, large estimates of additive genetic variability and small estimates of genotype by environment variability suggest that rapid progress from selection can be made with minimal allocation of testing resources. While most recently trained plant breeders will assume responsibilities for established plant breeding programs, most established programs begin with an evaluation of genetic variability using one of the many mating designs. Thus, we feel it is instructive to understand the genetic basis upon which these programs were established.

The choice of mating designs is based on:

1. The natural mode of reproduction and mating flexibilities of the species.
2. The objective(s) in estimating genetic variances such as:
    - General interest in knowledge of gene action for quantitative characters
    - Making a choice among alternative selection and breeding procedures
    - The prediction of response to selection.
3. Joint purposes such as estimating genetic variances and simultaneously selecting among progenies or evaluating hybrid combinations
4. The precision of the estimates.

Learning Objectives

Students will learn about methods used to evaluate the potential for genetic improvement in germplasm with unknown estimates of heritability through the application of the Variance-Covariance principle in various types of mating designs.

Design Setup

Setting up the treatment and experimental designs for mating designs creates unique challenges. Several things need to be considered:

- Ease of making crosses in the species.
- Inbreeding generation of the parents of the crosses.
- The number of parents that will be used (male and female).
- Fixed versus random parents.
- The type of mating design to be used.
- The type of experimental design to be used.
- The environmental design to be used.

Diallel Crosses

Diallel matings (Table 1) are used to make inferences regarding the types of gene effects controlling traits. Diallels are particularly important in cross-pollinated crops and for determining the importance of general combining ability and specific combining ability. Consider the following general mating scheme. This scheme is very similar in structure to the two-way tables we have seen for studying interactions.

Table 1 General mating scheme for diallel

Parents	**P1**	**P2**	**P3**	**P4**	**Pn**	**Totals**
P1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{1n}	$Y_{1.}$
P2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{2n}	$Y_{2.}$
P3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{3n}	$Y_{3.}$
P3	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{4n}	$Y_{4.}$
Pn	Y_{n1}	Y_{n2}	Y_{n3}	Y_{n4}	Y_{nn}	$Y_{n.}$
Totals	$Y_{.1}$	$Y_{.2}$	$Y_{.3}$	$Y_{.4}$	$Y_{.n}$	$Y_{..}$

Number of Diallel Crosses and Entries

Let us consider the number of diallel crosses for n parents with and without reciprocal crosses. The number of entries is the number that would have to be evaluated if the parents were included in the experiment (Table 2).

Table 2 Number of crosses and entries possible with different numbers of parents with and without reciprocal matings.

Without Reciprocals			With Reciprocals	
No. of Parents	**No. of Crosses**	**Number of Entries**	**No. of Crosses Including Reciprocals**	**Number of Entries**
n	$\frac{n(n-1)}{2}$	$\frac{n(n-1)}{2}$	$n(n-1)$	$n(n-1)$
5	10	15	20	20
6	15	21	30	30
7	21	28	42	42
8	28	36	56	56
9	36	45	72	72
10	45	55	90	90
11	55	66	110	110
12	66	78	132	132
13	78	91	156	156
14	91	105	182	182
15	105	120	210	210
20	190	210	380	380
50	1225	1275	2450	2450
100	4950	5050	9900	9900

Types of Diallel Analysis

Table 3 Fixed versus random effects analysis based on method and entries makeup.

Model	Method	Parents Included	Crosses	Reciprocals
I (Fixed)	1	Yes	Yes	Yes
I (Fixed)	2	Yes	Yes	No
I (Fixed)	3	No	Yes	Yes
I (Fixed)	4	No	Yes	No
II (Random)	1	Yes	Yes	Yes
II (Random)	2	Yes	Yes	No
II (Random)	3	No	Yes	Yes
II (Random)	4	No	Yes	No

Common Diallel Experiment

The most common diallel experiment is conducted with selected parents, which means a fixed effects analysis where only gene effects and not variance will be estimated (Table 3). The reason for this is simple: It is very hard to sample a population adequately with a diallel. Diallels are useful mating designs, however, despite this limitation.

Therefore, we will not present any analyses related to estimating variance components — only gene effects. This makes this section somewhat out of place, but it fits in with the other mating designs from the structural point of view. The analyses we will present are a combination of those presented by Griffing (1956) and Gardner and Eberhart (1966).

Methods 2 and 4 are the most common types of diallels. Most scientists grow the parents and the crosses or just the crosses. The method 4 analysis is, however, the most commonly used analysis because Griffing assigns specific combining ability effects to the parents per se, and these are hard to interpret relative to Sprague and Tatum's (1942) definitions of general and specific combining ability.

The general model underlying the diallel can be written as in Equation 1:

$$Y_{ijk} = \mu + g_i + g_j + s_{ij} + r_k + e_{ijk},$$

Equation 1 General linear model for diallel design experiments.

where:

μ = the mean,
g_i = the general combining ability effect (marginal effect) of the i^{th} parent,
g_j = the general combining ability effect (marginal effect) of the j^{th} parent,
s_{ij} = the specific combining ability effect (interaction effect) of the i^{th} and j^{th} parents,
r_k = the effect of the k^{th} replication,
e_{ijk} = the residual (or error).

An ANOVA Table for Diallels is shown in Table 4.

Table 4 ANOVA Table for Diallels.

Source	**df**	**df (n = 10)**	**SS**	**MS**	**EMS (Model I – Fixed Effects)**
Replications	$r-1$	$r-1$	n/a	n/a	n/a
Entries	$\frac{n(n-1)}{2}-1$	44	$S2$	$M2$	$\sigma_\varepsilon^2 + r\delta_t^2$
Among Margins	$n-1$	9	$S21$	$M21$	$\sigma_\varepsilon^2 + (\frac{n-2}{n-1})\sum_i g_i^2$
Among cells/ Margins	$\frac{n(n-3)}{2}$	35	$S22$	$M22$	$\sigma_\varepsilon^2 + (\frac{2r}{n(n-3)})\sum_{i\prec j} s_{ij}^2$
Error	$(r-1)(\frac{n(n-1)}{2}-1)$	44 $r-1$	$S1$	$M1$	σ_ε^2

F-Tests

Model I F-Tests: For among cells/margins and among margins are, respectively,

F = \frac{M_{22}}{M_1} \text{, and }

$$F = \frac{M_{21}}{M_1}$$

These F-tests evaluate whether differences among the parents and crosses within parents are significant. Also, it is possible to show that the effects can be estimated using Equation 2:

$$\hat{u} = \frac{2}{n(n-1)} Y..,$$

$$\hat{g}_i = \frac{1}{n(n-2)} (nY_{i.} - 2Y..),$$

$$\hat{s}_{ij} = Y_{ij} - \frac{1}{n-2}(Y_{i.} Y_{.j}) \frac{2}{(n-1)(n-2)} Y..,$$

Equation 2 Formulae for estimating mean, gca, and sca effects.

where:
$\hat{u}$ = estimated mean,
n = number of parents,
$\hat{g}_i$ = estimate of gca effect of genotype i,
$\hat{s}_{ij}$ = estimate of sca effect of genotypes *i* and *j*,
$Y_{..}$ = grand total,
$Y_{i.}$ = sum of parent *i* across all parents,
$Y_{.j}$ = sum of parent *j* across all parents,
Y_{ij} = phenotype of cross *ij* .

The variances of the effects can be estimated with Equation 3:

$$V(\hat{u}) = \frac{2}{n(n-1)} \hat{\sigma}^2_{\bar{Y}}$$

$$V(\hat{g}_i) = \frac{1}{n(n-2)} \hat{\sigma}^2_{\bar{Y}}$$

$$V(\hat{s}_{ij}) = \frac{n-3}{n-1} \hat{\sigma}^2_{\bar{Y}}, (i \neq j)$$

and

$$\hat{\sigma}^2_{\bar{Y}} = \frac{\hat{\sigma}^2_{\varepsilon}}{r} = \frac{M_1}{r}$$

Equation 3 Formulae for estimating variances of estimated of mean, gca, sca effects and error,

where:

$\hat{\sigma}^2_{\bar{Y}}$ = estimates variance of average phenotype,
other terms are as defined previously.

Gardner and Eberhart Diallel Analysis II

The Gardner and Eberhart Analysis II for the diallel is a more general analysis designed for the case of when the diallel includes random mating varieties. The model is best laid out by starting with the following single locus theory for the j^{th} variety and i^{th} locus (Table 5):

Table 5 Frequencies and genotypic values for genotypes.

Frequency	Genotype	Genotypic value
p_{ji}^2	AA	$\mu' a_i$
$2p_{ji}(1-p_{ji})$	Aa	$\mu' \delta_i$
$(1-p_{ji})^2$	aa	$\mu' - a_i$

Where, $\mu' = \dfrac{AAaa}{2}$.

The population mean can be written as in Equation 4:

\mu' \sum_i(2p_{ji}-1) \alpha_{i} + 2\sum_i(p_{ji}-p_{ji}^{2})\delta_i

Equation 4 Formula for calculating the population mean,

where:
$p, 1-p$ = frequencies of the two allele.
μ' = average genotypic value,
α_i, δ_i = coded genotypic values.

Equations

Let $a'_j = \sum_i (2p_{ji} - 1)\alpha_i; \quad \bar{a} = \dfrac{1}{n}\sum_j a'_j; \quad a_j = a'_j - \bar{a}; and \quad \mu = \mu' + \bar{a}.$

Similarly, let $d_j = 2\sum_i (p_{ji} - p_{ji}^2)\delta_i; \quad h_{jj'} = \sum_i (p_{ji} - p_{ji}^2)^2 \delta_i.$

Then, the population mean can be written as in Equation 5:

$$\text{pop mean} = \mu' + a'_j + d_j = \mu + a_j + d_j$$

Equation 5 Formula for the population mean.

where:
μ = the mean,
μ' = average genotypic value of **AA** & **aa**,
a =genotypic value of **AA**, **aa** genotypes,
d = genotypic value of **Aa** genotype.

A population cross mean can be written as in Equation 6:

$$C_{jj'} = \mu + \frac{1}{2}(a_j a_{j'}) + \frac{1}{2}(d_j + d_{j'} + h_{jj'}).$$

Equation 6 Formula for the population cross mean.

where:
$C_{jj'}$ = mean of the "variety cross" from two parents,
μ = the mean of all crosses,
a = the additive effect,
d = the dominance effect,
$h_{jj'}$ = the heterosis effect.

If the varieties, varieties selfed, population crosses, population crosses selfed, and population crosses random mated are included in the analysis, then all of these genetic effects can be estimated. Usually, this **is not** the case, and only varieties and variety crosses are included in the analysis, which are confounded, and they have to be estimated together. We can then define the following parameters:

The mean of all parental varieties included in the analysis is written as in Equation 7:

$$\mu_v = \mu \frac{1}{n} \sum_j d_j; \quad = \mu + \bar{d}$$

Equation 7 Formula for mean when all parental varieties are included.

where:
μ_v = the mean of all parental varieties included in the analysis,
μ = the mean,

d = dominance effect,
$\bar{d}$ = average dominance effects.

The variety effect when parents are included in the analysis is written as in Equation 8:

$$v_j = a_j + (d_j + \bar{d})$$

Equation 8 Formula for estimating variety effects with parents included.

where:
v = the variety effect when parents are included,
a, d are as defined previously.

Models

We can then fit the following four models to the data (Equation 9):

$$\begin{aligned}
Y_{jj'} &= \mu_v + \frac{1}{2}(v_j v_{j'}) \\
Y_{jj'} &= \mu_v + \frac{1}{2}(v_j v_{j'}) + \gamma\bar{h} \\
Y_{jj'} &= \mu_v + \frac{1}{2}(v_j v_{j'}) + \gamma\bar{h} + \gamma(h_j h_i') \\
Y_{jj'} &= \mu_v + \frac{1}{2}(v_j v_{j'}) + \gamma\bar{h} + \gamma(h_j h_{j'}) + \gamma s_{jj'}
\end{aligned}$$

Equation 9 Linear models for estimating different genetic effects on phenotype,

where:
$y = \left\{ \begin{array}{l} 0 \; when \; j = j' \\ 1 \; when \; j \neq j' \end{array} \right\}$,
$Y_{jj'}$ = phenotype of j by j' progeny.

ANOVA Table

The following ANOVA table can be written as in Table 6:

Table 6 ANOVA Table for Gardner and Eberhart Diallel Analysis II.

Source	df	Sum of squares
Populations	$[n(n-1)/2]-1$	S'
Varieties (v_j)	$n-1$	$S'_1 = (B'G)_1 - CF$
Heterosis$(h_{jj'})$	$n(n-1)/2$	$S'_1 = (B'G)_1 - (B'G)$
Average $(\bar{h})$	1	$S'_{21} = (B'G)_2 - (B'G)_1$
Variety (h_j)	$n-1$	$S'_{22} = (B'G)_3 - (B'G)_2$
Specific $(s_{jj'})$	$\frac{n(n-3)}{2}$	$S'_{23} = (B'G)_4 - (B'G)_3$

Equivalent Analysis

An equivalent analysis can be made with just the crosses as follows:

The mean of crosses in the diallel can be estimated as follows:

$$Let: \ \mu_c = \mu_v \bar{h} = \mu \bar{d} \bar{h};$$

The variety effect in crosses = general combining ability effect = $g_j = \frac{1}{2} v_j h_j$, then the mean of crosses is written as in Equation 10:

$$C_{jj'} = \mu_c + g_j + g'_j + s_{jj'}$$

Equation 10 Model for analysis with only crosses included.

where:

$s_{jj'}$ = specific heterosis from variety j by variety j' mating,

$\sum_j g_j = 0$,

$\sum_{j \neq j'} s_{jj'} = 0$.

Analysis III of Gardner and Eberhart

The following ANOVA (Table 7) can be written (Analysis III of Gardner and Eberhart)

Table 7 ANOVA Table for Gardner and Eberhart Diallel Analysis III.

Source	Degrees of Freedom	Sum of squares
Population	$[n(n-1)/2]-1$	n/a
Varieties (v_j)	$n-1$	S''_1
Varieties vs. crosses ($\bar{h}$)	1	S''_2
Crosses($x_{jj'}$)	$[n(n-1)/2]-1$	S''_3
GCA (g_j)	$n-1$	S''_{31}
SCA ($s_{jj'}$)	$\frac{n(n-3)}{2}$	S''_{32}

The analysis of Crosses, GCA, and SCA is all that can be done if only the crosses are included in the analysis. This analysis is equivalent to the Model 4 analysis of Griffing. If varieties or parents are also included, then the analysis, Varieties, and Varieties vs. Crosses can also be calculated.

Analysis III is related to Analysis II in the following ways that the ($s_{jj'}$) are the same in the two analyses $S'_{21} = S''_2$, meaning that average heterosis is simply a contrast of the mean of the varieties with the mean of the crosses (Equation 11).

$$S''_1 + S''_{31} = S'_1 + S'_{22}, \text{since } g_j = \frac{1}{2} v_j h_j$$

Equation 11 Formula for calculating average heterosis.

where:
S''_1 = effects of variety 1,
S''_2 = effects of variety 2,
S''_3 = effects of crosses,
S''_{31} = GCA effects,
S''_{32} = SCA effects.

North Carolina Design I

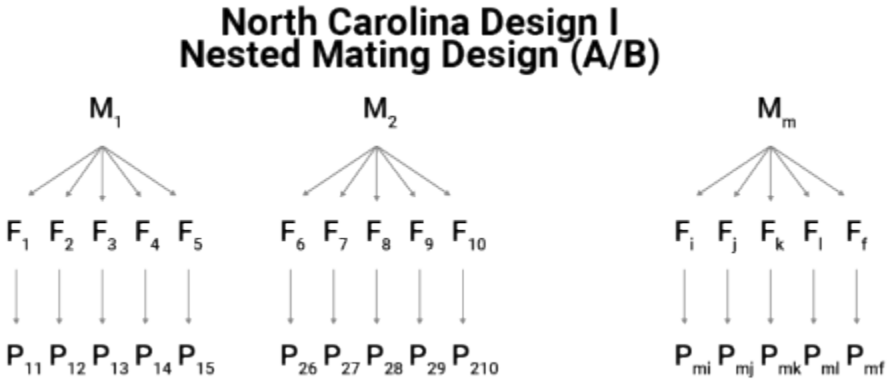


Fig. 1 North Carolina Design I (NC I), Nested Mating Design (A/B).

- Consider m male plants:
- each of which is mated to *f* female plants,
- to produce *n* full-sib families within each male,
- for a total of mf half-sib families.
- There is a total of m half-sib families.
- Different female plants are used to cross with each male.
- The progeny P are grown in a replicated experiment design.

The model for analysis is written as in Equation 12:

$$Y_{ijk} = \mu + m_i + f_{ij} + r_k + e_{ijk}$$

Equation 12 General linear model for NC I experiments.

where:
μ = the mean,
m_i = the effect of male **i**,
f_{ij} = the effect of female **j** when crossed to male **i**,
r_k = replication effect,
e_{ijk} = the residual.

ANOVA Table

Then the ANOVA (Table 8) can be written as:

Table 8 ANOVA Table for Gardner and Eberhart Diallel Analysis III.

Source of Variation	d.f.	MS	EMS
Replications	$r-1$	n/a	n/a
Males	$m-1$	**M4**	$\sigma + r\sigma^2_{f(m)} + rf\sigma^2_m$
Females/Males	$m(f-1)$	**M3**	$\sigma^2 + r\sigma^2_{f(m)}$
Error	$(mf-1)(r-1)$	**M2**	σ^2
Total	$rmf-1$	n/a	n/a
Within	$rmf(k-1)$	**M1**	σ^2_W

The table can be rewritten in terms of the covariance of relatives as follows (Table 9):

Table 9 ANOVA Table for Gardner and Eberhart Diallel Analysis III.

Source of Variation	d.f.	MS	EMS
Replications	$r-1$	n/a	n/a
Males	$m-1$	**M4**	$\sigma^2 + r[Cov(FS) - Cov(HS)] + rfCov(HS)$
Females/Males	$m(f-1)$	**M3**	$\sigma^2 + r[Cov(FS) - Cov(HS)]$
Error	$(mf-1)(r-1)$	**M2**	σ^2
Total	$rmf-1$	n/a	n/a
Within	$rmf(k-1)$	**M1**	σ^2_W

$$And \quad \sigma^2_u = \frac{\sigma^2_{\mu\varepsilon}[\sigma^2_G - Cov(FS)]}{k}$$

Variance Estimates

Estimation of variance of the various components is as in Equation 13:

$$\hat{\sigma}^2_m = Cov(HS) = \frac{M_4 - M_3}{rf}; \quad \hat{\sigma}^2_{f(m)} = Cov(FS) - Cov(HS) = \frac{M_4 - M_3}{rf}$$

Equation 13 Formulae for calculating variance components for NC I.

where:
$\hat{\sigma}^2_m$ = variance of males,
$\hat{\sigma}^2_{f(m)}$ = variance of males within females,
$Cov(HS)$ = covariance of half-sibs,
$Cov(FS)$ = covariance of full-sibs.

So, ignoring epistasis, the variances are written as in Equation 14:

$$\hat{\sigma}^2_m = \frac{(1-F_m)}{4}\sigma^2_A; \quad \hat{\sigma}^2_{f(m)} = \frac{(2+F_m+F_f)}{4}\sigma^2_A + \frac{(1+F_m)(1+F+f)}{4}\sigma^2_D - \left[\frac{(1+F_m)}{4}\sigma^2_A\right]$$

.

Equation 14 Alternative formulae for calculating variance components.

where:
σ^2_A = the additive variance,
σ^2_D = the dominance variance,
F_m = the inbreeding coefficient of the male parent,
$\hat{\sigma}^2_{f(m)}$ = estimated variance of male within females,
F_f = the inbreeding coefficient of the female parent,
f = effects of females.

Consider the case when all the parents are noninbred, i.e., $F_m = F_f = 0$. The variances are written as in Equation 15:

$$\hat{\sigma}^2_m = \frac{1}{4}\sigma^2_A; \quad \hat{\sigma}^2_{f(m)} = \frac{1}{2}\sigma^2_A + \frac{1}{4}\sigma^2_D - [\frac{1}{4}\sigma^2_A] = \frac{1}{4}(\sigma^2_A + \sigma^2_D); \quad \hat{\sigma}^2_A = 4\hat{\sigma}^2_m; \quad \hat{\sigma}^2_D = 4(\hat{\sigma}^2_{f(m)} - \sigma^2_m)$$

Equation 15 Formulae for calculating variance component when all the parents are noninbred.

where:
$terms$ are as defined previously.

When both the male and female parents are inbred, i.e., $F_m = F_f = 1$, then the variances can be estimated as written in Equation 16:

$$\hat{\sigma}^2_m = \frac{1}{2}\sigma^2_A; \quad \hat{\sigma}^2_{f(m)} = \sigma^2_A + \sigma^2_D - [\frac{1}{2}\sigma^2_A] = \sigma^2_A + \frac{1}{2}\sigma^2_D; \quad \hat{\sigma}^2_A = 2\sigma^2_m; \quad \hat{\sigma}^2_D = (\hat{\sigma}^2_{f(m)} - 2\hat{\sigma}^2_m)$$

Equation 16 Formulae for calculating variance component when all the parents are inbred.

where:
$terms$ are as defined previously.

North Carolina Design II

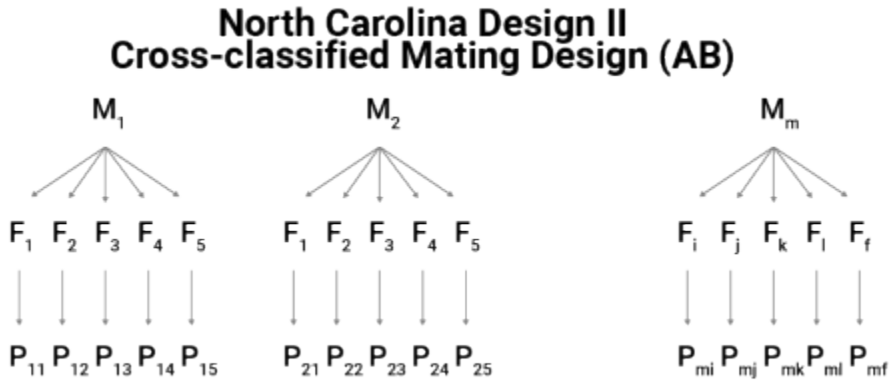


Fig. 2 North Carolina Design II (NC II), Mating Design (AB).

- Consider m male plants,
- each of which is mated to f female plants,
- to produce f full-sib families within each male,
- for a total of mf half-sib families.
- There is a total of m f half-sib families.
- The same female plants are crossed with each male.
- The progeny P are grown in a replicated experiment design.

The design is related to the diallel and another simpler way to represent the design is (Table 10):

Table 10 North Carolina Design II arrangement.

Parents	**M1**	**M2**	**M3**	**M4**	**Totals**
F5	Y_{15}	Y_{25}	Y_{35}	Y_{45}	$Y_{.5}$
F6	Y_{16}	Y_{26}	Y_{36}	Y_{46}	$Y_{.6}$
F7	Y_{17}	Y_{27}	Y_{37}	Y_{47}	$Y_{.7}$
F8	Y_{18}	Y_{28}	Y_{38}	Y_{48}	$Y_{.8}$
Totals	$Y_{1.}$	$Y_{2.}$	$Y_{3.}$	$Y_{4.}$	$Y_{..}$

Model

The model for analysis is written as in Equation 17:

$$Y_{ijk} = \mu + m_i + f_{ij} + mf_{ij} + r_k + e_{ijk}$$

Equation 17 Formulae for calculating variance components for NC II.

where:
μ = mean,
m_i =the effect of male **i**,
f_j = the effect of female **j**,
mf_{ij} = the interaction effect of female **j** when crossed to male **i**,
r_k = replication effect,
e_{ijk} =the residual.

ANOVA Table

The ANOVA is shown in Table 11.

Table 11 ANOVA Table for North Carolina Design II.

Source of Variation	d.f.	MS	EMS
Replications	$r-1$	n/a	n/a
Males (M)	$m-1$	**M5**	$\sigma^2 + r\sigma^2_{mf} + r\sigma^2_m$
Females (F)	$f-1$	**M4**	$\sigma^2 + r\sigma^2_{mf} + rm\sigma^2_f$
M x F	$(m-1)(f-1)$	**M3**	$\sigma^2 + r\sigma^2_{mf}$
Error	$(mf-1)(r-1)$	**M2**	σ^2
Total	$rmf-1$	n/a	n/a
Within	$rmf(k-1)$	**M1**	σ^2_w

Covariance of Relatives

The table can be rewritten in terms of the covariance of relatives as follows (Table 12):

Table 12 A general ANOVA table for covariance of relatives.

Source of Variation	d.f.	MS	EMS
Replications	$r-1$	n/a	n/a
Males (M)	$m-1$	M5	$\sigma^2 + r[Cov(FS) - Cov(HS_m) - Cov(HS_f)] + rfCov(HS_m)$
Females (F)	$f-1$	M4	$\sigma^2 + r[Cov(FS) - Cov(HS_m) - Cov(HS_f)] + rmCov(HS_f)$
M x F	$(m-1)(f-1)$	M3	$\sigma^2 + r[Cov(FS) - Cov(HS_m) - Cov(HS_f)]$
Error	$(mf-1)(r-1)$	M2	σ^2
Total	$rmf-1$	n/a	n/a
Within	$rmf(k-1)$	M1	σ_w^2

$$\text{And} \quad \sigma_u^2 = \frac{\sigma_{u\varepsilon}^2[\sigma_G^2 - Cov(FS)]}{k}$$

Estimation

Variance components are estimated as written in Equation 18:

$$\hat{\sigma}_m^2 = Cov(HS_m) = \frac{M5 - M3}{rf}; \quad \hat{\sigma}^2 = Cov(HS_f) = \frac{M4 - M3}{rm}; \quad \hat{\sigma}_{mf}^2 = [Cov(FS) - Cov(HS_m) - Cov(HS_f)] = \frac{M3 - M2}{r}$$

Equation 18 Formula for estimating covariance of relatives,

where:
$\hat{\sigma}_m^2$ = estimated variance of males,
$\hat{\sigma}_f^2$ = estimated variance of females,
$\hat{\sigma}_{mf}^2$ = estimated variance of male by female cross (full-sibs),
$Cov(FS)$ = covariance of full-sibs,
$Cov(HS_m)$ = covariance of half-sibs with common male,
$Cov(HS_f)$ = covariance of half-sibs with common female.

Variance Estimates

Ignoring epistasis, variance components are estimated as written in Equation 19.

$$\hat{\sigma}^2 m = \frac{(1+F_m)}{4}\sigma^2_A; \quad \hat{\sigma}^2_f = \frac{(1+F_f)}{4}\sigma^2_A; \quad \hat{\sigma}^2_{mf} = \frac{(2+F_m+F_f)}{4}\sigma^2_A + \frac{(1+F_m)(1+F_f)}{4}\sigma^2_D - \left[\frac{(1-F_m)}{4} + \frac{1+F_f)}{4}\right]\sigma^2_A$$

Equation 19 Formula for calculating variance estimates, ignoring epistasis.

where:
$terms$ are as defined previously.

Consider the case when all the parents are noninbred, i.e., $F_m = F_f = 0$. Variance components are estimated as written in Equation 20:

$$\hat{\sigma}^2_m = \frac{1}{4}\sigma^2_A; \quad \hat{\sigma}^2_f = \frac{1}{4}\sigma^2_A; \quad \hat{\sigma}^2_{mf} = \cancel{\frac{1}{2}\sigma^2_A} + \frac{1}{4}\sigma^2_D - \cancel{[\frac{1}{4} + \frac{1}{4}]\sigma^2_A} = \frac{1}{4}\sigma^2_D; \quad \hat{\sigma}^2_A = 4(\frac{\hat{\sigma}^2_m + \hat{\sigma}^2_f}{2}) = 2(\hat{\sigma}^2_m + \hat{\sigma}^2_f); \quad \sigma^2_D = 4\hat{\sigma}^2_{mf}$$

Equation 20 Formulae for estimating variance components when male and female parents are noninbred,

where:
$terms$ are as defined previously.

When both the male and female parents are inbred, i.e., $F_m = F_f = 1$, then variance components are estimated as written in Equation 21:

$$\hat{\sigma}^2_m = \tfrac{1}{2}\sigma^2_A; \quad \hat{\sigma}^2_f = \tfrac{1}{2}\sigma^2_A; \quad \hat{\sigma}^2_{mf} = \sigma^2_A + \sigma^2_D - [\tfrac{1}{2} + \tfrac{1}{2}]\sigma^2_A = \sigma^2_D; \quad \sigma^2_A = \hat{\sigma}^2_m + \hat{\sigma}^2_f; \quad \sigma^2_m = \hat{\sigma}^2_{mf}$$

Equation 21 Formulae for estimating variance components when male and female parents are inbred,

where:
$terms$ are as defined previously.

North Carolina Design III

The main use of Design III is for estimating the average degree of dominance.

North Carolina Design III Backcross Design is shown in Fig. 3.

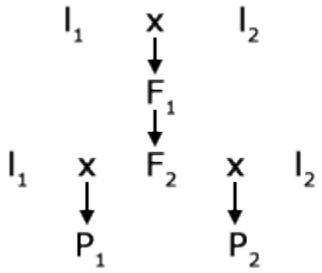


Fig. 3 North Carolina Design III (NC III), Mating Design (F2 backcrossed to inbred parents).

This design involves crossing two inbred lines and obtaining the F_1 and F_2 generations. An individual F_2 plant is then backcrossed to each of the inbred parents generating a pair of progeny using the F_2 plants and pollen parents. Then for n F_2 plants, there are 2n progenies produced, and the model is as written in Equation 22:

$$Y_{ijk} = \mu + l_i + m_j + ml_{ij} + r_k + e_{ijk}$$

Equation 22 Linear model for estimating average degree of dominance.

where:
μ = the mean,
l_i = contrast of the inbred parents ***i*** = **1, 2**
m_j = the effect of $\mathbf{F_2}$ parent ***j***,
ml_{ij} = the interaction effect of inbred parent ***i*** and $\mathbf{F_2}$ plant ***i***,
r_k = replication effect,
e_{ijk} =the residual.

ANOVA Table

An ANOVA Table for North Carolina Design III is shown in Table 13.

Table 13 ANOVA Table for North Carolina Design III.

Source of Variation	d.f.	MS	EMS
Replications	$r-1$	n/a	n/a
Inbred Lines	1	n/a	n/a
F_2 parents	$n-1$	**M3**	$\sigma^2 + 2r\sigma^2_m$
F_2 parent x inbred line	$n-1$	**M2**	$\sigma^2 + r\sigma^2_{ml}$
Error	$(2n-1)(n-1)$	**M1**	σ^2
Total	$rmf-1$	n/a	n/a

Estimation

Variance components are estimated as written in Equation 23:

$$\hat{\sigma}^2_m = \frac{M3-M1}{2r} = \frac{1}{8}\sum_i a_i^2, \quad \hat{\sigma}^2_{ml} = \frac{M2-M1}{r} = \frac{1}{4}\sum_1 d_i^2$$

Equation 23 Formulae for estimating effects of parents and the interaction effect of inbred parents and **F_2** plants,

where:
$\sum$ = the summation is over i loci,
$\hat{\sigma}^2_m$ = estimated variance of **F_2** parents ***i*** = **1, 2**
$\hat{\sigma}^2_{ml}$ = estimated variance of the interaction effect of the parent and **F_2** plant.

F-Tests

Remember that in an F_2 population, additive and dominance variances are written as in Equation 24

$$\sigma^2_A = \frac{1}{2}\sum_i a_i^2, and \quad \sigma^2_D = \frac{1}{4}\sum_i d_i^2$$

Equation 24 Formula for total genetic variance for and F_3 population,

so that variance components are estimated as written in Equation 25,

$$\hat{\sigma}^2_m = \frac{1}{4}\sigma^2_A, and \quad \hat{\sigma}^2_{ml} = \sigma^2_D$$

Equation 25 Formula for total genetic variance for and F_3 population.

where:
$\hat{\sigma}^2_m; \hat{\sigma}^2_{ml}$ are as defined previously,
σ^2_A = additive variance,
σ^2_D = dominance variance.

Note that this design is very specialized for the specific case of F_2 populations when **p** = **q** = **0.5**. This design provides exact F-tests of two important hypotheses:

1. The null hypothesis of no dominance. This is tested by: F=M2/M1, and if this F-test is significant, then it means that $d \succ |0|$ and there is no dominance.
2. The null hypothesis is that dominance is complete. If there is complete dominance, then the ratio, M3/M2=1.

A significant departure of this ratio from one indicates that d departs significantly from 1.

References

Gardner, C. O, and A. S., Eberhart. 1966. Analysis and interpretation of the variety cross diallel and related populations. Biometrics, 22(3):439-52.

Griffing, B. 1956. Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. Aust. J. Biol. Sci. 9: 463-93.

Sprague, G. F., and Tatum, L. A. 1942. General Vs Specific Combining Ability in Single Crosses of Corn. J. Amer. Soc. Agron. 34: 923-32.

How to cite this chapter: Beavis, W., K. Lamkey, and A. A. Mahama. 2023. Mating Designs. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 9: Selection Response

William Beavis; Kendall Lamkey; and Anthony Assibi Mahama

Selection is the crux of crop improvement and makes use of the art and science concepts to ensure the success (gains derived) of a breeding program or specific project. This chapter explains principles and practices related to genetic gain.

Learning Objectives

- Explain the role of selection on genetic improvement.
- Explain all of the components of realized and predicted genetic gain.
- Explain why realized genetic gains are always less than predicted genetic gains.
- Explain the role of replication in multi-environment tests on predicted and realized genetic gains.

Underlying Theory of Selection

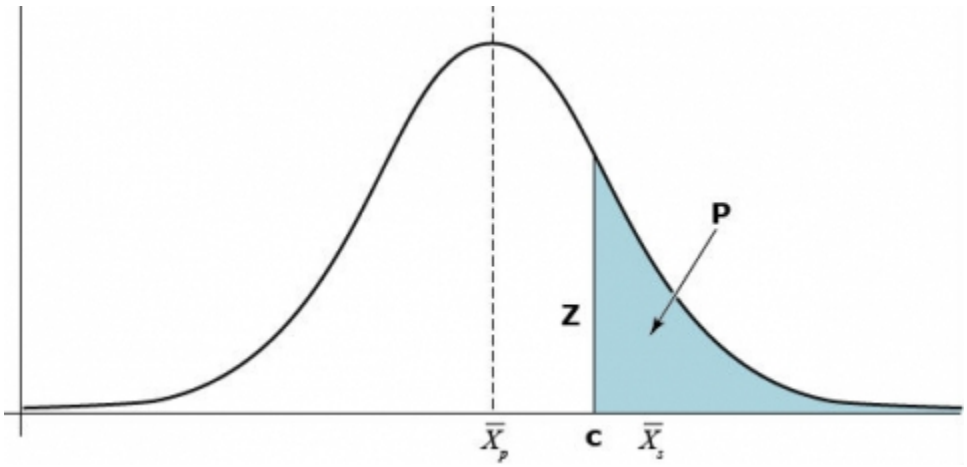


Fig. 1 The normal distribution.

Let $\bar{X}_p$ be the mean phenotypic value of a quantitative trait that is normally distributed in a large random mating population (Fig. 1). Also, designate $\bar{X}_s$ as the mean of a selected proportion **P** of this population, where **c** is the truncation point of selection and **Z** is the height of the ordinate at the selection truncation point.

The selection differential is defined as in Equation 1:

$$S = \bar{X}_s - \bar{X}_p.$$

Equation 1 Formula for calculating selection differential.

where:
S = selection differential,
$\bar{X}_s$ = mean of selected proportion,
$\bar{X}_p$ = mean of the population.

If σ_p^2 is the phenotypic variance in the population, then the standardized selection differential can be written as in Equation 2:

$$i = \frac{S}{\sigma_p} = \frac{\bar{X}_s - \bar{X}_p}{\sigma_p}.$$

Equation 2 Formula for calculating standardized selection differential.

where:
i = standardized selection differential; also, is the number of standard deviations represented by the selection differential, **S**,
σ_p = square root of phenotypic variance in the population,
$\bar{X}_s$ = mean of selected proportion,
$\bar{X}_p$ = mean of the population.

Selection Response

While $\bar{X}_s$ may be distinctive relative to $\bar{X}_p$, of greater interest are the phenotypes of the progeny derived from crosses among the selected parents $\bar{X}_s$. The predicted response of progeny to the selection of their parents can be derived from the relationship between parent and offspring as follows (Fig. 2). Designate ***R*** as the response to selection measured in the offspring (represented as a deviation from the population mean). ***S*** is the selection differential (represented as a deviation from the population mean) as described in the previous section.

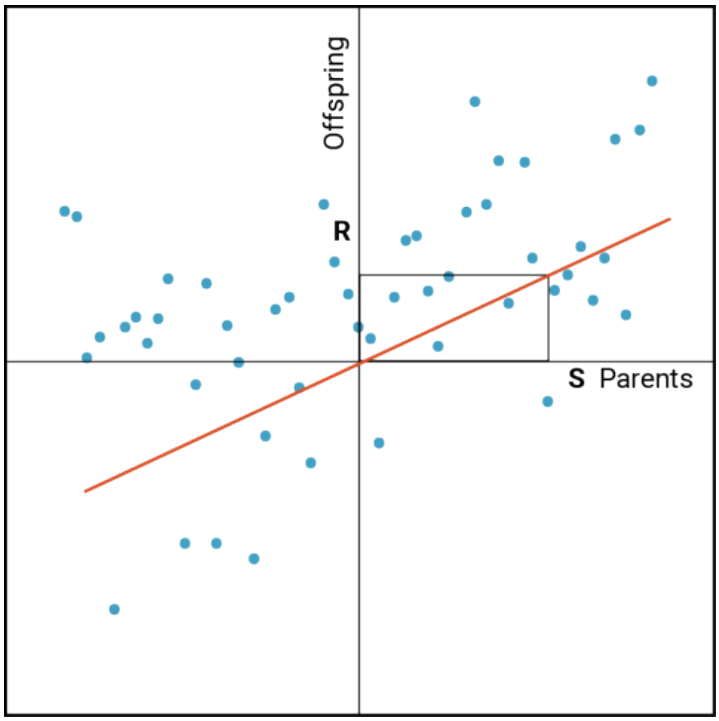


Fig. 2 Parent-offspring regression plot.

Genetic Gain

The response to selection (R) can be written simply as in Equation 3:

$$R = b_{o\bar{p}}\, S.$$

Equation 3 Formula for calculating response to selection.

where:
R = response to selection,
$b_{o\bar{p}}$ = the regression coefficient of offspring on the mid-parent value,
S = the selection differential.

The regression coefficient of offspring on the mid-parent value can be written as in Equation 4:

$$b_{o\bar{p}} = \frac{Cov(o,\bar{p})}{Var(\bar{p})} = \frac{Cov(o,\bar{p})}{\frac{1}{2}\sigma_P^2}$$

Equation 4 Formula for calculating the regression coefficient.

where:

$Cov(o, \bar{p})$ = the covariance of offspring on mean of parents,
$Var(\bar{p})$ = the variance of mid-parent,
σ_P^2 = the phenotypic variance.

Equation 4 is written that way because:

$$Var(\bar{p}) = Var(\frac{P_m + P_f}{2}) = \frac{1}{4}Var(P_m + P_f) = \frac{1}{2}\sigma_P^2$$

Equation 5 Formula for calculating the mean of parent phenotype.

where:
P_m = the phenotype of the male parent,
P_f = the phenotype of the female parent.

Also, we can show that: $Cov(o, \bar{p}) = \frac{1}{2}\sigma_A^2$, so $b_{O\bar{P}} = \frac{\frac{1}{2}\sigma_A^2}{\frac{1}{2}\sigma_P^2} = \frac{\sigma_A^2}{\sigma_P^2} = h^2$

Therefore, R is written as in Equation 6:

$$R = h^2 S = ih^2\sigma_p = \Delta G_c$$

Equation 6 Formula for calculating the change in genetic gain.

where:
ΔG_c = the rate of genetic gain per cycle,
h^2 = the narrow sense heritability,
σ_p = the standard deviation of the phenotype,
other terms are as defined previously.

R is the selection response or Genetic Gain, as Lush defined it in 1940. This equation for ΔG, also known as the **Breeder's Equation**, based on the regression of offspring values on mid-parent values, is difficult to apply directly to plant breeding systems because plant breeders typically evaluate hundreds of replicated individuals representing thousands of genotypes grown in replicated plots in dozens to hundreds of environments. Unlike most animal systems, it is possible to replicate progeny genotypes due to the diversity of reproductive biology that is available to plant breeders: clonal propagation, doubled haploids, and tolerance to inbreeding through self-pollinations for multiple generations. In the last example, the response units can be several generations removed from the parental (crossing) generation. The type of reproductive biology will affect the details of how we estimate the response to selection, $\mathbf{\Delta G_c}$, also referred to as the "**Rate of Genetic Gain**", per cycle.

Heritability on an Entry-Mean Basis

Recall plant breeders often report heritability from field experiments on an entry-mean basis represented as in Equation 7:

$$H = \frac{\sigma^2_G}{\sigma^2_G + \left(\frac{\sigma^2_{GE}}{E}\right) + \left(\frac{\sigma^2_\varepsilon}{rE}\right)}$$

Equation 7 Formula for calculating heritability on entry mean basis.

where:
σ^2_G = the genotypic variance,
σ^2_{GE} = the genotype by environment variance,
σ^2_ε = the residual or error variance,
r = the number of replications
E = number of environments.

Although Equation 7 is similar to Lush's **broad sense heritability**, it is not exactly the same concept because it can be 'adjusted' by adding replicates and environments to reduce the impact of σ^2_{GE} and ε on the estimated phenotypic variance.

The problem for plant breeders is that the concept of evaluating individual plants and the performance of their progeny to obtain an estimate of heritability simply is of no practical use for most crops where plot performance is the basis for selection. Hanson attempted to address this by framing the multiple concepts of heritability within the context of genetic gain (1963).

Hanson defined heritability as "the fraction of the selection differential expected to be gained when selection is practiced on a defined reference unit." Given the standard definition for selection response is $\Delta G = i\frac{\sigma^2_G}{\sigma^2_{\bar{y}}}\sigma_{\bar{y}} = ih^2\sigma_{\bar{y}} = R$, we can then solve for h^2 using the expression in Equation 8:

$$h^2 = \frac{R}{i\sigma_{\bar{y}}}$$

Equation 8 Formula for estimating realized heritability,

where:
terms are as defined previously.

That is the **standardized response to selection** or **realized heritability**.

Context of Heritability

Within the framework of genetic gain, Hanson defined heritability in such a manner as to be consistent with the original concept while at the same time taking into consideration that it has little meaning unless the selection units (entry means) and response units are defined. Thus, when plant breeders wish to communicate information about heritability, they should specify:

1. A reference population of genotypes.
2. A reference population of environments. i.e., the target environments.
3. Selection units
4. Response units

This context emphasizes the purpose of obtaining variance component estimates, usually for the purpose of comparing genetic gains (ΔG) under various possible breeding procedures. The results are used to make decisions about which procedure to employ. Indeed, it is in this context that variance components of heritability are used as "plug-in values" (Sprague and Eberhart, 1977) for a six-step decision-making algorithm that uses ΔG as an arbiter for comparing breeding methods (Fehr, 1994; Chapter 17). Actually, this back-of-the-envelope algorithm is fairly insensitive to the estimated heritability values, and there are more effective means of optimizing genetic gain, number of generations, and costs.

Holland's Synthesis

A thorough review of heritability and how it should be interpreted to compare ΔG by plant breeders was given by Holland et al, (2003). The review was essentially an update to a review by Nyquist (1991), where the updates were based on computational techniques, REML in particular, for obtaining appropriate estimates of variance components. He indicated that plant breeders have traditionally used the method of moments (covered in later slides) to estimate genotypic and phenotypic correlations between traits on the basis of a multivariate analysis of variance (MANOVA) and pointed out the key drawbacks of using the method that include the possibility of obtaining estimates outside of parameter bounds, reduced estimation efficiency, and ignorance of the estimators' distributional properties when data are missing.

With Hanson's response, the response to selection can be rewritten as $R = \beta_{SR} S$, where β_{SR} is the regression coefficient of the response units on the selection units and is equal to

$$\frac{Cov(R,S)}{Var(S)}.$$

Family Structure

Assume our selection and response units are represented by some family structure, say half-sibs, or full-sibs, or recombinant inbred lines, as examples. Also, recall that we can equate the genotypic variance component, designated as *f* for family relationships, to the genetic covariance of relatives. Thus, the *Cov(R,S) = Var(f)*. Also, note that the *Var(S)* is the phenotypic variance among the entry means. Thus, B_{SR} is the proportion of variance among family units relative to the phenotypic variance among entry means. We might refer to this as the heritability of the family units represented in Equation 9:

$$h_f^2 = \frac{\sigma_f^2}{\sigma_p^2}$$

Equation 9 Formula for estimating heritability of family units,

where:
σ_f^2 = the family unit variance,
σ_P^2 = the phenotypic variance.

If the replicated plots consist of half-sibs from a random mating population, then the variance

component among half-sibs on an entry mean basis is equal to the covariance of the half-sibs (Equation 10), ignoring epistasis:

$$Cov(HS) = \frac{1}{4}(1+F)\,\sigma_A^2$$

Equation 10 Formula for estimating covariance of half-sibs, ignoring epistasis,

where:
F = the inbreeding coefficient
σ_A^2 = the additive variance.

Narrow-Sense Heritability of Half-Sibs

Thus, it is possible to utilize the estimated variance components from an ANOVA to estimate a "narrow sense heritability", h^2, by simply multiplying this variance component by 4/(1+F) and plugging the value into Equation 7 as in Equation 11; all terms are as defined previously:

$$h^2 = \frac{\sigma_A^2}{\sigma_G^2 + \left(\frac{\sigma_{GE}^2}{E}\right) + \left(\frac{\sigma_e^2}{rE}\right)}$$

Equation 11 Formula for estimating narrow sense heritability of half-sibs.

Notice that this is not the same as the original narrow sense heritability as defined by Lush (1940), but is a narrow sense heritability for a population of half-sibs.

Next, consider the numerator in Equation 11 above. In the case of half-sibs, we have learned that the variance of family units is represented as in Equation 12.

$$\sigma_f^2 = Cov(HS) = \sigma_{HS} = \frac{(1+F)}{4}\sigma_A^2 + \left[\frac{(1+F)}{4}\right]^2 \sigma_{AA}^2 + \ldots$$

Equation 12 Alternative formula for estimating family units variance,

where:
σ_{AA}^2 = the additive by additive interaction variance.

Covariance Estimation

Again, if the data are not balanced, the variance component will not be estimated correctly unless REML is used. Let us assume that we obtain a 'best' estimate for σ_{HS}, either because our data are balanced or we have used REML. Should we use the previous equation for the Cov(R,S)? To answer this, we have to recognize that there is a genetic relationship between selection units and response units, i.e., there is a pedigree relationship or coefficient of coancestry between the selection and response units, and Equation 12 does not take this into consideration. In the case where both selection units and response units are half-sibs, the Cov(R,S) is represented as in Equation 13

$$Cov(R,S) = \frac{(1+F)}{4}\sigma^2_A + \frac{1}{32}[1+F]^2\sigma^2_{AA} + \dots$$

Equation 13 Formula for estimating covariance of response units and selection units,

where:
terms are as defined previously.

Note that if Equation 13 is used, a slightly biased estimate of heritability will result even if the best estimates of variance components are obtained. This is due to epistatic variance. For other types of progeny, the bias in the numerator can be much larger. Let us look at estimates based on Equation 13 for some example cases/progeny types.

Example A

Estimation of Narrow sense heritability from a half-sib family experiment with data obtained on **individual plants** in **one** environment.

1. Heritability on an individual plant basis

    - Selection among individual plants
    - 1 Replication in 1 environment
    - Response is measured in outbred progeny

Example A

$$h_1^2 = \frac{(\frac{4}{1+F_P})\hat{\sigma}_F^2}{\hat{\sigma}_p^2} = \frac{\hat{\sigma}_A^2 + (\frac{1+F_P}{4})\hat{\sigma}_{AA}^2}{\hat{\sigma}_p^2}$$

$$\hat{\sigma}_p^2 = \hat{\sigma}_F^2 + \hat{\sigma}_{FE}^2 + \hat{\sigma}_\epsilon^2 + \hat{\sigma}_\omega^2$$

$$Bias = \frac{\frac{1}{4}(F_p - 1)\sigma_{AA}^2}{\sigma_p^2}$$

$$\sigma_F^2 = \frac{(1+F_p)}{4}\sigma_A^2 + \frac{(1+F_P)^2}{16}\sigma_{AA}^2$$

where: $\hat{\sigma}_{FE}^2$ is the estimated family by environment interaction variance; $\hat{\sigma}_\epsilon^2$ is the estimated error variance, and $\hat{\sigma}_\omega^2$ is the estimated within family variance.

Example B

Estimation of Narrow sense heritability from a half-sib family experiment with data obtained on **individual plants** in **multiple** independent environments.

2. Family heritability on a plot basis (half-sib family, single plot mean values)

    - Selection among plot means
    - 1 Replication in 1 environment
    - Response is measured in outbred progeny

Example B

$$h_1^2 = \frac{\hat{\sigma}_F^2}{\hat{\sigma}_p^2} = \frac{\frac{1}{4}(1+F_p)\hat{\sigma}_A^2 + \frac{1}{16}(1+F_p)^2\,\hat{\sigma}_{AA}^2}{\hat{\sigma}_p^2}$$

$$\hat{\sigma}_p^2 = \hat{\sigma}_F^2 + \hat{\sigma}_{FE}^2 + \hat{\sigma}_\epsilon^2 + \frac{\hat{\sigma}_\omega^2}{n}$$

$$Bias = \frac{\frac{1}{32}(1+F_p)^2\hat{\sigma}_{AA}^2}{\sigma_p^2}\sigma_p^2$$

$$\hat{\sigma}_F^2 = \frac{(1+F_p)}{4}\sigma_A^2 + \frac{(1+F_p)^2}{16}\sigma_{AA}^2$$

where: n is the number of entries or plots; all other terms are described in example A.

Computational Considerations

Example C

Estimation: Narrow sense heritability estimated from a half-sib family experiment with data obtained on **individual plants** in **multiple** independent environments.

3. Family heritability

    - selection among half-sib family mean averaged over environments
    - outbred progeny

$$Example\ C$$

$$h_1^2 = \frac{\hat{\sigma}_F^2}{\hat{\sigma}_p^2} = \frac{\frac{1}{4}(1+F_p)\hat{\sigma}_A^2 + \frac{1}{16}(1+F_p)^2\hat{\sigma}_{AA}^2}{\hat{\sigma}_p^2}$$

$$\hat{\sigma}_p^2 = \hat{\sigma}_F^2 + \frac{\hat{\sigma}_{FE}^2}{e} + \frac{\hat{\sigma}_\epsilon^2}{er} + \frac{\hat{\sigma}_\omega^2}{ern}$$

$$Bias = \frac{\frac{1}{32}(1+F_P)^2\hat{\sigma}_{AA}^2}{\sigma_p^2}$$

$$\hat{\sigma}_F^2 = \frac{(1+F_p)}{4}\sigma_A^2 + \frac{(1+F_p)^2}{16}\sigma_{AA}^2$$

The only way to remove the bias is to include both selection units and response units in the analyses. This **is not the same thing as including both groups in the same sets of environments**.

Method of Moments

Next, let us explore the computational nuances of these concepts in the context of plant breeding populations. Consider first the evaluation of half-sibs from a random mating population in a replicated Multi-Environment Trial. Let the phenotypic variance of the selection units be designated σ_p^2. From an introductory course in statistics, we were taught that the phenotypic variance on an entry means basis can be obtained directly from Ordinary Least Squares (OLS) ANOVA by equating the estimated Mean Squares (MS) with Expected Mean Squares (EMS). This is also known as the Method of Moments (MoM). Thus, an estimate of phenotypic variance, σ_p^2 represented as in Equation 14:

$$\frac{MS_f}{er} = \sigma^2_f + \frac{\sigma^2_{fE}}{e} + \frac{\sigma^2_e}{er}$$

Equation 14 Formula for estimating phenotypic variance,

where:

MS_f = the mean square, i.e., family variance,

other terms are as defined previously.

When to Use Method of Moments

It turns out that the application of MoM is appropriate only if the data are from a balanced experiment, i.e., the number of genotypes, in this case, families or genotypic entries, is the same across reps and environments. Recall that *lsmeans* are useful for estimates of entry means in the case of unequal replication per environment. Next, we need to learn how to obtain estimates of the variance components for unbalanced data sets.

The most obvious problem is that the coefficients of the variance components are not equal to the products of the numbers of reps and environments in the EMS. Addressing this problem is fairly straightforward (Milliken and Johnson, 1992). A more difficult problem is that the estimates of the variance components themselves are no longer the "best" estimates. The solution, as described by Holland et al (2003) is to obtain Restricted Expected Maximum Likelihood (REML) estimates in a Mixed Model Procedure (MMP).

REML

For example, let us consider the case of half-sib progeny. Recall Equation 12:

$$\sigma^2_f = Cov(HS) = \sigma_{HS} = \frac{(1+F)}{4}\sigma^2_A + [\frac{(1+F)}{4}]^2\sigma^2_{AA} + \ldots$$

If the data are not balanced, the variance component will not be estimated correctly unless REML is used. Let us assume that we obtain a 'best' estimate for σ_{HS}; either because our data are balanced or we have used REML. Should we use Equation 8 for the Cov(R,S)? To answer this, we have to recognize that there is a genetic relationship between selection units and response units, i.e., there is a pedigree relationship or coefficient of coancestry between the selection and response units, and Equation 12 does not take this into consideration. In the case where both selection units and response units are half sibs, the Cov(R,S) is represented as in Equation 13:

$$Cov(R,S) = \frac{(1+F)}{4}\sigma^2_A + \frac{1}{32}[1+F]^2\sigma^2_{AA} + \ldots$$

Thus, if Equation 13 is used, a slightly biased estimate of heritability will result even if REML-based estimates of variance components are obtained. For other types of progeny, the bias in the numerator can be much larger. Thus, the predicted genetic gain that might be used for planning purposes or comparison of possible breeding methods will be overestimated.

References

Fehr, W. R. 1993. Principles of Cultivar Development. Vol. 1. Theory and Techniques. Macmillian Publishing Company.

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. Plant Breed. Rev. 2003:9–112.

Lush, J. L. 1940. Intra-sire correlations or regressions of offspring on dam as a method of estimating heritability of characteristics. Am. Soc. Anim. Prod. Proc. 33: 293-301.

Milliken, G. A., and D. E. Johnson. 1992 Analysis of Messy Data: Vol I, Design Experiments, Chapman & Hall/CRC, London.

Nyquist, W. E. 1991. Estimation of heritability and prediction of selection response in plant populations. Crit. Rev. Plant Sci. 10:235–322.

Sprague, G. F., and S. A. Eberhart. 1977. Corn breeding. p. 305-362. In. G. F. Sprague and J. W. Dufley (ed) Corn and corn improvement Agron. Monogr. 18. ASA, CSSA, and SSSA, Madison, WI.

How to cite this chapter: Beavis, W., K. Lamkey and A. A. Mahama. 2023. Selection Response. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics in Plant Breeding*. Iowa State University Digital Press.

Chapter 10: G x E

William Beavis; Kendall Lamkey; Katherine Espinosa; and Anthony Assibi Mahama

One of the most difficult aspects of plant breeding involves making decisions about environments to target for the development of new cultivars. This challenge is not especially difficult if cultivars are adapted to large geographic regions with little variability among environments or if there is significant variability among environments, but potential cultivars respond similarly to the environmental differences. However, if potential cultivars do not respond similarly to environmental differences within a targeted set of environments, i.e., there are genotype-by-environment interactions, decisions on which genotypes to develop can be difficult. Herein, we explore the impacts of environments on genotypes in cultivar development programs.

Learning Objectives

- Conceptual types of GxE interactions.
- Decompose GxE interactions into heterogeneous variability and inconsistent rankings.
- Leverage information on heterogeneous variance and inconsistent rankings to meet breeding objectives.

Environmental Components of Variance

Recall that our working model for the phenotype includes genotypic and non-genotypic (environmental) sources of variability (Equation 1):

$$P = \mu + G + E$$

Equation 1 Linear model for sources of variability in phenotype.

where:
P = phenoype,
μ = overall mean,
G = genotype effects ,
E = non-genetic of environment effects.

Briefly, we consider the components of E as shown in Fig. 1.

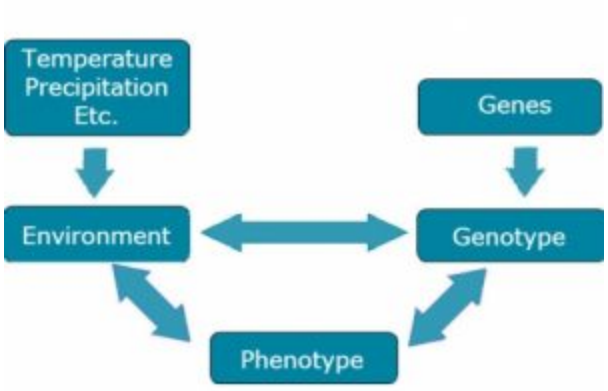


Fig. 1 Illustration of relationships among environments, genotypes, and phenotypes.

Many Meanings of Environment

Micro-environmental effects: the environment of a single organism as opposed to that of another growing at the same time and in almost the same place.

- All things except genotype affect a plant's development. Note that the probability that two plants in the same field will experience the same environment is infinitesimally small.
  - Physical and chemical properties of the soil
  - Climatic variables (rain, temperature, etc)
  - Solar radiation
  - Biotic stresses

Macro-environmental effects: the general environment associated with a field site over a period of time.

- Different class of environments in one area or time than another
- A collection of micro-environments

Environmental Sources of Variation

Environmental sources of variation can also be hierarchically modeled to consist of variability among environments and within environments (Equation 2):

$$\sigma_E^2 = \sigma_{among}^2 + \sigma_{within}^2$$

Equation 2 Formula for among and within sources of variability.

where:
σ^2_E = environmental variance,
σ^2_{among} = among environment variance,
σ^2_{within} = within environment variance.

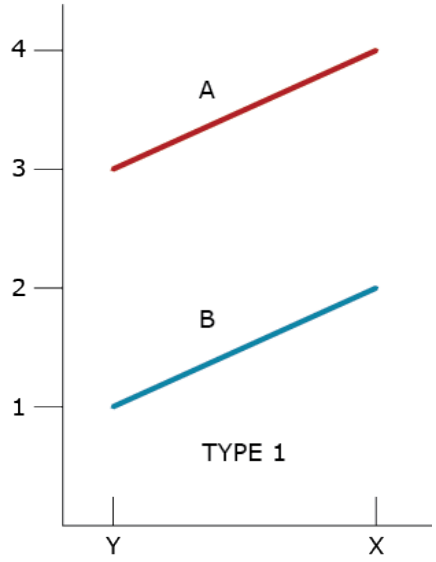


Fig. 2 No interaction type phenotype response in two environments.

$$P_{ij} = \mu + G_i + E_j$$

Equation 3 Model for phenotypic response.

where:
P_{ij} = phenotype,
μ = overall mean,
G_i = genotype effect,
E_j = environment effect.

From the beginning of field assessments of clonally propagated plants, we have been able to recognize variability within and among locations (environments). As soon as we could evaluate more than one replicated genotype, we also began to recognize patterns of phenotypic responses to environments. In 1964, Allard and Bradshaw provided a simple classification scheme of the types of phenotypic responses that could happen using two genotypes (designated A and B) and two environments (designated X and Y) and modeled as in Equation 3. The first type of response (Type 1) reveals that there is a difference of 2 units between the genotypes and a difference of 1

unit between the environments (Fig. 2). They also recognized a second type of response (Type 2) in which the difference between genotypes was one unit while the difference between environments was 2 units. Both types of responses indicate no interaction.

Simple Types of GxE Interactions

These types of interaction can all be modeled as (Equation 4):

$$P_{ij} = \mu + G_i + E_j + GE_{ij}$$

Equation 4 Model for phenotype response with GxE present.

where:
P_{ij} = phenotypic response,
μ = overall mean,
G_i = genotype effects,
E_j = environment effects,
GE_{ij} = genotype by environment interaction effect.

Type 3 GxE Interaction

Allard and Bradshaw recognized that there could be a lack of consistent responses by two different genotypes in two environments. The lack of consistent responses by genotypes to different environments had been recognized as genotype by environment interactions since the beginning of replicated trials, and Allard and Bradshaw classified these into four types of GxE for two genotypes in two environments.

A type 3 GxE response (Fig. 3) was based on the heterogeneity of genotypic variability between environments. Assuming that larger phenotypic values are desired, in GxE types 1,2, and 3, genotype A is better adapted to both types of environments.

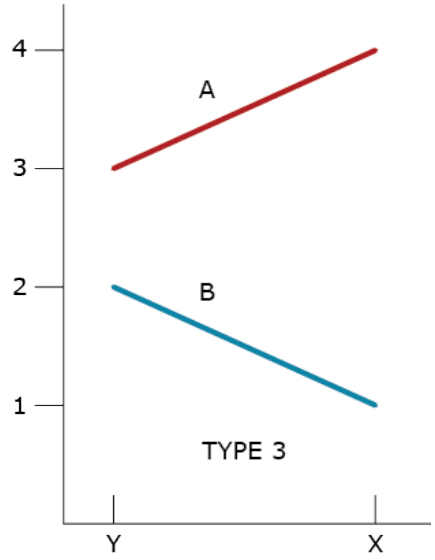


Fig. 3 Genotype response with interaction in two environments.

Type 4 GxE Interaction

A type 4 GxE interaction is due to a combination of heterogeneous genotypic variability and a failure of the genotypes to have correlated responses (change of rank) across the environments (Fig. 4).

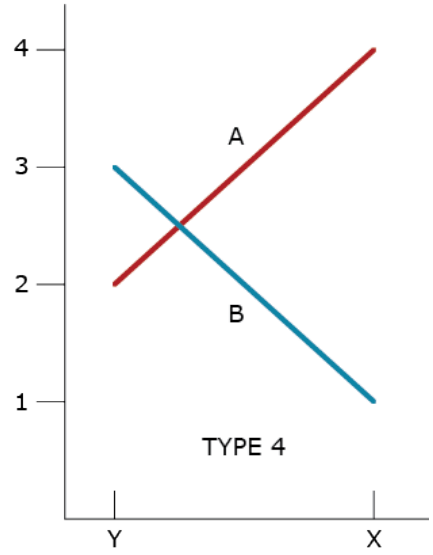


Fig. 4 Genotype response with interaction (lines intersect) in two environments.

Type 5 GxE Interaction

A type 5 GxE is due to a failure of the genotypes to have correlated responses across the environments, while the genotypic variability is homogeneous between the environments (Fig. 5). If environments X and Y represent typical types of environments in a market, then there are unique best genotypes for each type of environment; neither is broadly adapted to both environments.

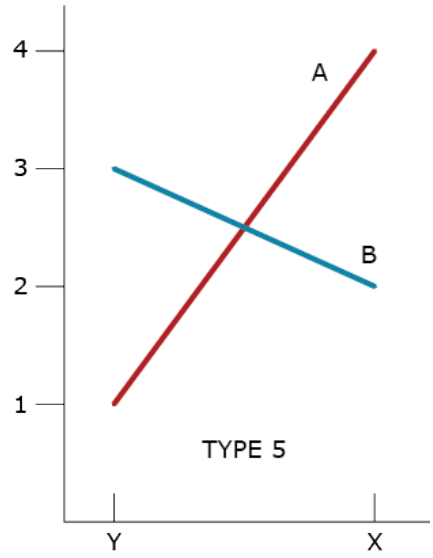


Fig. 5 Genotype response with interaction (lines intersect) in two environments.

Type 6 GxE Interaction

A type 6 GxE interaction illustrates a 'racehorse' response by Genotype A. It takes full advantage of favorable environment X while failing under the stressful environment Y. This type of GxE also illustrates a more 'stable' response by Genotype B. The question for the plant breeder is whether to develop both types of cultivars or just one.

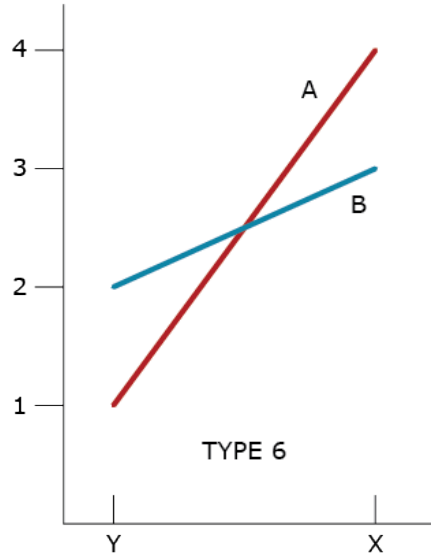


Fig. 6 Genotype response with interaction (lines intersect) in two environments.

Complex Types of GxE Interactions

As the number of environments and genotypes increases, the ability to sort out the types of GxE interactions becomes more difficult. For example, consider the data and plot in Fig. 7. It is likely that all types of GxE interactions are present in these data. If all types of GxE are present, are there prevalent types of GxE? Do the genotypes behave the same way in some pairs of environments? What is the nature of GxE between all pairs of environments (1:2, 1:3, 2:3 …)?

To answer these questions, multi-variate **statistical techniques**, sometimes referred to as **pattern analyses**, are used to discover and summarize consistent patterns in large data sets.

Name	Env1	Env2	Env3	Env1-Env2	Env1-Env3
45	55.4	71.31	46.31	-15.91	9.09
55	49.8	90.66	65.66	-40.86	-15.86
60	55.8	82.49	57.49	-26.69	-1.69
141	40	69.11	44.11	-29.11	-4.11
215	29.7	79.43	54.43	-49.73	-24.73
218	31	83.42	58.42	-52.42	-27.42
314	51.2	91.24	66.24	-40.04	-15.04
390	31.1	80.69	55.69	-49.59	-24.59
391	27.4	79.43	54.43	-52.03	-27.03
399	27.9	73.22	48.22	-45.32	-20.32
414	29	70.92	45.92	-41.92	-16.92
415	39.8	68.01	43.01	-28.21	-3.21
416	33.3	64.49	39.49	-31.19	-6.19
417	33.4	72.22	47.22	-38.82	-13.82
422	35.7	76.8	51.8	-41.1	-16.1
429	41.4	82.8	57.8	-41.4	-16.4
440	54.8	81.38	56.38	-26.58	-1.58
449	50.2	85.1	60.1	-34.9	-9.9
477	57.2	73.97	48.97	-16.77	8.23
486	43.2	72.24	47.24	-29.04	-4.04
497	30.5	76.21	51.21	-45.71	-20.71
501	30.2	90.47	65.47	-60.27	-35.27
511	30.9	88.72	63.72	-57.82	-32.82
537	50.7	82.72	57.72	-32.02	-7.02
628	47.1	65.3	40.3	-18.2	6.8
650	74	71.13	46.13	2.87	27.87
687	41	85.61	60.61	-44.61	-19.61
690	37.9	97.34	72.34	-59.44	-34.44
715	56.5	78.96	53.96	-22.46	2.54
771	43.7	76.11	51.11	-32.41	-7.41
779	59.3	78.17	53.17	-18.87	6.13
792	50.9	79.76	54.76	-28.86	-3.86
803	56.2	79.3	54.3	-23.1	1.9
804	51.9	82.52	57.52	-30.62	-5.62
828	32.2	78.55	53.55	-46.35	-21.35
967	54.6	83.29	58.29	-28.69	-3.69

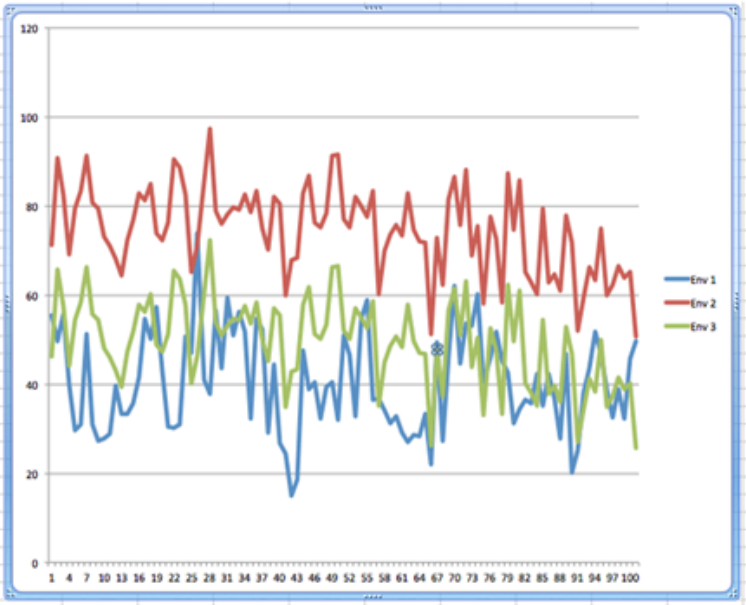


Fig. 7 Sample of yield values from ~100 entries taken at three environments from a 6500-entry trial grown at 20 environments.

Pattern Analysis Methods

A partial listing of these pattern analysis methods would include Cluster Analyses (CA), Principal Component Analysis (PCA), Additive Main and Multiplicative Interaction (AMMI) models, Sites Regression models, Partial Least Squares, Factorial Regression, Linear Bilinear Mixed Models, Generalized Linear Models, Support Vector Machines, Bayesian Networks, Reproducing Kernel Hilbert Spaces, etc.

The development of these methods was motivated by the need to find patterns in physical and chemical spectra 25 to 50 years ago. These methods began to be applied by ecologists in the 1970s, GxE challenges in plant breeding during the 1990s, and to find patterns in 'omics' data during the 2000s. The application and interpretation of the methods in GxE continue to be an active area of research and well beyond the scope of an introduction to GxE.

Cluster Analyses

Herein we introduce an application of multi-variate techniques to find patterns in GxE interactions using Cluster Analyses (CA). The purpose of applying CA is to either: **1.** organize the environments into homogeneous groups of environments so that there are no GxE interactions within environments and to emphasize (maximize) the differences between homogeneous groups

of environments or **2.** organize the genotypes into groups with no GxE within the groups and maximize our ability to identify genotypes that have different responses to environments.

Cluster analyses require a metric that quantifies dissimilarities among all possible pairs of units that we wish to cluster. There are many possible distance metrics that could be used. The most commonly used metric for CA of environments, based on GxE, is the Euclidean distance which is based on the Pythagoras' theorem (Fig. 8).

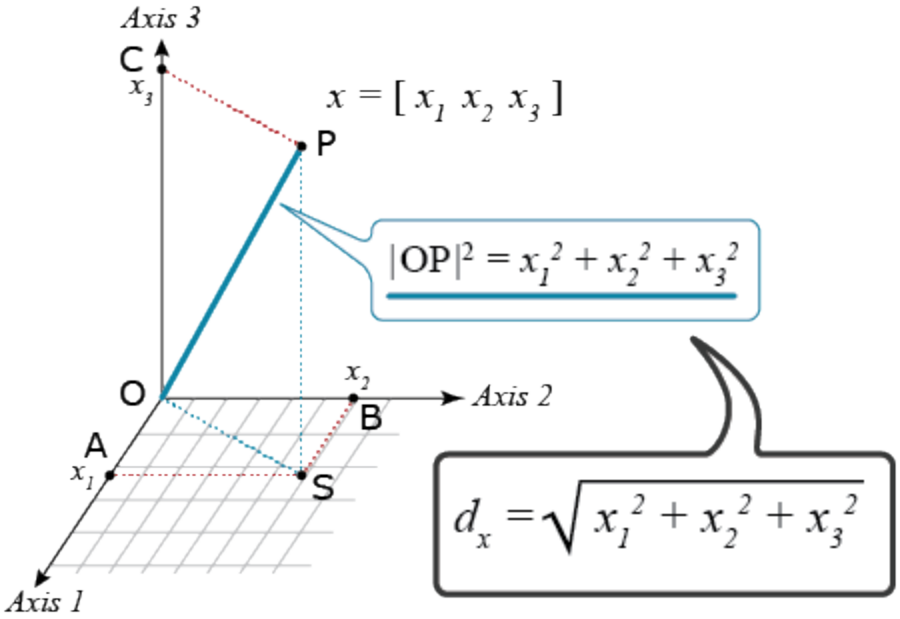


Fig. 8 Pythagoras' theorem extended into three-dimensional space.

Euclidean Distance

For a trait such as yield, it is advisable to first standardize the values so that all of the yield values are on the same scale: Calculate the average and standard deviation for each location, subtract the average value from the genotypic value, and divide by the standard deviation for each of the genotypes by location. Next, calculate the Euclidean Distance of the standardized yield values between all possible pairs of environments.

CA also requires that we choose a grouping algorithm. There are dozens of clustering algorithms, and none should be considered 'best' because there are no objective criteria that can be applied to all data sets. For purposes of interpretation using yield data with evidence of GxE, I prefer to use an agglomerative hierarchical clustering technique such as Ward's (also known as **incremental sums of squares**) or the Unweighted Pair Group Method with Arithmetic mean (UPGMA, also known as average linkage clustering). Cooper and DeLacy (1994) prefer Ward's, but for the novice,

it is usually good to try several grouping algorithms for purposes of comparison based on the goals of the breeding project.

Variation Flux

One of the fundamental questions that a breeding project needs to decide is whether to develop broadly adapted cultivars or specific cultivars for specific environments. Often this is determined by production and marketing considerations, but there is also an issue of identifying the types of environments that the crop will encounter within a marketing region. In order to assess the types of environments, the breeder needs to sample the total population of macro-environments using a sample of genotypes. There will clearly need to be trade-offs between these two sampling objectives. Decisions on the trade-offs could actually bias the results that one obtains because genotypic variances can be confounded with GxE variances and vice-versa.

To illustrate, consider Fig. 9, where A represents a population of macro-environments and S is a subset of macro-environments.

Let A serve as our reference population of environments.

It can be shown (with a little algebra) that $\sigma^2_{GA} = \sigma^2_G + \sigma^2_{GS}$

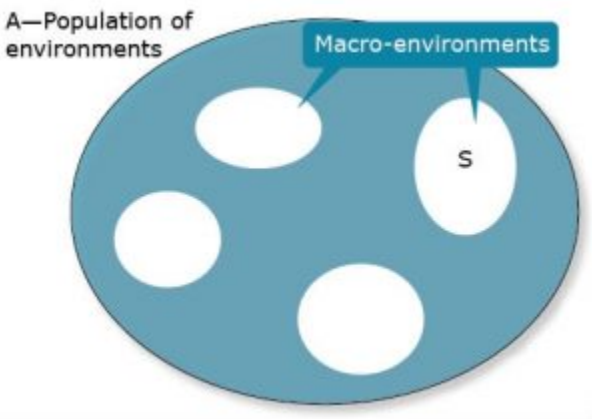


Fig. 9 Illustration of targeted sets of environments.

A consequence is that if the subset population of environments, S, is made more homogeneous (smaller subsets of the total), then genotypic variance will increase because GS interaction variance will decrease. Alternatively, expansion of the targeted subset S of environments will result in a more heterogenous subset which will, in general, increase GS interaction variance at the expense of genetic variance. The challenge is to subdivide an original set of environments

so that subdivisions are clearly delineated and substantially more homogeneous. If the market analysis then reveals that multiple sub-environments should be served, it will require an increase in the breeding effort since one breeding program needs to be replaced by multiple breeding projects.

Partition of GxE Variances

GxE variances can be partitioned into two components:

- due to heterogeneity of genotype variance among environments, and
- due to lack of correlation of genetic performance among environments.

Muir et al (1992), provided the means for calculating these two components.

Muir's Partition Method

Given a model for the phenotype (Equation 5):

$$Y_{ijk} = \mu + G_i + E_j + GE_{ij} + \varepsilon_{(ij)k}, \quad i = 1, \ldots, g; \;\; j = 1, \ldots, e; \;\; k = 1, \ldots n$$

Equation 5 Model for calculating components of phenotype.

where:
Y_{ijk} = phenotype,
μ = overall mean,
G_i = genotype effect,
E_j = environment effect,
GE_{ij} = genotype by environment interaction effect,
$\varepsilon_{(ij)k}$ = residual (error).

Then SS(GE) is determined as (Equation 6):

$$SS(GE) = n \sum_{i}^{g} \sum_{j}^{e} (\bar{Y}_{ij.} = \bar{Y}_{i..} = \bar{Y}_{.j.} + \bar{Y}_{...})^2, \quad = \frac{n \sum_{i \neq i'}^{g} (S_i^2 + S_{i'}^2 - 2S_{ii'})}{2g};$$

$$\text{where,} \quad S_i^2 = \sum_{j}^{e} (\bar{Y}_{ij.} - \bar{Y}_{i..})^2, \quad S_{ii'} = \sum_{j}^{e} (\bar{Y}_{ij.} - \bar{Y}_{i..})(\bar{Y}_{i'j.} - \bar{Y}_{i'..})$$

Equation 6 Model for calculating sum of squares GxE, SS(GE).

where:
$\bar{Y}_{ij.}$ = mean of ij's for all plots,
$\bar{Y}_{i..}$ = mean of i's for all jk's,
$\bar{Y}_{.j.}$ = mean of j's for all ik's,
$\bar{Y}_{...}$ = grand mean,
S_i^2 = variance of i,
$S_{i'}^2$ = variance of i',
$S_{ii'}$ = variance of ii',
g = all genotypes.

GE Interaction Equation

The GE interaction can then be expressed as:

$$\frac{\sum_{i<i'}^{g}\left[(S_i - S_{i'})^2 + 2(1 - r_{ii'})S_i S_{i'}\right]}{g}$$

These results then allow the GE interaction sums of squares to be partitioned into a term due to heterogeneous variance, $SS(HV)_{ii'}$, and that due to imperfect positive correlation of the pair, $SS(IC)_{ii'}$ (Equation 7).

$$SS(HV)_{ii'} = \frac{n(S_i - S_{i'})^2}{g}, \quad SS(IC)_{ii'} = \frac{2n(1 - r_{ii'})S_i S_{i'}}{g}$$

Equation 7 Model for calculating sum of squares GxE, SS(GE).

where:
S_i = variance of i,
$S_{i'}$ = variance of i',
$r_{ii'}$ = correlation between i and i',
g = all genotypes.

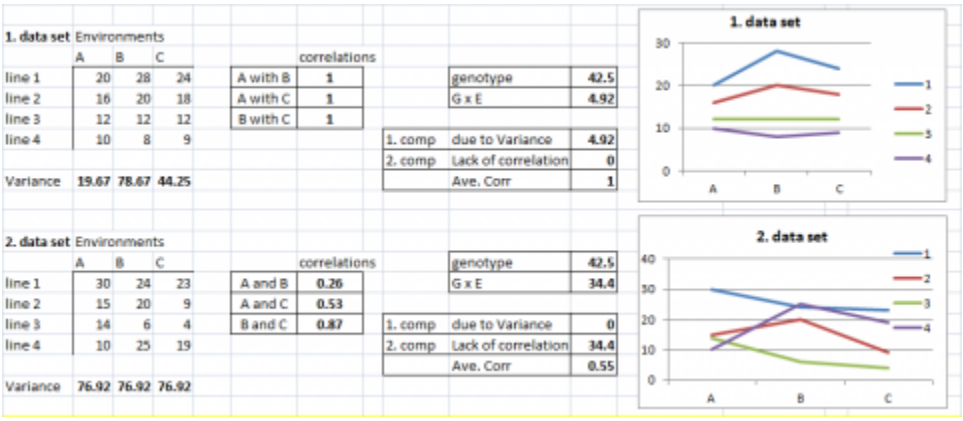
Interaction Components

While the first component can be derived from ANOVA results of single environments using the

genotypic variance in each environment and their average, the second component can then be calculated using the estimated variance component (EMS) for G x E over all environments and genetic variance component over all environments.

Example Data Sets 1 and 2

To visualize the relative influence of heterogeneity of genotypic responses and lack of correlations among genotypes, consider the following example data sets.



1. data set	Environments		
	A	B	C
line 1	20	28	24
line 2	16	20	18
line 3	12	12	12
line 4	10	8	9
Variance	19.67	78.67	44.25

	correlations
A with B	1
A with C	1
B with C	1

	genotype	42.5
	G x E	4.92
1. comp	due to Variance	4.92
2. comp	Lack of correlation	0
	Ave. Corr	1

2. data set	Environments		
	A	B	C
line 1	30	24	23
line 2	15	20	9
line 3	14	6	4
line 4	10	25	19
Variance	76.92	76.92	76.92

	correlations
A and B	0.26
A and C	0.53
B and C	0.87

	genotype	42.5
	G x E	34.4
1. comp	due to Variance	0
2. comp	Lack of correlation	34.4
	Ave. Corr	0.55

Fig. 10 GxE due to heterogeneity of environments (data set 1) and lack of correlation among lines (data set 2).

In the first two examples, data sets, G x E is fully explained by one of its two components. In the first case, all G x E is due to the heterogeneity of genotypic variance among environments. In the second case, G x E is completely due to a lack of correlation of genotypic performance among environments.

Example Data Set 3

A more realistic example can be found in data set 3, where the GxE is due to a mixture of both components. It is, however, worthwhile to look at their partial contribution; genotypic heterogeneity explains only 18% of GxE interaction, and lack of correlation (consistent ranking) explains 82%.

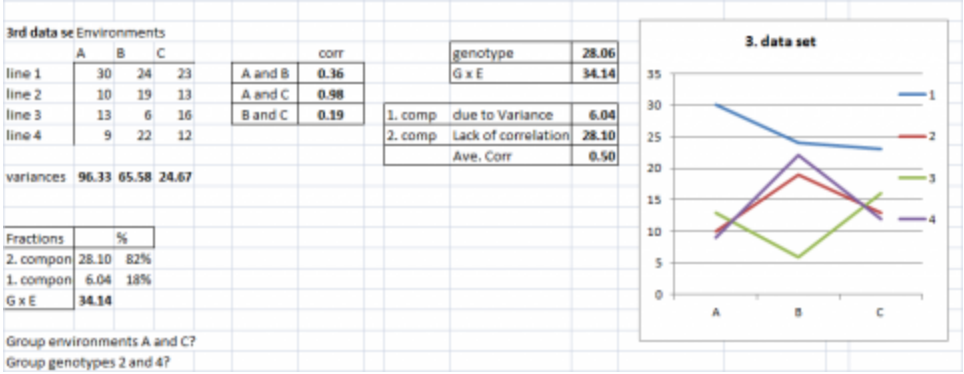


Fig. 11 GxE due to both heterogeneity of environments and lack of correlation among lines.

In the early stages of a breeding project, thousands of genotypes might be evaluated in two or three environments, or a few hundred genotypes might be evaluated in dozens of environments. In such situations, the simple graphical representations (Figs. 2-6) and partitioning of GxE variances (Equations 1-5) become very difficult to interpret.

Alternative Analyses

There are alternative analyses and visualization techniques that are used to interpret data from large numbers of genotypes grown in large numbers of environments. For example, pattern analyses employ measures for similarity or dissimilarity to group environments and lines for interpretable graphical representations of either genotypic or environmental performance. Consider the following as illustrated examples based on the data represented in Fig. 11.

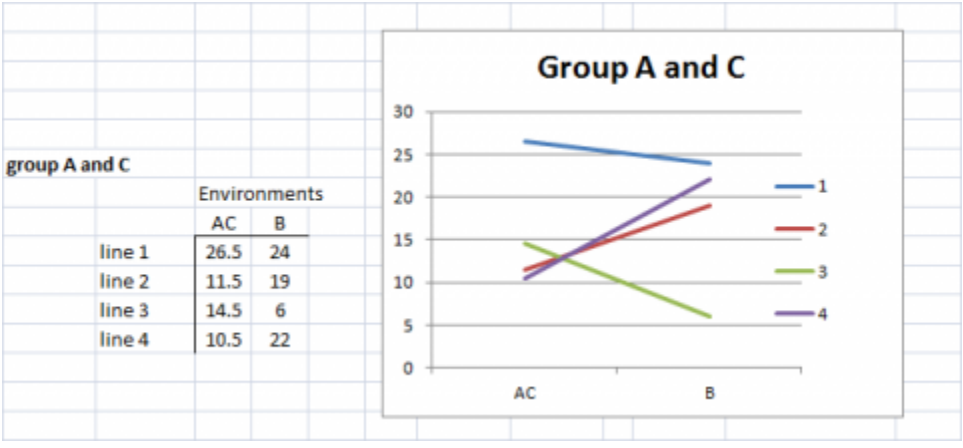


Fig. 12 Grouping of environments based on similar genotypic responses.

In Fig. 12, we have clustered Environments A and C because the genotypes respond to these

environments almost identically and very differently from the manner in which they respond to environment B.

Grouping Similar Responses

In Fig. 12, we also recognize that Genotypes (lines) 2 and 4 respond to the environments in a similar manner, so we cluster these together and represent the response patterns of genotypes as 3 distinct patterns.

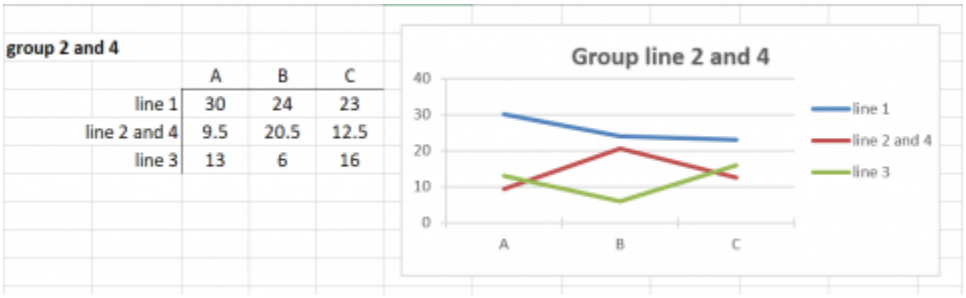


Fig. 13 Grouping of lines with similar responses in 3 environments.

Grouping of Lines and Environments

In Figures 12 and 13, either lines or environments are grouped for similar performance. In Fig. 14, both groupings are shown in the same graph. Simple means of groups were taken to give an example for simplification of G x E interactions.

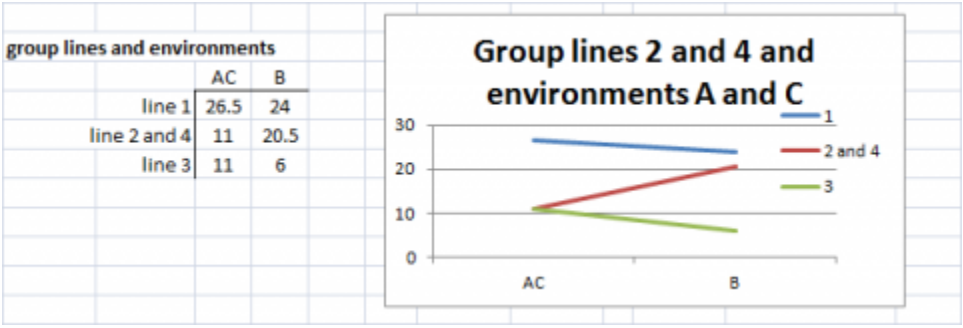


Fig. 14 Grouping of lines and environments.

For more complex data sets, measures for similarity and dissimilarity of the performance of genotypes can be used to summarize differences in genetic performance of the genotypes in environments j and j'. We can denote such difference measures as $D_{g(jj\text{-})}$. We can also consider a measure of a difference, designated as $D_{e(ii'\text{-})}$, between environments j and j_m, the way in which they discriminate between the genetic performance of genotypes. DeLacy and Cooper (1990) and

DeLacy et al. (1990a) discussed alternative forms of $D_{e(jj')}$, which have been used for pattern analysis of relationships among environments in METs (Cooper and Delacey 1994).

Flux between Genotypic Variance and GE Interaction Variance

One of the fundamental questions that a breeding project needs to decide is whether to develop broadly adapted cultivars or specific cultivars for specific environments. Often this is determined by production and marketing considerations, but there is also an issue of identifying the types of environments that the crop will encounter within a marketing region. In order to assess the types of environments, the breeder needs to sample the total population of macro-environments using a broad sample of genotypes. There will clearly need to be trade-offs between these two sampling objectives. Decisions on the trade-offs could actually bias the results that one obtains because genotypic variances can be confounded with GxE variances and vice-versa.

To illustrate, consider Fig. 9 above, where A represents a population of macro-environments and S is a subset of macro-environments.

Let A serve as our reference population of environments.

It can be shown (with a little algebra) that $\sigma^2_{GA} = \sigma^2_G + \sigma^2_{GS}$.

A consequence is that if the subset population of environments, S, is made more homogeneous (a smaller subset of the total), then genotypic variance will increase because GS interaction variance will decrease. Alternatively, expansion of the targeted subset S of environments will result in a more heterogeneous subset which will, in general, increase GS interaction variance at the expense of genetic variance. The challenge is to subdivide an original set of environments so that subdivisions are clearly delineated and substantially more homogeneous. If the market analysis then reveals that multiple sub-environments should be served, it will require an increase in the breeding effort since one breeding program needs to be replaced by multiple breeding projects.

Impact of Multiple Environments

From an introductory course in statistics, we were taught that the phenotypic variance on an entry means basis can be obtained directly from Ordinary Least Squares (OLS) ANOVA by equating the estimated Mean Squares (MS) with Expected Mean Squares (EMS). This is also known as the Method of Moments (MM see Review Module on Statistical Inference). Thus, an estimate of phenotypic variance, S^2_p, on an entry mean basis is equal to

$$\frac{MS_g}{er} = \sigma_g^2 + \frac{\sigma_{ge}^2}{e} + \frac{\sigma_\varepsilon^2}{er}$$

Equation 8 Formula for estimating phenotypic variance on entry mean basis.

where:
MS_g = mean square value of genotype (g) in the ANOVA table,
σ_g^2 = genotypic variance,
σ_{ge}^2 = genotype by environment interaction variance,
σ_ε^2 = error variance of the error term, ε,
r = number of replications.

The ability to replicate genotypes and grow them within plots that can be replicated enables the plant breeder to "adjust" their precision around their estimates.

Variance Component Estimation Example

Let us consider a typical plant breeding field trial in which location and year combinations are considered unique environments. (Table 1) Let there be *e* combinations of years and locations. Also, assume the genotypes, *g*, are grown at the same locations each year.

Table 1 Expected means squares for estimating variance components.

Source	d.f.	Mean Square	EMS
Environments (E)	$e-1$	n/a	n/a
Reps within E (R)	$e(r-1)$	n/a	n/a
Genotypes (G)	$g-1$	**M1**	$\sigma_e^2 + r\sigma_{\bar{g}e}^2 + re\sigma_{\bar{g}}^2$
G x E	$(g-1)(e-1)$	**M2**	$\sigma_e^2 + r\sigma_{\bar{g}e}^2 \sigma_e^2 + r\sigma_{\bar{g}e}^2$
Residual	$e(g-1)(r-1)$	**M5**	σ_e^2

Estimators

Employing the MM approach, we can obtain estimates of the variance components and, thus, an estimate of the Covariance of the genotypic units (Table 2).

Table 2 Estimating variances.

Function	Variance Estimated
$F_2 = \frac{(M1 - M5)}{r}$	$\sigma^2_{\overline{ge}}$
$F_1 = \frac{(M1 - M2)}{re}$	$\sigma^2_{\overline{g}}$

Application Notes

Application of the MM is appropriate only if the data are from a balanced experiment, i.e., the number of genotypic units is the same across reps and environments. In the review chapter, we employed lsmeans to obtain adjusted estimates of entry means in the case of unequal replication per environment. However, we did not learn how to obtain estimates of the variance components for unbalanced data sets.

If there are only a few missing values (say < 5%) from some replicates, then the impact on the estimates of variance components will not be very great. However, we often design experiments to take advantage of seed supplies which may vary greatly among our genotypic units. In such cases, the coefficients of the variance components are not equal to the products of the numbers of reps and environments represented in the EMS. Addressing this issue is fairly straightforward (Milliken and Johnson, 1992). A more difficult problem is that the estimates of the variance components themselves are no longer the “best” estimates. The solution, as described by Holland et al (2003) is to obtain Restricted Maximum Likelihood (REML) estimates in a Mixed Model Procedure (MMP).

References

Cooper, M., and J.H. Delacy. 1994. Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theoretical and Applied Genetics 88: 561–572.

DeLacy, I. H., and M. Cooper. 1990. Pattern analysis for the analysis of regional variety trials. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding. Louisiana State University, Baton Rouge, Louisiana, pp. 301–334.

DeLacy I. H., M. Cooper, and P. Lawrence. 1990a. Pattern analysis over years of regional variety

trials: relationship among sites. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding. Louisiana State University, Baton Rouge, Louisiana, pp 189–213.

Holland, J.B., W.E. Nyquist, and C.T. Cervantes-Martínez. 2003. Estimating and interpreting heritability for plant breeding: An update. Plant Breed. Rev. 2003:9–112.

Milliken, G. A., and D. E. Johnson. 1992 Analysis of Messy Data: Vol I, Design Experiments, Chapman & Hall/CRC, London.

How to cite this chapter: Beavis, W., K. Lamkey, K. Espinosa, and A. A. Mahama. 2023. G x E. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 11: Multiple Trait Selection

William Beavis; Kendall Lamkey; and Anthony Assibi Mahama

Most cultivar improvement programs involve selection for multiple traits. Depending on the program or project goals, one of three types of multiple-trait selection can be employed. A brief description of these types follows.

- **Multistage selection:** Selection for different traits at different stages during cultivar development.
- **Tandem selection:** Selection for one trait until that trait is improved, then for a second, etc., until finally each has been improved to the desired level.
- **Independent culling levels:** A certain level of merit is established for each trait, and all individuals below that level are discarded regardless of values for other traits.
- **Index selection:** Select for all traits simultaneously by using some index of net merit.

Learning Objectives

- Explain the role of selection for multiple traits on genetic gain.
- Learn methods for selecting multiple traits.
- Learn how to evaluate the efficiency and effectiveness of selection on multiple traits.

Index Selection

An **index** is the best linear prediction of an individual's or line's breeding value and takes the form of the multiple regression of breeding value on all sources of information.

The objective of an index is to find the linear combination of phenotypic values that maximizes the expected gain or, equivalently, that maximizes the correlation between the index value and true worth (breeding value).

Index Selection Theory

Define phenotype as affected by components in Equation 1.

$$P_i = G_i + E_i (i = 1, \ldots, n)$$

Equation 1 Linear model for sources of variability in phenotype.

where:
P_i = the observed value of the attribute i for an individual or line,
G_i = the average of phenotypic values over a population of environments,
E_i = non-genotypic contributions from environments.

Note: Genotype x environment interactions are permitted, assuming genotypes and environments are associated entirely at random; such interactions are incorporated into $\boldsymbol{E_i}$ (Equation 1). If GxE are not random, then see Cooper and DeLacy (1994).

Assume $\boldsymbol{G_i}$ is composed entirely of additive effects of genes (breeding values). Define the genotypic economic value, H, of an individual as written in Equation 2:

$$H = \sum_{i=1}^{n} a_i G_i$$

Equation 2 Linear model for sources of variability in phenotype.

where:
a_i = known relative economic values.

Assume the quantities $\boldsymbol{P_i}$ and H are such that the regression of $\boldsymbol{P_i}$ on H is linear. Selection will then be based on the linear function, I (Equation 3).

$$I = \sum_{i=1}^{n} b_i P_i = \sum_{i=1}^{n} b_i (G_i + E_i)$$

Equation 3 Linear model for index selection.

where:
b_i = regression coefficient,
P_i, G_i, E_i are as defined previously.

Assumptions

Assume an equal amount of information on all individuals to be evaluated and selected. Also, assume that the distributions of $\mathbf{P_i}$, $\mathbf{G_i}$, and $\mathbf{E_i}$ are unknown but that the mean and covariances are known.

Then,

$$P_i \tilde{} NID(0, {}^{P}\sigma_{ii})(i = 1, \ldots, n)$$
$$G_i \tilde{} NID(0, {}^{G}\sigma_{ii})(i = 1, \ldots, n)$$
$$E_i \tilde{} NID(0, {}^{E}\sigma_{ii})(i = 1, \ldots, n)$$
$$Cov(G_i, E_i) = 0\ (i = 1, \ldots, n)$$
$$Cov(P_i, G_i) = {}^{G}\sigma_{ii}(i = 1, \ldots, n)$$
$$Cov(P_i, E_i) = 0\ (i = 1, \ldots, n)$$
$$Cov(P_i, P_j) = {}^{P}\sigma_{ij}\ (i = j = 1, \ldots, n)$$
$$Cov(G_i, G_j) = {}^{G}\sigma_{ij}\ (i = j = 1, \ldots, n)$$
$$Cov(E_i, E_j) = {}^{E}\sigma_{ij}\ (i = j = 1, \ldots, n)$$

Mean and Covariance of H and I

With these assumptions, we can derive the mean and covariance of H and I as in the set of equations written in Equation 4.

$$E(H) = \sum_{i=1}^{n} a_i E(G_i) = 0, \quad V(H) = \sigma_H^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \, {}^{G}\sigma_{ij}$$

$$E(I) = \sum_{i=1}^{n} b_i E(P_i) = 0, \quad V(I) = \sigma_I^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \, {}^{P}\sigma_{ij}$$

$$Cov(I, H) = \sigma_{IH} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \, {}^{G}\sigma_{ij}, \quad Cov(G_i, I) = \sigma_{GiL} = \sum_{j} b_j^{G}\sigma_{ij}$$

Equation 4 Formulae for estimating mean, variance, and covariance of H, I, and G_i.

where:
terms are as defined previously.

The objective of a selection index is to use some linear combination of trait values (I) to predict true genetic worth (H).

This can be accomplished by:

- maximizing expected genetic gain.
- maximizing the correlation of the sample index (I) with true worth (H).
- maximizing the probability of correct selection.
- minimizing the $E(I-H)^2$.

Williams (1962) showed that maximizing the correlation between I and H also maximizes the expected genetic gain and the probability of correct selection.

Derivation of the Optimum Index

Maximizing correlation of I with H (Equation 5):

$$r_{IH} = \frac{Cov(I,H)}{\sqrt{V(I)V(H)}} = \frac{\sum_i \sum_j a_i b_j \, {}^{G}\sigma_{ij}}{\sqrt{(\sum_i \sum_j b_i b_j \, {}^{G}\sigma_{ij})(\sum_i \sum_j a_i a_j \, {}^{G}\sigma_{ij})}}$$

Equation 5 Equation for estimating the correlation between *I* and *H*.

where:
terms are as defined in equation 4.

It can be shown that maximizing r_{IH} is equivalent to maximizing $\log(r_m)$ (Equation 6).

$$\log r_{IH} = \log\left(\sum_i \sum_j a_i b_j \, {}^{G}\sigma_{ij}\right) - \frac{1}{2}\log\left(\sum_i \sum_j b_i b_j \, {}^{P}\sigma_{ij}\right) - \frac{1}{2}\log\left(\sum_i \sum_j a_i a_j \, {}^{G}\sigma_{ij}\right)$$

Equation 6 Equation for maximizing correlation of I with H.

where:
terms are as defined previously.

Using least squares and differentiating with respect to b_j, we get Equation 7:

$$\frac{\partial \log(r_{IH})}{\partial_{bj}} = \frac{\sum a_i \, {}^{G}\sigma_{ij}}{\sigma_{IH}} - \frac{1}{2}\frac{2\sum_i b_i^{P}\sigma_{ij}}{\sigma_I^2} = 0 \quad (j = 1, \ldots, n), \quad \frac{\sum_i b_i \, {}^{P}\sigma_{ij}}{\sigma_I^2} = \frac{\sum_i a_i \, {}^{G}\sigma_{ij}}{\sigma_{IH}} \quad (j = 1, \ldots, n)$$

Equation 7 Calculating the correlation of I with H using least squares techniques.

where:
all terms are as defined previously.

Rewriting the Normal Equations

These equations are called the normal equations (unrelated to normal distribution) and constitute n equations in n unknowns.

They can be rewritten as in Equation 8:

$$\sum_i b_i \, {}^P\sigma_{ij} = \sum_i a_i \, {}^G\sigma_{ij} \left(\frac{\sigma_I^2}{\sigma_{IH}}\right) \; j = (1, \ldots, n)$$

Equation 8 Calculating the regression coefficient and observed value.

where:
all terms are as defined previously.

Because we are only interested in relative values of $\boldsymbol{b_i}$, the constant term can be dropped, resulting in Equation 9:

$$\sum_i b_i \, {}^P\sigma_{ij} = \sum_i a_i \, {}^G\sigma_{ij} \; j = (1, \ldots, n)$$

Equation 9 Dropping the constant values before calculating the regression coefficient and observed value.

where:
all terms are as defined previously.

Further Calculations

Considering 2 traits (n=2), Equation 9 is written as:

$$b_1 \, {}^P\sigma_{11} + b_2 \, {}^P\sigma_{21} = a_1 \, {}^G\sigma_{11} + a_2 \, {}^G\sigma_{21} \quad (j = 1); \; b_2 \, {}^P\sigma_{12} + b_2 \, {}^P\sigma_{22} = a_1 \, {}^G\sigma_{12} + a_2 \, {}^G\sigma_{22} \quad (j = 2)$$

Solving the above equations, we get Equation 10:

$$b_1 = \frac{a_1\left[{}^G\sigma_{11} \, {}^P\sigma_{22} - {}^G\sigma_{12} \, {}^P\sigma_{21}\right] + a_2\left[{}^G\sigma_{21} \, {}^P\sigma_{22} - {}^G\sigma_{22} \, {}^P\sigma_{21}\right]}{{}^P\sigma_{11} \, {}^P\sigma_{22} - {}^P\sigma_{12} \, {}^P\sigma_{21}};$$

$$b_2 = \frac{a_1\left[{}^G\sigma_{12} \, {}^P\sigma_{11} - {}^G\sigma_{11} \, {}^P\sigma_{12}\right] + a_2\left[{}^G\sigma_{22} \, {}^P\sigma_{11} - {}^G\sigma_{21} \, {}^P\sigma_{11}\right]}{{}^P\sigma_{11} \, {}^P\sigma_{22} - {}^P\sigma_{12} \, {}^P\sigma_{21}}.$$

Equation 10 Calculating the regression coefficient,

where:
all terms are as defined previously.

Minimizing E(I-H)^2

The same equations can be derived to get Equation 9 by minimizing E(I-H)^2 as written below:

$$E(I-H)^2 = E\Big[(\sum_i b_i P_i) - (\sum_i a_i G_i)\Big]^2$$
$$E(\sum_i b_i P_i)^2 - 2E\Big[(\sum_i b_i P_i)(\sum_i a_i G_i)\Big] + E(\sum_i a_i G_i)^2$$
$$\sum_i \sum_j b_i b_j \, {}^P\sigma_{ij} - 2\sum_i \sum_j a_i b_j \, {}^G\sigma_{ij} + \sum_i \sum_j a_i a_j \, {}^G\sigma_{ij}$$

because,

$$E(\sum_i b_i P_i)^2 = V(I);\ E\Big[(\sum_i b_i P_i)(\sum_i a_i G_i)\Big] = Cov(I,H);\ E(\sum_i a_i G_i)^2 = V(H)$$

Applying least squares:

$$\frac{\partial E(I-H)^2}{\partial_{ij}} = 2\sum_i b_i \, {}^P\sigma_{ij} - 2\sum_i a_i \, {}^G\sigma_{ij} = 0 \ (j = 1, \ldots, n)$$

Dividing through by 2 and rearranging, we get the normal equations:

$$\sum_i b_i \, {}^P\sigma_{ij} = \sum_i a_i \, {}^G\sigma_{ij} \ \ j = (1, \ldots, n)$$

which are identical to the equations presented previously.

Expected Genetic Gains

To derive the expected genetic gains, we need to make assumptions about the distributions of P_i, G_i, and E_i.

Assume:

1. P_i, G_i, and E_i are distributed normally with the mean and covariance structure given earlier.
2. Truncation selection on I.

Then, the expected genetic gain is estimated with Equation 11:

$$\Delta H = E(\bar{H}_S - \bar{H}) = \beta_{HI}(\bar{I}_S - \bar{I})$$

Equation 11 Calculating the genetic gain of change in genetic worth,

where:
ΔH = the genetic gain, that is, the change in genetic worth,
β_{HI} = the regression coefficient of **H** on **I**, which gives the mean value of **H** or any value of **I**.

This is the standard way to calculate predicted gains from univariate selection. See, for example, Empig et al. (1972).

Truncation Selection

The situation with regard to truncation selection is based on the following where,
Ī is the mean value of the index in the population; **c** is the truncation point; **z** is the height of ordinate of the standard normal curve at the truncation point **c**;
P is the proportion of the population selected, and $\bar{\mathbf{I}}_\mathbf{s}$ is the mean of the selected individuals.

Then, the value of the frequency of the index is estimated with Equation 12:

$$f(I) = \frac{1}{\sigma_I \sqrt{2\pi}} e^{-\left[\frac{(I-\bar{I})^2}{2\sigma^2}\right]}$$

Equation 12 Calculating the genetic gain of change in genetic worth,

where:
$f(I)$ = the frequency of individuals with index value **I**
other terms are as defined previously.

Selection Relationships

The proportion saved is related to the truncation point by $S = \int_c^\infty f(I)\partial I$, and the mean value of the selected group is $\bar{I}_S = \frac{1}{S}\int_c^\infty I f(I)\partial I$.

The selection differential (D) is given by Equation 13:

$$D = (\bar{I_S} - \bar{I}) = \frac{1}{S}\left[\int_c^\infty If(I)\partial I - \bar{I}\right] = \frac{1}{S}\int_c^\infty (I - \bar{I})f(I)\partial I$$

Equation 13 Formula for calculating the selection differential, D,

where:
S, I are as defined previously.

Standardized Regression Coefficient

Let,

$$i = \frac{I - \bar{I}}{\sigma_I};\ c = \frac{C - \bar{I}}{\sigma_I};\ \partial i = \frac{\partial I}{\sigma_I};$$

then, the standardized regression coefficient is derived as in Equation 14:

$$DD = \frac{\sigma_I}{S}\frac{1}{\sqrt{2\pi}}\int_c^\infty ie^{-[\frac{i^2}{2}]};$$

$$\partial i = \frac{\sigma_I}{S}\frac{1}{\sqrt{2\pi}}e^{-[\frac{c^2}{2}]};$$

$$D = \frac{\sigma_i}{s}z$$

Equation 14 Calculating the standardized regression coefficient,

where:
z = the height of the ordinate at the truncation point.

Typically, this is represented as $\frac{S}{\sigma_I} = \frac{z}{P} = i$, where, **k** is the standardized regression coefficient.

Expected Gain

Expected gain can then be represented as in Equation 15:

$$\Delta H = kb_{HI}\sigma_I,$$

Equation 15 Formula for expected genetic gain,

where:

$$B_{HI} = \frac{Cov(H,I)}{\sigma_I^2},$$

$$Cov(H,I)\sum_i\sum_j a_i b_j \, {}^{G}\sigma_{ij} = \sum_i b_j \sum_i a_i \, {}^{G}\sigma_{ij}$$

From the normal equations derived earlier, we have

$$\sum_i b_i \, {}^{P}\sigma_{ij} = \sum_i a_i \, {}^{G}\sigma_{ij} \quad j = (1,\ldots,n)$$

Substituting these terms,

$Cov(H,I) = \sum_j b_j \sum_i b_i \, {}^{P}\sigma_{ij} = \sum_i\sum_j b_i b_j \, {}^{P}\sigma_{ij};\ Cov(H,I) = \sigma_I^2$, we get expected genetic as written in Equation 16:

$$\Delta H = \frac{kCov(H,I)}{\sigma_I^2}\sigma_I = k\frac{\sigma_I^2}{\sigma_I^2}\sigma_I = k\sigma_I,$$

Equation 16 Alternative formula for expected genetic gain,

where:

$terms$ are as defined previously.

Predicted Gain

The predicted gain is more useful when written in terms of the correlation between **H** and I designated as $\mathbf{r_{HI}}$ (Equation 17):

$$\Delta H = \frac{kr_{IH}\sigma_H\sigma_I^2}{\sigma_I^2} = kr_{IH}\sigma_H,$$

Equation 17 Formula for predicted genetic gain,

where:

$r_{HI} = \frac{Cov(H,I)}{\sigma_I\sigma_H} = \frac{\sigma_I}{\sigma_H}$, as defined previously.

In the selection index literature, $\mathbf{r_{HI}}$ is called the **accuracy of selection** because it is a measure of how well the index, **I**, measures the true worth, **H**.

Alternative selection indices, **I**, can be compared using $\mathbf{r_{HI}}$ as long as the selection goal, **H**, remains the same for each of the indices.

Expected Genetic Gains for Each Trait

Let $\mathbf{\Delta G_i}$ be the expected genetic gain in trait ***i*** when selection is on **I**. From representations in previous equations as below, the expected genetic gains can be obtained as written in Equation 18.

$$\Delta H = \sum_i a_i \Delta G_i,$$

Equation 18 Formula for expected genetic gains for each trait,

where:

$$\Delta G_i = kB_{GiI}\sigma_I (i = 1, \ldots, n) = k\frac{Cov(G_i, I)\sigma_I}{\sigma_I^2}\left(\frac{{}^G\sigma_{ii}}{{}^G\sigma_{ii}}\right) = k\frac{Cov(G_i, I)}{\sigma_I\ {}^G\sigma_{ii}}\ {}^G\sigma_{ii} = kr_{GiI}\ {}^G\sigma_{ii}$$

,

$$Cov(G_i, I) = E[G_i, \sum_j b_j P_j] = \sum_j b_j\ {}^G\sigma_{ii},$$

$$\sum_i a_i \mathrm{Cov}(G_i, I) = \sum_i \sum_j a_i b_j\ {}^G\sigma_{ij} = \mathrm{Cov}(H, I).$$

This index requires that you know the true values of the population parameters. However, estimates of the population parameters are often substituted for the true values, and the resulting index is called the **estimated index** or the **Smith-Hazel index**.

Matrix Representation of Selection Indices

With this notation, the normal equations can be written as $Pb = Ga, and\ b = P^{-1}Ga$.

Some Results

$$\sigma_I^2 = \underline{b}'P\underline{b};\ \ \sigma_H^2 = \underline{a}'G\underline{a};\ \ \sigma_{IH} = b'Ga = b'Pb = \sigma_I^2;\ \mathrm{Cov}(G_i, I) = \sigma_{GiI} = b'G_i,$$

where G_i is the ith row of G. The genetic gain can be written as in Equation 19:

$$\Delta H = k\sigma_I = k\sqrt{\underline{b}' P \underline{b}}; \quad \underline{\Delta} = \begin{pmatrix} \Delta G_1 \\ \Delta G_2 \\ \vdots \\ \Delta G_n \end{pmatrix}; \quad \Delta = \frac{G\underline{b}}{\sqrt{\underline{b}' P \underline{b}}}; \quad \Delta H = \underline{a}' \, \underline{\Delta},$$

Equation 19 Matrix notation for expected genetic gains,

where:

$r_{IH} = \sqrt{f \frac{\underline{b}' G \underline{a}}{\underline{a}' G \underline{a}}},$

other terms are as defined previously.

Construction of a Selection Index

Optimum Index

The normal equations can be written for an optimum index as $P\underline{b} = G\underline{a}; \; \underline{b} = P^{-1} G \underline{a},$ where **P**, **G**, and **a** are known without error, and the index is as in Equation 20:

$$I = b' \underline{x}; \; H = \underline{a}' \, \underline{y},$$

Equation 20 Formula for optimum index,

where:

x = trait x,

y = trait y.

Smith-Hazel Index

This index is the same as the optimum index; only in this case (Equation 21), we use estimates of P, G, and a, designated as $\hat{P}$, $\hat{G}$, and $\hat{a}$, respectively.

$$\hat{P}\hat{\underline{b}} = \hat{G}\hat{\underline{a}}; \;\; \hat{\underline{b}} = \hat{P}^{-1} \hat{G} \hat{a}; \;\; \hat{I} = \hat{b}' \underline{x}; \; H = \hat{\underline{a}}' \underline{y},$$

Equation 21 Formula for Smith-Hazel index,

where:
$terms$ are as defined previously.

Base Index

The Base Index was apparently first suggested by Brim et al. (1959) and named the Base Index by Williams (1962). The base index is constructed simply by allowing $\boldsymbol{b}$ = $\boldsymbol{a}$ and written as in Equation 22, with all terms defined earlier.

$$I = \underline{a}'\underline{x}; \quad H = \underline{a}'\underline{y},$$

Equation 22 Formula for Smith-Hazel index,

Some results for this index include:

$$\sigma_I^2 = \underline{a}'P\underline{a}; \;\; \sigma_H^2 = \underline{a}'\underline{a}; \;\; \mathrm{Cov}(I,H) = \underline{a}'G\underline{a}; \;\; r_{IH} = \frac{\underline{a}'G\underline{a}}{\sqrt{(a'Pa)(a'Ga)}} = \sqrt{\frac{\underline{a}'G\underline{a}}{\underline{a}'P\underline{a}}}$$

The foremost attribute of this index is its simplicity of construction and interpretation. Also, this index does not require the estimation of genetic parameters.

Multiplicative Index

The multiplicative index was first proposed by Elston (1963). This index is also sometimes called the weight-free index because it does not require the specification of index weights or economic values.

The general form of this index is as in Equation 23:

$$I = (X_1 - k_1)(X_2 - k_2)\cdots(X_n - k_n),$$

Equation 23 Formula for multiplicative index,

where:
k_1 = the minimum value of trait $\mathbf{X_1}$ set by the breeder.

In addition to being weight-free, this index also does not require the estimation of genetic or phenotypic parameters. Because this is a curvilinear index, theory is not available to predict gains. Baker (1974) found that this index can be approximated by using a linear index, where the weights are the reciprocals of the phenotypic standard deviations of the traits in the index.

This essentially amounts to an index with equal weighting per phenotypic standard deviation. Approximate, predicted gains can then be obtained for this index using the Smith-Hazel index theory.

Desired Gain Index

The desired gain index was suggested by Pesek and Baker (1969). This index allows the breeder to specify a vector of desired gains, $\underline{q}$, and then substitute this into the predicted gain equation and solve for $\underline{b}$the index weights. The solution for $\underline{b}$ is as in Equation 24:

$$\underline{b} = G^{-1}\underline{q},$$

Equation 24 Formula for desired gain index,

where:
$\underline{q}$ = the vector of desired gains,
$I = \underline{b'x}$ is as defined previously.

This index was proposed to eliminate the need to specify economic weights. However, in practice, there are some difficulties with the index in specifying the vector of desired gains.

This index will result in maximum gains in each trait according to the relative importance assigned by the breeder in specifying the desired gains.

Predicted gains can be obtained by substituting the vector of index weights into the conventional Smith-Hazel predicted gain equations.

Restricted Selection Index

Restricted selection indices were first derived by Kempthorne and Nordskog (1959). Since then, various restricted indices have been derived by Cunningham et al. (1970) and James (1968). See Lin (1978) for a complete list. Basically, restricted selection indices involve holding the genetic gains in one or more traits to a constant or zero while changing the means of other traits in the desired direction. The basic method is to impose the restriction on the index equations that $Cov(G_i, I) = 0$.

The simplest procedure to accomplish this was given perhaps by Cunningham et al. (1970). Their method involved solving the following set of equations in Equation 25:

$$\begin{pmatrix} P & G_i \\ G_i & O \end{pmatrix} \begin{pmatrix} \underline{b} \\ \underline{b_d} \end{pmatrix} = \begin{pmatrix} G \\ O \end{pmatrix} [a],$$

Equation 25 Formula for restricted selection index,

where:

$b = [I - P^{-1}G_i(G_i'P^{-1}G_i)^{-1}G_i']P^{-1}G\underline{a}$,

$b_d = \underline{b_d} = (G_i'P^{-1}G_i)^{-1}G_i'P^{-1}G\underline{a}$,

other terms are as defined previously.

$\underline{\boldsymbol{b}}$ is the vector of index weights to use in the index equation as $I = \underline{b'x}$, as before. The dummy variable is not used in the index equation.

This method has the interesting consequence that the value obtained for the dummy variable is the negative of the economic weight needed to produce zero change in that trait in an unrestricted selection index.

Rank Summation Index

The rank summation index was first suggested by Mulamba and Mock (1978). Basically, this index involves obtaining the ranks of each of the traits to be included in the index and then calculating the index by summing up the trait ranks, represented in Equation 26.

$$I = \sum_{i=1}^{n} \text{rank}(X_i).$$

Equation 26 Formula for the rank summation index,

where:

$terms$ are as defined previously.

The primary advantages of this index are that genetic parameters need not be calculated, it transforms the data so that the variances for each trait are identical, and it does not require the specification of economic weights, although they can be used.

As with the multiplicative index, predicted gains cannot be calculated for this index. However, Crosbie et al. (1980) found that the same prediction equation used for the multiplicative index provides a reasonably good approximation of the predicted gains for the rank summation index.

Selection Index Efficiency

Methods to Compare Selection Index Efficiency

Cunningham (1969) provided a method for comparing the relative efficiencies of selection indices. He was primarily interested in deleting traits from the index so that their relative contribution to the gain in the true worth (H) could be determined. Dropping traits from the index means that fewer genetic parameters need to be estimated, providing considerable cost savings.

Define the index containing all the traits of interest as the original index and define the index with one trait dropped out as the i^{th} reduced index. Then the efficiency of the i^{th} reduced index relative to that of the original index is the ratio of their standard deviations. Cunningham showed this to be as in Equation 27:

$$\sqrt{\frac{\underline{b}' P \underline{b} - \frac{b_i^2}{W_{ii}}}{\underline{b}' P \underline{b}}}$$

Equation 27 Formula for selection index efficiency,

where:
b_i = the i^{th} weighting factor in the original index,
W_{ii} = the corresponding diagonal element in the inverse of P.

A more usual procedure is to compare the gain for the i^{th} trait when selection is on I, relative to the single trait selection for the i^{th} trait.

Effect of Correlations on Index Weights

To determine the effects of correlation on index weights, we need to derive the index equations $P\underline{b} = G\underline{a}$ in terms of genetic and phenotypic correlation coefficients (Equations 28, 29, 30) following the series of derivations as follows. Let,

$$X_i^* = \frac{X_i}{\sqrt{{}^P\sigma_{ii}}} = \frac{G_i + E_i}{\sqrt{{}^P\sigma_{ii}}} \ (i = 1, \ldots, n)$$

Then:

$$V(X_i^*) = {}^P\sigma_{ii} = 1$$

$$\text{Cov}(X_i^*, X_j^*) = \frac{{}^P\sigma_{ij}}{\sqrt{{}^P\sigma_{ii}\ {}^P\sigma_{ij}}} = {}^Pr_{ij} \ (i = j = 1, \ldots, n)$$

$$V(G_i^*) = V(\frac{G_i}{\sqrt{{}^P\sigma_{ii}}}) = \frac{{}^G\sigma_{ii}}{{}^P\sigma_{ii}} = h_i^2$$

$$Cov(G_i^*, G_j^*) = \frac{{}^G\sigma_{ij}}{\sqrt{{}^P\sigma_{ii}\ {}^P\sigma_{jj}}} = h_{ij}$$

Derivation

$$P^* = \begin{pmatrix} 1 & {}^Pr_{12} & \cdots & {}^Pr_{1n} \\ {}^Pr_{21} & 1 & \cdots & {}^Pr_{2n} \\ \vdots & \vdots & & \vdots \\ {}^Pr_{n_1} & {}^Pr_{n_2} & \cdots & 1 \end{pmatrix}; \quad G^* = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix}; \quad \underline{\hat{b}}^* = (P^*)^{-1} G^* \underline{a}$$

Equation 28 Formula for phenotypic and genotypic correlations,

where:
$terms$ are as defined previously.

For 2 traits (n=2),

$$\underline{\hat{b}}^* = \begin{pmatrix} a_1({h_1}^2 - {}^Pr_{12}h_{12}) + a_2(h_{12} - {}^Pr_{12}{h_2}^2) \\ a_1(h_{12} - {}^Pr_{12}{h_1}^2) + a_2({h_2}^2 - {}^Pr_{12}h_{12}) \end{pmatrix} (\frac{1}{1 - ({}^Pr_{12})^2})$$

Equation 29 Formula for estimated correlation between two traits,

where:
$terms$ are as defined previously.

When correlations are zero: ${}^{\mathbf{P}}\mathbf{r_{12}} = {}^{\mathbf{G}}\mathbf{r_{12}} = 0$,

$$\hat{b}^* = \begin{pmatrix} a_1 h_1^2 \\ a_2 h_2^2 \end{pmatrix}; \quad P = \begin{pmatrix} {}^P\sigma_{11} & 0 \\ 0 & {}^P\sigma_{22} \end{pmatrix}; \quad P^{-1} = \begin{pmatrix} \frac{1}{{}^P\sigma_{11}} & 0 \\ 0 & \frac{1}{{}^P\sigma_{22}} \end{pmatrix}; \quad G = \begin{pmatrix} {}^G\sigma_{11} & 0 \\ 0 & {}^G\sigma_{22} \end{pmatrix}$$

Equation 30 Formula for estimated correlation between two traits when phenotypic and genetic correlations are zero,

where:
$terms$ are as defined previously.

When | r | < 0.30, the use of the above index is nearly as efficient as using the Smith-Hazel index.

References

Brim, C. A., H. W. Johnson, and C. C. Cockerham. 1959. Multiple selection criteria in soybeans. Agron J 51: 42-46.

Crosbie, T. M., J. J. Mock, and O. S. Smith. 1980. Comparison of gains predicted by several selection methods for cold tolerance traits in two maize populations. Crop Sci. 20:649-655.

Cunningham, E. P. 1969. The relative efficiencies of selection indexes. Acta Agriculturae Scandinavica 19(1): 45-48.

Elston, R. C. 1963. A weight-free index for the purpose of ranking or selection with respect to several traits at a time. Biometrics 19(1): 85-97.

James, J. W. 1968. Index selection with restrictions. Biometrics 24, 1015-1018.

Kempthorne, O., and A. W. Nordskog. 1959. Restricted selection indices. Biometrics 15:10-19.

Lin, C. Y. 1978. Index selection for genetic improvement of quantitative characters. Theoretical and Applied Genetics, 52(2): 49-56.

Mulamba, N N., and J. J. Mock. 1978. Improvement of yield potential of the Eto Blanco maize (Zea mays L.) population by breeding for plant traits. Egypt. J. Genet. Cytol. 7: 40-51.

Pesek. J., and R. J. Baker. 1969. Desired improvement in relation to selection indices. Can J Plant Sci, 9: 803–804.

Williams, J. S. 1962. The evaluation of a selection index. Biometrics 18:375-393.

How to cite this module: Beavis, W., K. Lamkey, and A. A. Mahama. 2023. Multiple Trait Selection. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics in Plant Breeding*. Iowa State University Digital Press.

Chapter 12: Multi Environment Trials: Linear Mixed Models

William Beavis and Anthony Assibi Mahama

The general linear model (Equation 1) can be applied to replicated trial data for basic prediction purposes. It is, however, not adequate in all experimental situations, especially where trials are conducted in more than one environment. This chapter explores the role of mixed linear models, BLUEs (Best Linear Unbiased Estimates), and BLUPs (Best Linear Unbiased Predictors) in the analysis of multiple environment trial data to characterize and select among entries.

$$Y_{ij} = \mu + g_i + r_j + \varepsilon_{ij}$$

Equation 1 General linear model for basic prediction of replicated trials.

where:
Y_{ij} = the phenotypic response,
μ = population (or overall) mean,
g_i = genotypic units,
r_j = number of replicates of the genotypic units,
ε_{ij} = residual source of variability.

Learning Objectives

- Conceptual basis of mixed linear models
- Review matrix algebra
- The meaning of BLUE and BLUP

Henderson's Concept

C.R. Henderson recognized the challenge of prediction using models such as Equation 1 and addressed it using the concept of shrinkage estimators for the genotypic units in the model. Note that the fitted regression line provides predictions that are 'shrunken' to the line rather than scattered around the line. Henderson's idea, first published in 1963, was framed in the context of the matrix form of Equation 2, which can be explained using scalar algebra.

First, let us obtain phenotypic averages for each genotypic unit. Next minimize the difference

of $\boldsymbol{N_1}$, where **E** represents the expectation, that is, the average for the genotypic unit, **m** is the population mean and $\mathbf{g_i}$ is the genotypic value from the scalar version of model Equation 2. In this case, we need to find a value that will ensure that the sum of the squared differences is minimal. As with Equation 2, a little knowledge of how to obtain partial derivatives provides the answer:

$$w_i = \frac{\sigma_g^2}{\left(\frac{\sigma_g^2+\sigma_e^2}{r}\right)}$$

Equation 2 Formula for calculating intra-class correlation.

where:
w_i = intra-class correlation coefficient or broad sense heritability,
σ_g^2 = genotypic variance,
σ_e^2 = residual (or error) variance,
r = number of replications.

This is known as the **intra-class correlation coefficient**. It is also known as the **broad sense heritability**, but for now we will refer to it as a **shrinkage factor**. When $\mathbf{w_i}$ is multiplied by ($\mathbf{Y_i}$ – **m**) it will provide the Best Linear Unbiased Predictor of $\mathbf{g_i}$. Notice that the predictions of genotypic values are scaled towards the mean BV, which by definition is zero.

Example Prediction 1

If the overall mean is the only fixed effect (one environment), all lines are unrelated to each other, and the data are balanced, then the predicted genotypic value is obtaining using Equation 3:

$$\hat{u}_j = w\left(Y_j - \hat{Y}\right)$$

Equation 3 Formula for calculating predicted genotypic value.

where:
$\hat{u}_j$ = predicted genotypic value of genotype***j***,
w = the Shrinkage factor,
Y_j = the phenotype of genotype ***j***,
$\hat{Y}$ = the predicted phenotype.

If w is equal to zero, $\hat{u}_j$ would be zero.

If w is equal to one, $\hat{u}_j$ equals the phenotypic values.

Let us demonstrate this with a simple data set in which four unrelated lines (**A, B, C, D**) were evaluated for yield (t/ha) in hybrid combination with a single tester (**Z**) in single rep tests at **N** environments (Table 1). For this simple example we are only interested in the impact of number of environments (replicates) on $\mathbf{w_i}$ and its subsequent impact on the predicted value for each $\mathbf{g_i}$. Also, assume that the residual variance, $\boldsymbol{\sigma_e^2}$= **40**.

Summary Data

Table 1 Summary data of four inbreds evaluated in hybrid combination with one tester (Z) in single rep tests at 10 environments.

Hybrid	$\bar{Y}_{i.}$	$\bar{Y}_{..}$	$\bar{Y}_{i.} - \bar{Y}_{..}$	N_1	$w_i(Y_{i.} - \bar{Y}_{..})$	N_2	$w_i(Y_{i.} - \bar{Y}_{..})$
AxZ	7	10	-3	10	-2.5	2	-1.5
BxZ	9	10	-1	10	-0.83	2	-.05
CxZ	11	10	1	10	0.83	2	0.5
DxZ	13	10	3	10	2.5	2	1.5

Prove for yourself that the estimated σ_e^2 = 20.

Some things to notice from the table:

- The data are from balanced trials, i.e., all genotypic units are evaluated in the same number of environments (either 2 or 10).
- With a large N, the observed differences will be equal to the predicted values.
- For balanced trials, shrinkage does not change the relative ranking.

In essence the shrinkage predictor provides us with a value that not only includes the difference relative to the mean, but also weights it by our confidence in the magnitudes of the differences from the overall mean.

We need to consider how to obtain predictions for genotypic units in the more likely situations where not all genotypic units (lines, cultivars, hybrids) will be evaluated equally in all environments. Indeed, we now find it possible with marker technologies to predict the values of the genotypes before they have been grown.

Best Linear Unbiased Prediction

Henderson's shrinkage predictor can now be considered in the context of the matrix form of the mixed model equation (Equation 4):

$$Y = X\beta + Z\mu + e$$

Equation 4 Mixed model equation for predicting phenotype.

where:
Y = Vector of observations (phenotypes),
X = Design matrix for fixed effects,
β = Vector of unknown fixed effects (to be estimated),
Z = Design matrix of random effects,
μ = a vector of random effects (genotypic values to be estimated),
e = a vector of residual errors (random effects to be estimated).

The random effects are assumed to be distributed as
$u \sim MVN(0, A),\ and\ \varepsilon \sim MVN(0, R)$.

Just as estimates for β in the matrix form of Equation 4 can be found using the normal equations, the normal equations for Equation 4 can be used to find least squares estimates for the parameters in Equation 5.

$$\begin{bmatrix} \hat{\beta} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}X + A^{-1}(V_r/V_A) \end{bmatrix}^{-} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Equation 5 Normal equations in matrix notation.

where:
$\hat{\beta}$ = the fixed effects parameters,
$\hat{v}$ = the random effects for the parameters,
X = incidence matrix,
y = a vector of observed phenotype (e.g., yield),
A = the additive relationship matrix,
R = the diagonal matrix,
Z = incidence matrix,
V_A = the additive variance,
VR = the residual variance.

BLUEs and BLUPs

The values for $\hat{\beta}$ represent the Best Linear Unbiased Estimators (BLUE) of the fixed effects, while the values for $E\left(w_i\left[\bar{Y}_i - \mu\right] - g_i\right)^2$ represent the Best Linear Unbiased Predictors (BLUP) of the random effects. It is important to remember that BLUE's and BLUP's are **not** methods; they are statistical properties of methods (there are many) that are capable of producing such values. These statistical properties include:

- **Best:** the sampling variance of what is being estimated or predicted is minimized.
- **Linear:** estimates or predictions are linear functions of the observations.
- **Unbiased:** in BLUE indicates that the expected values of the estimates are equal to their true values. In BLUP, indicates that the sum of the predictions have an expectation of zero.
- **Estimators and Predictors:** refer to algorithms that generate the estimated or predicted values.

For BLUE's the effects are considered fixed. Examples include the overall mean, effects of different soil types, fertilizer treatments, etc. From a practical perspective, fixed effects do not have a covariance structure. Due to the practical perspective, we often consider environments as fixed effects.

Effects of BLUPs

The effects of BLUPs are considered random and it is possible to define covariance structures associated with these effects. Examples include breeding values, dominance effects, tester effects, etc. The challenge for application of methods that provide BLUPs is that Equation 3 assumes covariances and variances are known. The truth is that the variances of genetic and non-genetic random effects are not known. Rather in practice we estimate these values. Thus, all implementations of methods that provide BLUPs from mixed linear model equations provide only approximations of the unknown vector values.

Nonetheless, BLUP values are useful in practical plant breeding trials where designs are unbalanced. Indeed, a method that produces a BLUP value enables the estimation of genetic variances without having to resort to mating designs to obtain estimates of heritability. A typical trial will have different numbers of genotypic units from different families evaluated in different sets of environments, some replicated some not. BLUPs utilize covariance structures (covariances among genotypic units grown in the same sets of environments and covariances among relatives) to maximize information on the traits of interest. Thus, the true purpose of a plant breeding trial (to compare genotypes for purposes of selection), is enabled with the best possible values

for comparison because BLUPs maximize the correlation between the true genotypic values and predicted values.

Example

While Equation 5 initially appears to be daunting, with the little bit of matrix algebra, introduced above, you have the skill to do these analyses using EXCEL. For example, consider the simple data set in Table 2 (adapted from Chapter 11 of Bernardo, 2010).

Note: This is an example of a self-pollinating crop (barley) and the number of environments does not factor into solving of the equation.

Table 2 Sample data from Chapter 11 of Bernardo, 2010.

Environments	No. of Env	Line	Yield
Low yield	18	1	4.45
Low yield	18	2	4.61
Low yield	18	4	5.27
High yield	9	2	5.00
High yield	9	3	5.82
High yield	9	4	5.79

We desire to translate this into the following model in Equation 6.

$$Y_{ijk} = \mu + G_i + E_j + GE_{ij} + \epsilon_{(ij)k},\ i = 1, \cdots, g; \quad j = 1, \cdots, e; \quad k = 1, \cdots, n$$

Equation 6 General linear model for basic prediction of replicated trials.

where:
Y_{ijk} = the phenotypic response,
μ = population (or overall) mean,
G_i = genotypic effect,
E_j = environmental effect,
GE_{ij} = genotypic by environment interaction effect,
$\varepsilon_{(ij)k}$ = residual source of variability.

In matrix notation, the data are represented in the model, (Equation 4, $y = X\beta + Z\mu + \epsilon$) as in Equation 7:

$$\begin{bmatrix} 4.45 \\ 4.61 \\ 5.27 \\ 5.00 \\ 5.82 \\ 5.79 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{bmatrix}$$

Equation 7 Matrix notation of Equation 4 and Table 2 data.

Linear Mixed Model Solution

The LMM solution is represented in Equation 8:

$$\begin{bmatrix} \hat{\beta} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}X + A^{-1}(V_r/V_A) \end{bmatrix}^{-} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

Equation 8 Linear mixed model solution for Equation 7.

where:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \end{bmatrix}, X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, y = \begin{bmatrix} 4.45 \\ 4.61 \\ 5.27 \\ 5.00 \\ 5.82 \\ 5.79 \end{bmatrix}, Z = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$R = \begin{bmatrix} \frac{1}{18} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{18} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{18} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{9} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{9} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{9} \end{bmatrix}, A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Thus, **R** represents a matrix that weights the calculations by the number of observations that contribute to the estimated mean values of each cultivar in each type of environment.

Estimated Residual Variance

Assuming that the lines are unrelated to each other,

$A = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$, *and* $\frac{V_R}{V_A}$ is the ratio of the estimated residual variance (sometimes incorrectly referred to as the estimate of the experimental error) to the estimated additive genetic variance. For purposes of illustration let us consider this estimated ratio to be 5, i.e., the estimated additive genetic variance is 20% as large as the residual variability.

Calculations for the example have been implemented in an EXCEL spreadsheet: BLUEs and BLUPs of 4 barley varieties [XLSX].

As an exercise to conduct on your own, consider implementing the LMM.7 for the example on estimation of means using **lsmeans** discussed in the review module:
Review of EDA and Estimation [DOC]:

- Download R-code example eda aov lsmeans [TXT]
- Download R-code example for mixed models [TXT]
- Download R-code example Regression [TXT]

Reference

Henderson, C. R. 1963. Estimation of variance and covariance components. Biometrics. 9: 226.

How to cite this chapter: Beavis, W. and A. A. Mahama. 2023. Multi Environment Trials: Linear Mixed Models. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Chapter 13: Simulation Modeling

William Beavis and Anthony Assibi Mahama

Quantitative genetic models are used to represent, describe, and quantify the genetic contributions to natural phenomena. These models can be arbitrarily simple, e.g., additive linear models, or complex, e.g., non-additive, non-linear models. R.A. Fisher and Sewell Wright had a decades-long debate about which type of model should be considered in the study of natural and artificial selection. Fisher and his disciples argued that the more complex models were not needed. Sewell Wright and his students argued that biology was inherently complex and needed non-linear non-additive models to accurately understand adaptation and evolution. Of course, both were correct, and both were wrong. As George Box reminds us, all models are wrong; some are useful. The choice of an appropriate model depends upon the purpose of the research.

Prior to this chapter, we investigated the development of theoretical quantitative genetic models for the purposes of conducting and interpreting analyses of plant breeding experiments. As noted, without the theoretical models, there would be no genetic understanding of the results. Theory provides predictions, and predictions are the basis for generating testable hypotheses. In this chapter, we introduce a far more practical justification for theoretical models: With a theoretical model, it is possible to simulate many different data analysis techniques and breeding strategies prior to conducting expensive experiments. In other words, natural and artificial systems can be modeled *in silico* for purposes of predicting unknown outcomes. Many *in silico* experiments can be compared, and the most promising can be used to compare methods or processes. If comparisons are based on objective criteria, such as accuracy, power, precision, efficacy, and efficiency, and if the model used for data analyses is the same as the model used for simulating data sets, we can make rational decisions about which methods to implement in plant breeding experiments.

Learning Objectives

- Recognize limitations of experimental research
- Translate QG models to simulation models
- Translate simulation models to EXCEL software functions
- Build confidence in the use of simulation models

History of Simulations

Geneticists first used computers to implement simulation models to evaluate limits to artificial selection (Hill and Robertson, 1968; Bulmer, 1976) in closed breeding populations. By 1988, Oscar Kempthorne, one of R.A. Fisher's disciples, pointed out that the classical experimental and algebraic approaches were limited to unrealistic assumptions in breeding and evolutionary systems. Since 1988, plant and animal geneticists and breeders have used simulation models to evaluate the limits to emerging statistical methods (Beavis, 1994) and to choose among selection methods because experimental evaluation of breeding methods is time and resource-limited (Podlich and Cooper, 1998). To date, there have been over 15,000 publications in which the terms simulation and breeding occur in the title. Currently, there are numerous simulation software packages that have been developed and implemented for public and private research enterprises. Some are quite simple, while others are very flexible and complex. Generally, as the flexibility of the package increases, the learning curve associated with the complexity of the package also increases. Thus, some of these simulation packages require entire courses and years to master. While it is beyond the scope of this chapter to either advocate or teach any particular simulation package, we will learn how to implement the core quantitative and population genetic models that are part of every useful simulation package.

Core Elements

The core elements of simulation modeling include the genetic architecture of the trait(s), the structure of segregating generations derived from breeding populations, and the organization of the segregating genomes.

Before we decide on the genetic architecture of the trait, we need to know the **structure of the segregating generation derived from the breeding population**. In diploid organisms, there are usually three genotypes at a SNP locus: {aa, ac, cc} or {tt, tg, gg}. Let us consider a locus with the second triplet, {tt, tg, gg}. If we decide to simulate a random mated population, then each of the three genotypes {tt, tg, gg} will occur at a frequency of p^2, $2pq$, and q^2. To decide which genotype is going to be assigned to an individual, we should obtain a random sample of a number from the Uniform[0,1] distribution. If the random number is in the interval $[0,p^2]$, then we will assign the genotype 'gg' to individual *i*. If the random number is in the interval $[p^2, p^2+2pq]$, then we will assign the genotype 'tg' to individual *i; otherwise*, we will assign the genotype 'tt' to individual *i*.

Obtaining a random sample from any distribution will depend on the syntax of the software system we decide to use for simulating the data. Since most students have experience with spreadsheet types of software, we will first learn how to use Excel for simulating SNP genotypes at a single locus in a random mating population, where p (frequency of g) = 0.3. The frequency

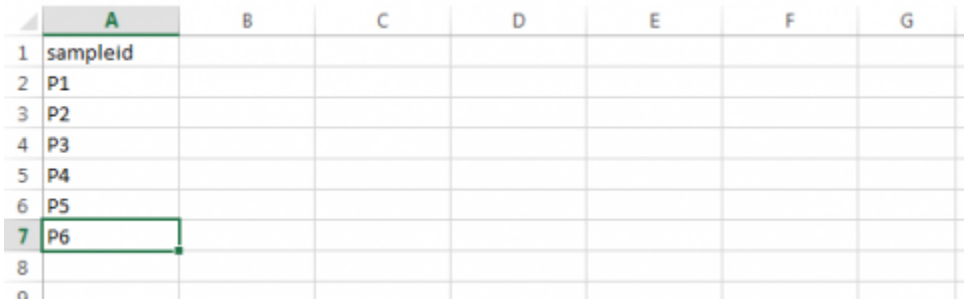
of 'gg' genotypes at this locus will be 0.09, the frequency of 'tt' genotypes will be 0.49, and the frequency of heterozygous genotypes will be 0.42. Thus, if we sample a random number from the uniform distribution in the interval [0, 0.09], then we will assign the genotype 'gg' to individual *i*; in the interval from (0.09, 0.51], we will assign the genotype 'gt' to individual *i*; otherwise, we will assign the genotype 'tt' to individual *i*.

Excel-Based Simulation

Step 1

With these parameters, we will use the Excel functions "IF" and "RAND" in the following steps:

1. **Assign a sampleid designator to the first column.**

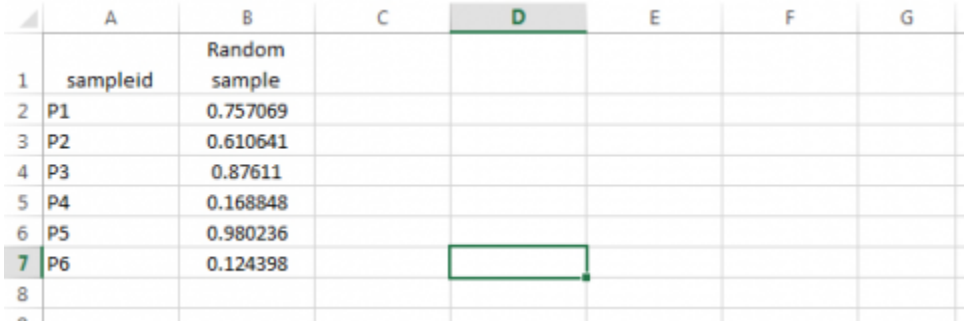


	A	B	C	D	E	F	G
1	sampleid						
2	P1						
3	P2						
4	P3						
5	P4						
6	P5						
7	P6						
8							

Step 2

With these parameters, we will use the Excel functions "IF" and "RAND" in the following steps:

2. **Obtain a random number from the Uniform Distribution for each of the sampleid's. Syntax: type** =RAND() **in cell B2, then drag across all cells in column B.**



	A	B	C	D	E	F	G
1	sampleid	Random sample					
2	P1	0.757069					
3	P2	0.610641					
4	P3	0.87611					
5	P4	0.168848					
6	P5	0.980236					
7	P6	0.124398					
8							

Step 3

With these parameters, we will use the Excel functions "IF" and "RAND" in the following steps:

3. **Based on the random number assigned to each sampleid, assign a genotype to the locus. Syntax: type** =IF(B2<=0.09,"gg",IF(B2>0.51,"tt","gt")) **in cell C1, then drag across all relevant cells. Note that values in columns B and C will likely differ from those in the example below.**

	A	B	C	D	E	F	G
1	sampleid	Random sample	locus1 genotype				
2	P1	0.757069	gt				
3	P2	0.610641	tt				
4	P3	0.87611	tt				
5	P4	0.168848	gg				
6	P5	0.980236	tt				
7	P6	0.124398	tt				
8							

Step 4

If we do not want the values generated by the RAND function to change as we add functions to the spreadsheet, we should plan to cut and paste a set of the actual values from one of the sampling events into new columns:

	A	B	C	D	E	F	G
1	sampleid	Random sample	locus1 genotype	fixed values	fixed genotype		
2	P1	0.72136	tt	0.147935	gt		
3	P2	0.506127	gt	0.657802	tt		
4	P3	0.336079	gt	0.940349	tt		
5	P4	0.090098	gt	0.022139	gg		
6	P5	0.461148	gt	0.764216	tt		
7	P6	0.856971	tt	0.735858	tt		
8							

Other Population Structures

There are many other possible breeding population structures; some are the result of designed crosses (see Chapter 8 on mating designs), but most population structures emerge from long-term breeding programs in which elite homozygous cultivars are crossed to promising homozygous lines through opportunistic networks of crosses. Simulating genotypes at segregating loci from any mating design or breeding program can be obtained in a manner described in the previous

paragraph. We need only decide on the frequencies of the genotypes in the segregating populations. For example, in the specific case of an F_2 generation derived from a cross of two inbred lines, $p = q = 0.5$, and if a random number obtained from a uniform distribution, U[0,1], is greater than ¼ and less than ¾ then the genotype of individual plant *i*, will be 'gt'. We could be interested in the case of recombinant inbred lines derived in the F_5 generation of a cross of two inbred lines. In this case, the frequency of homozygotes is now p^2+pqF and q^2+pqF, and the frequency of heterozygous lines is $2pq(1-F)$, where F = 0.875. Thus, if a random number obtained from U[0,1] is less than 0.4687, then RIL. *i* will be assigned the genotype 'gg'. It should be obvious that it should be possible to generate mixtures of segregating families from multiple independent or related crosses and simulate genotypes for any particular locus to all individuals in all families, regardless of how the families are derived.

Genetic Architecture of the Trait

Next, we need to decide how many loci will influence a trait and whether the alleles at the loci will interact. Let us begin with a single-locus additive quantitative trait, designated P. Further, consider a trait with an average phenotypic expression of 50 units and phenotypic variability in a diploid species that is due to additive genetic variability at a single locus and non-genetic variability. Initially, let us plan to let half of the phenotypic variability be due to segregation at the locus and half due to non-genetic sources of variability. The first step is to translate this brief description into a quantitative genetic model (Equation 1), preferably the same model that will be used in the eventual analysis of the phenotypic trait:

$$P_{ij} = \mu + G_i + \varepsilon_{ij}$$

Equation 1 Quantitative genetic model for phenotypic trait,

where:
P_{ij} = the phenotype,
G_i = the genotype effect,
i = one of three possible genotypes conferred by two alleles,
j = one of the repeated samples of the i^{th} genotypes,
ε_{ij} = non-genetic source of variability and is ~ i.id $N(0,\sigma_\varepsilon)$.

Thus, the variance model of Equation 1 is as written in Equation 2:

$$\sigma_P^2 = \sigma_G^2 + \sigma_\varepsilon^2$$

Equation 2 Variance model for phenotypic trait,

where:
σ^2_P = the phenotypic variance,
σ^2_G = the genotypic variance,
σ^2_ε = residual or non-genotypic variance.

Parameter Assignment

The next step is to assign values to each of the parameters in the model. Mean, μ is assigned the value of 50, and values for e_{ij} are sampled from a Normal distribution with a mean of zero and a standard deviation of σ_e. We decided that we want to simulate data in which $\sigma^2_P = \sigma^2_G + \sigma^2_\varepsilon$ (Equation 2 above).

We can choose any value for σ^2_P, but it is often best to choose a value that is ~ consistent with estimates from field trials for the crop of interest. In this case, let us say our field trials have typically produced estimates of phenotypic variance of ~ 98. Thus, both σ^2_G and σ^2_ε are ~ 49. Thus, we can obtain values for ε_{ij} by sampling a normal distribution with mean = 0 and standard deviation = 7.

Genotypic Values

We also need numeric values for each of the genotypes. Recall from Quantitative Genetic Models Theory we can assign coded genotypic values to each genotype as follows:

- Coded genotypic value of one homozygote (gg) = +a;
- Coded genotypic value of the other homozygote (tt) = -a;
- Coded genotypic value of heterozygotes (tg or gt) = d.

Since we are simulating an additive genetic model, the genotypic value of the heterozygotes (d) is midway between the two values for the homozygotes, i.e., d = 0. Thus,

G_i = a for i = “gg”; -a for i = “tt”; and 0 for i = “gt” or “tg”.

Now we need a numeric value for a.

Calculations

Recall that $\sigma^2_P = \sigma^2_G + \sigma^2_\varepsilon$ (Equation 2), and the additive portion of genetic variance is represented by Equation 3:

$$\sigma_A^2 = 2pq[a + d(q - p)]^2, \quad and \quad \sigma_D^2 = (2pqd)^2$$

Equation 3 Formula for calculating additive genetic variance.

where:
σ_A^2 = the additive genetic variance,
σ_D^2 = the dominance variance,
p and q = the frequency of the two alleles (*g* or *t*),
a and d = coded genotypic values.

Since we have decided to simulate d = 0, the genetic variance is all due to additive effects ((Equation 4):

$$\sigma_G^2 = \sigma_A^2 = 2pqa^2$$

Equation 4 Algebraic formula for calculating genotypic variance.

where:
All terms are as defined previously.

Thus, $a = \sqrt{\frac{49}{2pq}}$.

If we assume that the frequency of 't' (or 'g') in the population is $\frac{1}{4}$, then a reasonable value for $a = \sqrt{\frac{49}{2pq}}$ ~19.80.

Excel Application

Next, let us translate these values for the parameters into Excel functions.

Syntax for assigning 'a': Type `=IF(E3="gg",11.43, IF(E3="tt",-11.43,0))` in Cell G3, then drag across all relevant cells.

	A	B	C	D	E	F	G
1	sampleid	Random sample	random genotype	fixed values	fixed genotype	μ	*a*
2	P1	0.149166	gt	0.147935	gt	50	0
3	P2	0.212785	gt	0.657802	tt	50	-11.43
4	P3	0.769295	tt	0.940349	tt	50	-11.43
5	P4	0.255968	gt	0.022139	gg	50	11.43
6	P5	0.123113	gt	0.764216	tt	50	-11.43
7	P6	0.060002	gg	0.735858	tt	50	-11.43
8							

In order to fully understand how to sample from a normal distribution requires knowledge of probability density functions, cumulative density functions, and integral calculus that enables the translation between the two. These are topics beyond our current scope but worth exploring by those who wish to develop their own simulation capabilities.

Normal Distribution Interval

For our immediate purpose, the Syntax for obtaining values for ε_{ij} by sampling a Normal distribution with mean = 0 and standard deviation = 7 is the following:

Type `=NORMINV(RAND(),0,7)` in cell H3 and then drag across all relevant cells.

	A	B	C	D	E	F	G	H	I	J
1	sampleid	Random sample	random genotype	fixed values	fixed genotype	μ	G_i	ε_{ij}		
2	P1	0.694765	tt	0.147935	gt	50	0	-10.6475		
3	P2	0.274265	gt	0.657802	tt	50	-11.43	-3.17286		
4	P3	0.736703	tt	0.940349	tt	50	-11.43	0.861763		
5	P4	0.317299	gt	0.022139	gg	50	11.43	4.939074		
6	P5	0.772776	tt	0.764216	tt	50	-11.43	-12.0852		
7	P6	0.019678	gg	0.735858	tt	50	-11.43	-0.02329		
8										

Simulated Phenotypes

We now have values for all of the parameters in the model and need merely sum columns F, G, and H to obtain the simulated phenotypes (column I) for each of the sampleids.

	A	B	C	D	E	F	G	H	I	J
1	sampleid	Random sample	random genotype	fixed values	fixed genotype	μ	G_i	ε_{ij}	P	
2	P1	0.694765	tt	0.147935	gt	50	0	-10.6475	39.35253	
3	P2	0.274265	gt	0.657802	tt	50	-11.43	-3.17286	35.39714	
4	P3	0.736703	tt	0.940349	tt	50	-11.43	0.861763	39.43176	
5	P4	0.317299	gt	0.022139	gg	50	11.43	4.939074	66.36907	
6	P5	0.772776	tt	0.764216	tt	50	-11.43	-12.0852	26.48485	
7	P6	0.019678	gg	0.735858	tt	50	-11.43	-0.02329	38.54671	
8										

Keep in mind that if these were field trial data, we would only be able to obtain data found

in columns A and I. It should be immediately apparent that column F could be a mean value for a particular replication or environment of the sampleids. Thus, it should be possible to simulate data from multiple replicates and multiple environments with different mean values. It should also be apparent that the sampling of ε_{ij} could be derived from environments with a different plot-to-plot variability. For example, instead of using 7 in the function `=NORMINV(RAND(),0,7)`, we could designate the standard deviation for some environments to be 14 and thus create a type of GxE that we discussed in Chapter 12 on Multi Environment Trials.

Example Calculations

For the specific case of an F_2 generation derived from a cross of two inbred lines, p = q = 0.5,

$$a = \sqrt{\frac{49}{2(0.25)}} = 9.9$$

Alternatively, we could be interested in the case of recombinant inbred lines derived in the F_5 generation of a cross of two inbred lines. In this case, the additive portion of genetic variance is represented by Equation 5:

$$\sigma_A^2 = 2pq(1+F)[a+d(q-p)]^2$$

Equation 5 Formula for calculating additive genetic variance involving inbreeding coefficient.

where:
F = the coefficient of inbreeding,
All terms are as defined previously.

Again, let d = 0, p = q = 0.5, but F = \small 0.875 and a reasonable value to simulate for a as:

$$a = \sqrt{\frac{49}{2(1+F)pq}} \cong 7.23.$$

Polygenic Trait Simulation

Let us next simulate a polygenic trait *P* in which segregation at three loci will contribute additive genotypic values that are responsible for 30% of the phenotypic variability. In this case, the phenotype is modeled where *i*, *j*, *G*, and ε are as before, and *k* represents each locus, while n is

3. For simplicity, let us assume that segregation at each of the three loci contributes an equal amount to the genotypic variability in an F_2 population. Let us refer to these loci as quantitative trait loci (QTL). Using the quantitative genetic models we have already used, we learn that for each of the simulated QTL, ***a*** can be obtained using Equation 6:

$$a = \frac{\sqrt{\frac{\sigma_G^2}{2pq}}}{nQTL},$$

Equation 6 Formula for calculating additive genotypic value involving QTL,

where:
$< spanstyle =$" $text - align : initial; font - size : 1em$ "$>< /span >$ All terms are as defined previously.

If we want the total phenotypic variability to be ~98, as before, and the frequency of each of the alleles at all three loci is 0.5 (as in an F_2), then $\sigma^2_G = \sigma^2_A = 29$ and a = 2.56 for each of the loci. We would translate this information to the Excel spreadsheet as before, but now the spreadsheet will have three columns for genotypes and three columns for genotypic values at each of the loci.

QTL Simulations

How would you simulate genotypic effects if you wanted one of the QTLs to contribute 75%, a second QTL to contribute 20%, and the third to contribute 5% to the total genotypic variability?

For hybrid crops, the segregating progeny are often evaluated in testcross combination. For example, in maize, it is routine to generate doubled haploids (DHs) from a cross of two elite Stiff Stalk homozygous lines. The DH's are then crossed to an elite non-Stiff-Stalk homozygous 'tester'. The resulting sample of Testcrossed DH (TDH) will be evaluated in an initial field trial. Let us simulate this situation for TDHs, grown in a field trial in which the CV for yield is ~ 7% and the mean is ~ 225 bu/ac. In order to simulate TDH's we need to recall that the additive genetic variance for testcross progeny is represented by Equation 7:

$$\sigma^2_{A^T} = \frac{1}{2}(1 + F)(\alpha^T)$$

Equation 7 Formula for calculating additive genetic variance for testcross progeny,

where:

$\sigma^2_{A^T}$ = the additive genetic variance for testcross,
F = the coefficient of inbreeding,
α^T = the average effect of testcrossed allele.

Because the parents of the DH lines are fully homozygous, we can assume F=1. Thus, $\sigma^2_{A^T} = \alpha^T$
.
Otherwise, the simulations will be generated as before, except we now have a different mean and phenotypic variance.

References

Beavis, W. D.1994. The power and deceit of QTL experiments: lessons from comparative QTL studies, p.250-266. In Proceedings of the forty-ninth annual corn and sorghum industry research conference. American Seed Trade Association, Washington, DC.

Bulmer, M. G. 1976. The effect of selection on genetic variability: a simulation study. Genet. Res., Camb. (1976), 28, pp. 101-117

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226–231.

Podlich, D. W., and M. Cooper. 1998. Modelling plant breeding programs as search strategies on a complex response surface, p.171-178. In Asia-Pacific Conference on Simulated Evolution and Learning. Springer, Berlin, Heidelberg

How to cite this chapter: Beavis, W., and A. A. Mahama. 2023. Simulation Modeling. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Plant Breeding Basics

William Beavis and Anthony Assibi Mahama

The materials in this chapter cover the basics of plant breeding and data management and analysis concepts to serve as a refresher for helping the reader prepare to work through the main chapters of the book. Some readers may find it necessary to consult some introductory breeding and statistics texts for additional preparation.

Learning Objectives

Place plant breeding activities within a framework of three categories based on goals:

- Genetic improvement
- Cultivar development
- Product placement

Fig. 1 Plant breeding research activities at Makerere University in Uganda. Photo by Iowa State University.

Defining Plant Breeding

Plant Breeding has many definitions. A working definition to consider:

Plant Breeding is the genetic improvement of crop species.

This definition implies that a process (breeding) is applied to a crop, resulting in genetic changes that are valued because they confer desirable characteristics to the crop (Fig 1). Current breeding programs are the result of thousands of years of refinements that have been implemented through considerable trial and error. Refinements to the breeding processes are constrained by limited resources, technologies, and the reproductive biology of the species. Thus, the challenge of designing a plant breeding program might be thought of as the engineering counterpart to plant science (Fig. 2).

Fig. 2 Plant breeding research activities at Makerere University in Uganda. Photo by Iowa State University.

Other definitions

- Art of plant breeding: "... the ability to discern fundamental differences of importance in the plant material available and to select and increase the more desirable types..." Hayes and Immer (1942)
- "Plant breeding, broadly defined, is the art and science of improving the genetic pattern of plants in relation to their economic use." Smith (1966).
- "Plant breeding is the science, art, and business of improving plants for human benefit." Bernardo (2002).

Fig. 3 Individual plants of intermediate wheatgrass are tied into bundles to be cut and threshed in order to select the plants with the highest yield and largest seed. Photo by Dehaan; licensed under CC-BY-SA 3.0 via Wikimedia Commons.

Organization of Plant Breeding Activities

For the purposes of applying appropriate quantitative genetic models in plant breeding, it is important to understand the distinctions among three types of plant breeding projects: genetic improvement, cultivar development, and product placement. The distinctions among these three types of projects are nuanced aspects of every plant breeding program, yet the distinctions are critical for applying the correct models for data analyses used in decision-making.

Cultivar Development

The primary goal of a genetic improvement (red arrows) project is to identify lines to cycle into the breeding nursery (Fig. 3) for purposes of genetic improvement of the breeding population. Identification of lines to select is accomplished through assays of segregating lines (synthetics, hybrids) with trait-based markers and small plot field trials in single and Multi-Environment field Trials (METs) (Fig. 4). Data analyses will include analyses of binary traits with binomial and multinomial models and quantitative traits with mixed linear models, where the segregating lines will be modeled as random effects and the environments as fixed effects.

The primary goal of the cultivar development project (blue filters) is to identify cultivars that have the potential to be grown throughout a targeted population of environments. Thus, in a

cultivar development project, selected lines from segregating populations will be evaluated for quantitative traits in multi-environment trials. Data analyses in the Regional Trials of a cultivar development project will also be based on mixed linear models. However, in this case, lines are often modeled as fixed effects, while the environments are modeled as random effects.

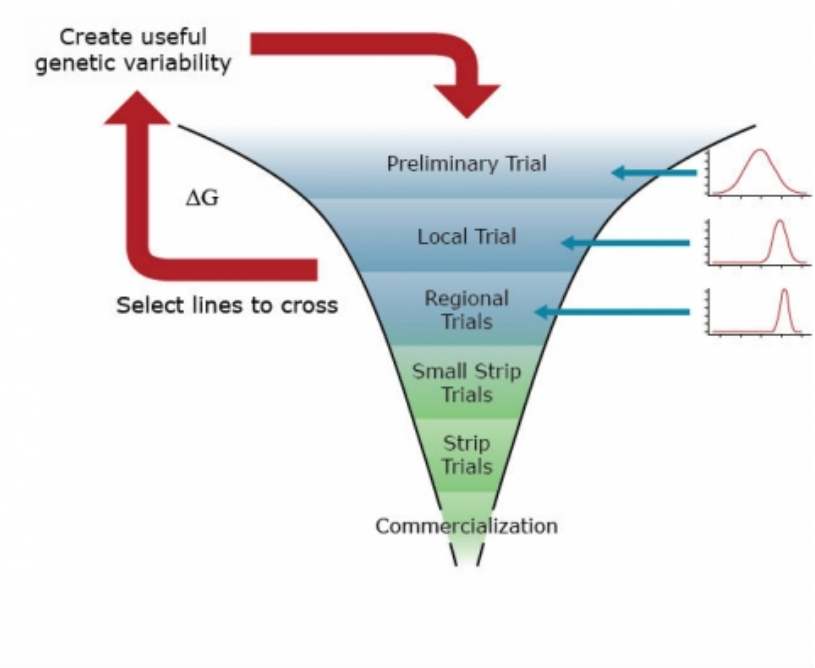


Fig. 4 A model of plant breeding activities.

The goals of a product placement project are again distinct from genetic improvement and cultivar development. In a product placement project, agronomic management practices, as well as cultivars, are selected for the field trials. These are often organized in hierarchical (split-plot) experimental designs. Thus, the parameters of a mixed linear model associated with agronomic practices and cultivars will be modeled as fixed effects, while various levels of residual variability associated with split-plot experimental units will be modeled as random effects.

For an introductory course on Quantitative Genetics, we will focus primarily on genetic improvement, a little bit on cultivar development projects, and no time will be spent on product placement projects.

Decision-Making Process

Conceptually, genetic improvement consists of a simple two-step, iterative decision-making process (Fig. 5):
1) selection of parents for crosses and 2) evaluation of their segregating progeny for the next

generation of parents and development of cultivars (Fehr, 1991). Operational implementation of genetic improvements for any given species requires far more detail.

For example, Comstock (1978) outlined the major activities involved in genetic improvement by plant breeders (below).

The details of any particular breeding program will likely consist of many activities. At the same time, it is important to categorize these activities according to the goals that transcend all plant breeding programs.

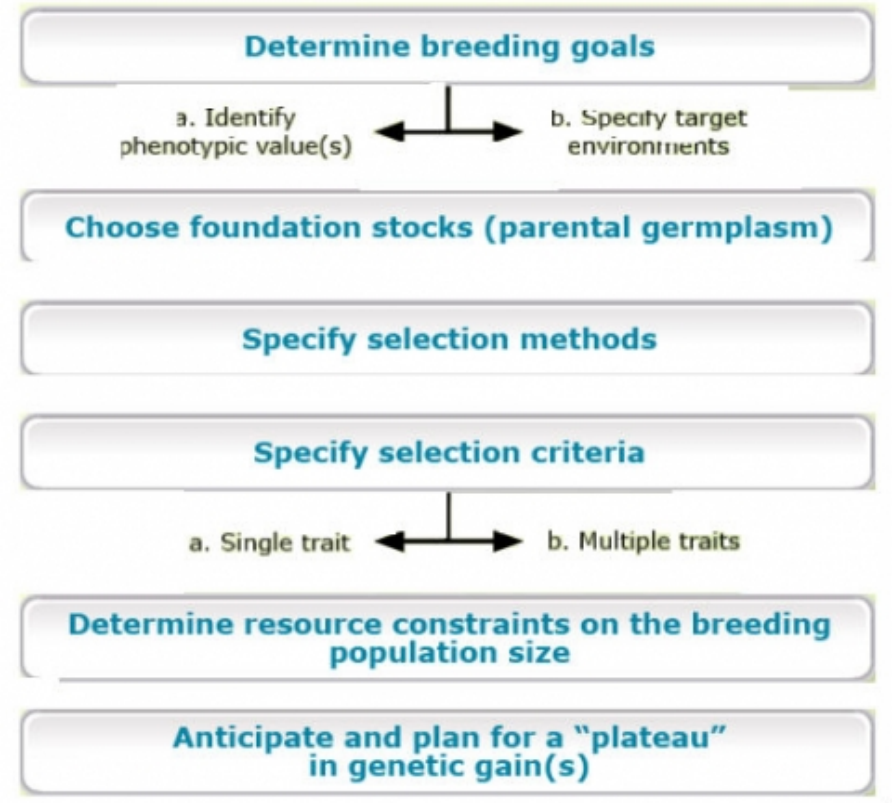


Fig. 5 Genetic Improvement Process by Plant Breeders

A Brief History of Quantitative Genetics

Quantitative genetics addresses the challenge of connecting traits measured on quantitative scales with genes that are inherited and measured as discrete units. This challenge was originally addressed through the development of theory between 1918 and 1947. The theory is now referred to as the **modern synthesis** and required another 50 years for technological innovations and experimental biologists to validate. Luminaries such as RA Fisher, Sewell Wright, JBS Haldane, and John Maynard Smith were able to develop this theory without the benefit of high throughput

'omics' technologies. Indeed, modern synthesis was developed before knowledge of the structure of DNA.

Unlike animal breeders, plant breeders implement breeding processes in organisms that cannot be protected from highly variable environments. Because plants are rooted in the sites in which they are planted, they have evolved unique adaptive mechanisms, including whole-genome duplications that enable biochemical diversity through secondary metabolism and multiple forms of reproductive biology.

Because of the reproductive and biochemical diversity in domesticated crops, plant breeders felt little need to develop quantitative genetics beyond initial concepts associated with the Analysis of Variance (**ANOVA**; Fisher, 1925; 1935). Thus, plant breeders focused their efforts on the development of field plot designs and careful plot management practices to ensure balance in field plot data for ANOVA.

Modern Synthesis Theory

The lack of reproductive and biochemical diversity in animal species created constraints that forced animal breeders to concentrate their efforts on the development of quantitative genetics beyond the ideas of Fisher (1918, 1928). JL Lush (1948) and his student CR Henderson (1975) realized that genetic improvement of quantitative traits in domesticated animal species could not take advantage of replicated field trials that are based on access to cloning, inbreeding, and the ability to produce dozens to thousands of progenies per individual. With these constraints, it was not possible to obtain precise estimates of experimental error or Genotype by Environment interaction effects using classical concepts from the ANOVA. So, they developed the statistical concepts of Mixed Model Equations (**MME**) to estimate **breeding values** of individuals with statistical properties of Best Linear Unbiased Prediction (**BLUP**). These very powerful statistical approaches were largely ignored by plant breeders until about 1995 (Bernardo, 1996).

Marker Technologies

The power of these methods is derived from knowledge of genealogical relationships. For some commercial plant breeding organizations, genealogical information had been carefully recorded for purposes of protecting germplasm. Thus, it was relatively easy for commercial breeding companies such as Pioneer and Monsanto to implement these methods. Next, international plant breeding institutes began to incorporate mixed linear models to estimate breeding values in their genetic improvement programs (Crossa et al, 2004); again, it was fairly easy to do this with extensive pedigree information. Since about 2005, many plant quantitative geneticists have published extensively on the benefits of this approach to genetic improvement of crops (Piepho,

2009), although there remain many academic plant breeding organizations that do not utilize MME to estimate breeding values with BLUP statistical properties, primarily because pedigrees of lines developed by academic programs have not been widely shared, nor aggregated into a shared repository. In parallel to the adoption of MME by plant breeders, there has been the development of relatively inexpensive genetic marker technologies. These have enabled the use of MME for Genomic Estimates of Breeding Values (**GEBV**; Meuwissen, 2001), thus overcoming the lack of genealogical knowledge for many crop and tree species.

Trait Measures

Objectives: Demonstrate ability to distinguish among the various types of phenotypic and genotypic traits that are assessed routinely in a plant breeding program.

Categorical Scales

In the context of plant breeding, quantitative genetics provides us with a genetic understanding of how quantitative traits change over generations of crossing and selection. Recall traits can be evaluated on **categorical** (Fig. 6) or **quantitative** scales. If the trait of interest is evaluated based on some quality, for example, disease resistance, flower color, or developmental phase, then it is considered a categorical trait. There are three further distinctions that can be made among categorical scales:

- **Binary** consists of only two categories such as resistant and susceptible or small and large.
- **Nominal** consists of unordered categories. For example, viral disease vectors might be categorized as insects, fungi, or bacteria.
- **Ordinal** consists of categorical data where the order is important. For example, disease symptoms might be classified as none, low, intermediate, and severe.

Fig. 6 Flower color variation in Aloe chabaudii from Manica province in Mozambique. Photo by Ton Rulkens, licensed under CC-SA 3.0 via Wikimedia Commons.

Quantitative Scales

Binary, nominal, and ordinal data are typically analyzed using Generalized Linear Models. Such models require that we model the error structures using Poisson or Negative Binomial distributions and are beyond the scope of introductory quantitative genetics. It is important to remember, however, that it is not advisable to apply General Linear Models to categorical data types.

There are two further distinctions of traits that are evaluated on quantitative scales:

- **Discrete** data occur when there are gaps between possible values. This type of data usually involves counting. Examples include flowers per plant, number of seeds per pod, number of transcripts per sample of a developing tissue, etc.
- **Continuous** data can be measured and are only limited by the precision of the measuring instrument. Examples include plant height, yield per unit of land, seed weight, seed size, protein content, etc.
    - In the context of measurement, **Precision** refers to the level of detail in the scale of the measurement.
    - Accuracy refers to whether the measurement represents the true value.

Types of Models

Objectives: Be able to place plant breeding activities within a framework of three categories based on goals:

- Genetic improvement
- Cultivar development
- Product placement

Definition and Purpose of Models

Models are representations or abstractions of reality. Some models can be very useful, e.g., prediction of phenotypes, even if they are not accurate. If data are modeled well, they can be used to generate useful graphics that will inform the breeder about data quality, integrity, and novel discoveries. Most often, predictive models are in the form of mathematical functions. Also, there are models for organizing data, analyses, processes, and systems. Yes, breeding systems and genetic processes can be represented as sets of mathematical equations. Historically the subject of optimizing a breeding system has been approached through ad hoc management activities that are often tested through trial and error. In the future, design and development of plant breeding systems will need to be treated with the same rigor that engineers use to design optimal manufacturing systems. Thus, it will be important to learn how to model breeding systems as mathematical functions.

Data Modeling

Even if it were possible to record data without error, as soon as we evaluate a trait and record the value of a living organism, we lose information about the organism. The challenge is to develop a data model that will minimize recording errors and loss of information.

What Is Data Modeling?

- Data modeling is the process of defining data requirements needed to support decisions.
- Data modeling is used to assure standard, consistent, and predictable management of data as a resource for making decisions.
- Data models support data and decision systems by providing definitions and formats. If the data are modeled consistently throughout a plant breeding program, then compatibility of data can be achieved.

If a single data structure is used to store and access data, then multiple data analyses can share data.

Steps for Modeling Data in a Plant Breeding Project

- Outline the plant breeding process.
- Determine the experimental or sampling units that will be evaluated at each step in the process.
- Determine the number of experimental or sampling units that will be evaluated.
- Characterize the experimental and sampling units as well as the traits that will be evaluated at each step in the process.

An experimental unit is defined as the basic unit to which a treatment will be applied. A sampling unit is defined as a representative of a population of interest. In quantitative genetics, we evaluate responses (traits) of experimental or sampling units on continuous scales, e.g., grain yield, plant height, harvest index, etc. Note that a measurement taken on a continuous scale is not the same as a continuously measured trait. Continuously measured traits such as grain fill, transpiration, disease progression, or gene expression are measured continuously over the growth and development of an organism. Historically, evaluation of continuously changing traits has been too labor-intensive to justify their expense. The emergence of 'phenomics' using image processing will overcome the limitations of acquiring the data. However, the need to store and manage 'big data' from phenomics is going to require novel data models and computational infrastructure, or else the acquisition of such data will be meaningless.

Organizing Data

Data models address the need to organize data for subsequent analysis.

A simple data model consists of a Row x Column matrix, where all experimental or sampling units are represented in rows, and the evaluated characteristics or attributes for each unit are recorded in the columns (Fig. 7):

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1c} \\ \vdots & \ddots & \vdots \\ a_{r1} & \cdots & a_{rc} \end{pmatrix} = \begin{pmatrix} \text{line}_1 & \cdots & \text{yield}_{\text{line}_1} \\ \vdots & \ddots & \vdots \\ \text{line}_r & \cdots & \text{yield}_{\text{line}_r} \end{pmatrix}$$

Fig. 7 A matrix, A, for representing data.

Alternatively, the $\mathbf{A}_{r \times c}$ matrix can be represented as:

- $A = \{a_{ij}\}$, for i = 1,2,3 … r and j = 1,2,3 … c
- *i* would represent line 1, line2, line3 … line r and
- *j* would represent location, replication, SNP locus 1, disease rating, … yield, etc.

Preserving Data

While the $\mathbf{A}_{(r \times c)}$ matrix is sufficient for small research projects, it is inadequate and cumbersome for breeding programs consisting of multiple types of evaluation trials at multiple stages of development. For such programs, relational databases are designed to optimize the ability to search and prepare data for analysis and interpretation using statistic and genetic models (Fig. 7). Further unless data in an $\mathbf{A}_{(r \times c)}$ matrix is disseminated through "read-only" access, there is potential for alteration of originally recorded data. Thus, the use of Excel files, too commonly used to store experimental data in an $\mathbf{A}_{(r \times c)}$ matrix, can create serious ethical issues. While such issues do not disappear with relational databases, relational databases enable more effective protection of data as originally recorded. Recently, a publicly available database designed for organizing data from plant breeding projects has been developed. Known as the Breeding Management System, it is part of the Integrated Breeding Platform designed and developed by the Generation Challenge Program of the Consultative Group of International Agricultural Research Centers.

A Relational Database

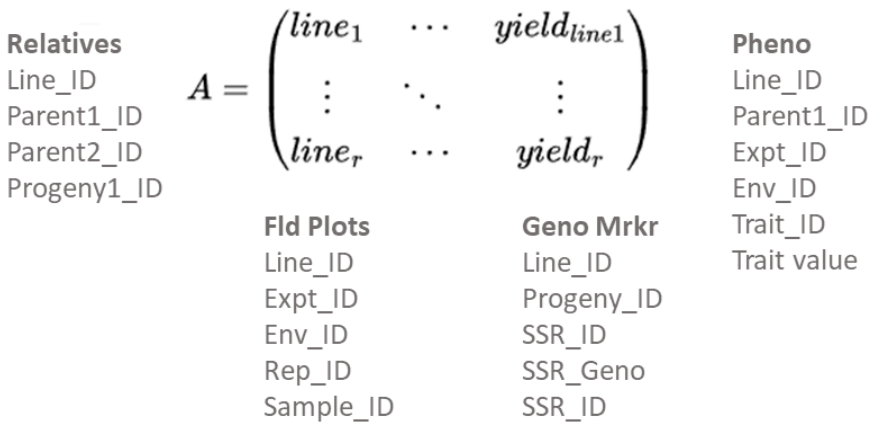


Fig. 8 A relational database for data management plant breeding research.

While the development of relational databases (Fig. 8) is outside of the scope of this course, it is important to note that plant breeders routinely work with database developers to design, implement, and populate relational databases.

Phenotypic Models

For the most part, plant breeders rely on linear models to represent measured traits. While we will concentrate on statistic and genetic models for continuous traits, it is important to recognize that there are well-developed data analysis methods for binary, nominal, and ordinal traits (see McCullagh and Nelder, 1989 or Christensen, 1997 for explanations of *Generalized* Linear Models). A general (not *Generalized*) linear model for the phenotype can be denoted as in Equation 1:

$$Y_i = \mu + e_i$$

Equation 1 General linear model for the phenotype.

where:
Y_i = phenotype of individual *i*,
μ = mean phenotype value of individual *i*,
e_i = random variability (or lack of precision) in the measurement of the phenotype of individual i.

Further, we often assume that the variability associated with each measurement of variables, e_i, is distributed as random identical and independent Normal variables. This simple model is typically

associated with the hypothesis that the only source of variability is due to chance (noise). We can extend the simple model to include genetic (G) and environmental (E) sources (signals) of variability, $Y = \mu + G + E + e$.

Two Linear Models

Scalar Notation

We will utilize two types of models to analyze data Equations 2 and 3):

$$Y_i = \beta_0 + \beta_1 G_1 + \epsilon_{ij}$$

Equation 2 Linear model for phenotype.

where:
β_0 = intercept at the y-axis,
β_1 = the slope of the regression line,
G_i = the genotypic value,
ϵ_{ij} = unspecified or residual sources of variability.

$$Y_{ij} = \mu + g_i + r_j + \epsilon_{ij}$$

Equation 3 Linear model for phenotype.

where:
Y_{ij} = the phenotypic response,
μ = the population mean,
g_i = the genotypic unit,
r_j = replicates of the genotypic units,
ϵ_{ij} = unspecified or residual sources of variability.

The parameters of Equation 2 represent the intercept and slope of a line that can be fit to data consisting of pairs of genotypic values, G_i, and phenotypic responses, where the genotypic values are continuous and known (i.e., measured without error) while the phenotypic data are measured with error in plots (experimental units). The parameters of Equation 3 represent a population mean, genotypic units, g_i, r_j replicates of the genotypic units, and the phenotypic, Y_{ij} responses. The genotypes are usually categorical designators of distinct segregating lines, hybrids, cultivars, clones, etc.

We typically estimate the parameters of Equations 2 and 3 using least squares methods. These methods are based on the idea of minimizing the squared differences between the model

parameters and the measured phenotypic value (Equation 4). For example, using Equation 2, we want to minimize the difference as:

$$min(Y_i - [\beta_0 + \beta_1 G_i])^2$$

Equation 4 Model for squared differences.

where:
terms are as described previously.

Taking the partial derivatives of Equation 3 with respect to β_0 and β_1 and setting the resulting two equations = 0, we find the slope using Equation 5:

$$\beta_1 = [V(G_i)]^{-1}[Cov(G_i Y_i)] = \frac{[Cov(G_i Y_i)]}{[V(G_i)]} \; and \; \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Equation 5 Formulae for calculating intercept and slope.

where:
VG_i = the genotypic variance,
$CovG_i Y_i$ = covariance of G and Y,
$\bar{Y}$ = mean of Y,
$\bar{X}$ = mean of X.

The result is a prediction equation (Equation 6):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$</>

Equation 6 Prediction model.

where:
$\hat{Y}, \hat{\beta}_0, \hat{\beta}_1$ = the estimates of these terms,
X_i = predictor variable, genotype, G_i.

Note that the predicted values are placed on the fitted line. Such values are sometimes referred to as 'shrunken' estimates because, relative to the observed values, they show much less variability.

If it were possible to obtain the true genotypic values, G_i, then we could routinely use Equation 2 to predict the phenotypic performance of individual *i*. Instead, plant breeders have used Equation 3 and its expanded versions to evaluate segregating lines and cultivars.

Matrix Notation

Equation 2 also can be represented by Equation 7:

$$y = X\beta + \epsilon \quad with\ solution\ as\ \ \beta = (X'X)^{-1}(X'\bar{Y})$$

Equation 7 Matrix model and solution for phenotype.

where:
y = phenotype,
X = matrix or vector of genotype,
ϵ = vector or residual or error.

And Equation 3 could be represented by Equation Equation 8:

$$y = Xr + Zg + v$$

Equation 8 Model for phenotype.

where:
y = phenotype ,
X = vector of replication,
Z = vector of genotype.

When represented this way, though, Equation 3 is usually misinterpreted by beginning students often, as the matrix form of the equation with an added set of the parameter **Z**. The matrix form of Equation 2 is actually a mixed linear model equation (Equation 9) and not a simple expansion of the matrix form of Equation 1.

$$y = X\beta + Zv + \varepsilon$$

Equation 9 Matrix model linear model for phenotype.

where:
y = vector of observations (phenotypes),
X = Incidence matrix for fixed effects,
β = vector of unknown fixed effects (to be estimated),
Z = Incidence matrix of random effects,
v = a vector of random effects (genotypic values to be predicted),
ε = a vector of residual errors (random effects to be predicted).

Exploratory Data Analysis (EDA)

Objectives

- Distinguish between descriptive and inferential statistics.
- Conduct and interpret exploratory data analyses.
- Distinguish parameters from estimators and estimates.
- Estimate means in both balanced and unbalanced data sets.
- Estimate variances, covariances, and correlations in balanced data sets.

Statistical Inference

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Statistical Inference consists of two components: estimation and hypothesis testing. In this section, we review some introductory estimation concepts.

Statistical Inference: Hypothesis Testing

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots, and simple descriptive statistics such as the range (maximum, minimum), quartiles, and the sample average, median, and mode. These exploratory data analysis (EDA) techniques should always be used prior to **estimation** and **hypothesis testing**. However, prior to conducting EDA, the phenotype should be **modeled** using the parameters in the experimental and sampling designs.

Exploratory Data Analyses

Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots, and simple descriptive statistics such as the range (maximum, minimum), quartiles, correlations, and coefficients of variation. These exploratory data analysis (EDA) techniques that provide ***descriptive statistics*** should always be used prior to ***estimation*** and ***hypothesis testing.***

Estimation: Sample Average

In population and quantitative genetics, ***parameters*** are quantities that are used to describe

central tendencies and dispersion characteristics of populations. Parameters are usually presented in the context of theoretical models used to describe quantitative and population genetics of breeding populations. Parameters of interest in population and quantitative genetics include frequencies, means, variances, and covariances.

Because populations often consist of an infinite or very large number of members, it may be impossible to determine these quantities. Instead, statistical inferences, i.e., ***estimates***, about the true but unknowable parameters are determined from samples. The rule by which a statistical estimate of a parameter is constructed is known as the ***estimator***. For example, the description of how to calculate***a sample average*** given by Equation 10:

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

Equation 10 Formula for estimating sample average.

where:
X_i = the sample, ***i***,
n = number of samples.

This average represents an estimator of the population mean, while the calculated value, e.g., 132.38, obtained from 25 (n) samples (X_i) from a population would be an estimate of the population average.

Estimation of Means

The most common inferential statistic is the estimate of a mean. Computing arithmetic means, either simple or weighted within-group averages represents a common approach to summarizing and comparing groups. Data from most agronomic experiments include multiple treatments (or samples) and sources of variability. Further, the numbers of observations per treatment often are not equal; even if designed for balance, some observations are lost during the course of an experiment. Thus, most data sets come from experiments that have multiple effects of interest and are not balanced. In such situations, the arithmetic mean for a group may not accurately reflect the "typical" response for that group because the arithmetic mean may be biased by unequal weighting among multiple sources of variability. The calculation of Least Square Means, ***lsmeans***, was developed for such situations. In effect, *lsmeans* are within-group means appropriately adjusted for the other sources of variability. The adjustments made by *lsmeans* are meant to provide estimates as though the data were obtained from a balanced design. When an experiment is balanced, arithmetic averages and *lsmeans* agree.

Estimation of Means: Example

Consider a data set consisting of 3 cultivars evaluated in a Randomized Complete Block Design consisting of 5 replicates at each of 3 locations (Table 1). Despite exercising best agronomic practices, note that some plots at some locations did not produce phenotypic values.

The estimated means and number of observations for each cultivar indicate that there is very little difference among the cultivars, although cultivar C appears to have the highest yield (Table 2).

A closer investigation of the data reveals that the means are unequally weighted by location effects. Recalculating the means for the cultivars indicates more distinctive differences among the cultivars once the differences among environments were taken into account (Table 3).

Table 1 Sample data with missing values.

Cultivar	Location	$Y_{j,k}$
A	Ames	17, 28, 19, 21, 19
A	Sutherland	43, 30, 39, 44, 44
A	Castana	-, -, 16, -, –
B	Ames	21,21, -, 24, 25
B	Sutherland	39, 45, 42, 47, –
B	Castana	-, 19, 22, -, 16
C	Ames	22, 30, -, 33, 31
C	Sutherland	46, -, -, -, –
C	Castana	25, 31, 25, 33, 29

Table 2 Trait average values for three cultivars, with the same sample number.

Cultivar	N	Average
A	11	29.1
B	11	29.2
C	11	30.2

Table 3 Unequally weighted trait average values.

Cultivar	lsmeans
A	25.6
B	28.3
C	34.4

Estimation of Variances

If we model a trait value as $Y_i = \mu + e_i$ (Equation 1), then the estimator of the variance of the population consisting of individuals, i = 1, 2, 3, …. , N is written as in Equation 11:

$$\sigma_y^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

Equation 11 Formula for calculating the variance of a population of individuals,

where:
σ_y^2 = the variance of the population of trait y values,
N = population size ,
μ = population mean.

Since it is not possible to evaluate a population of a crop species (think about it), we usually take a sample of individuals representing the population, i = 1, 2, 3, …, n, where n << N. The estimator of the sample variance from a sample of n values is represented in Equation 12:

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1}$$

Equation 12 Formula for estimating sample variance,

where:
s^2 = the sample variance,
Y_i = trait value of individual *i*,
$\bar{Y}$ = trait mean value.

Estimation of Covariance

The covariance is a measure of the joint variation between two variables. Let us designate one trait X and a second trait Y. We can model Y as before and we can model X in a similar manner, i.e.,

$$X_i = \mu + e_i$$

Equation 13 Formula for estimating trait value,

where:
X_i = i^{th} value of trait X,
μ = mean of trait X,
e_i = random variability in trait values.

and the estimator of the variance of X is obtained using Equation 14:

$$\sigma_x^2 = \frac{\sum (X_i - \mu)^2}{N}$$

Equation 14 Formula for estimating variance of the trait, X.

where:
X_i = i^{th} value of trait X,
μ = mean of trait X,
e_i = random variability in trait values.

Thus, the estimator of the covariance of X and Y is as in Equation 15:

$$\sigma_{X,Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)$$

Equation 15 Formula for estimating covariance of the traits, X and Y.

where:
$\sigma_{X,Y}$ = covariance of X and Y,
μ_X = mean of population of Xs,
μ_X = mean of population of Ys,
other terms are as defined previously.

Again, it is not possible to evaluate a population, so we usually take a sample of individuals

representing the population, i = 1,2,3 … n, where n << N. So, the estimator of a sample covariance is obtained using Equation 16:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Equation 16 Formula for calculating covariance of the traits, X and Y, from a sample.

where:
$Cov(X,Y)$ = the covariance of X and Y,
$\bar{X}$ = mean of sample of Xs,
$\bar{Y}$ = mean of sample of Ys.

Estimation of Variance Components

If we extend our simple model to include genetic and environmental sources of variability, as mentioned previously, as $y = \mu + G + E + e$, then, noting that **μ** is a constant and applying some algebra, we can show that the variance (V) of Y is as in Equation 17:

$$V(Y) = V(G) + V(E) + 2Cov(G,E) + V(e)$$

Equation 17 Formula for estimating variance components.

where:
$V(Y)$ = the variance of trait Y,
$V(G), V(E), V(e)$ = the genotypic, environmental, and random variability, respectively,
$Cov(G,E)$ = the covariance of G and E.

The assumption is that the errors are independently distributed. If we further assume that genotype and environment are independent and that there is no genotype x environment interaction, the variance and variance components are estimate with Equation 18:

$$V(Y) = V(G) + V(E) + V(e)$$

Equation 18 Formula for estimating variance components, in the absence of covariance.

where:
$terms$ are as defined previously.

Questions to Consider

A question to consider is whether the parameters of the linear model $Y = \mu + G + E + e$ represent fixed or random effects, because this determination will affect the way in which we estimate variance components and whether each is contributing significantly to the overall phenotypic variability. This determination depends on the inference space to which results are going to be applied. Fixed effects denote components of the linear model with levels that are deliberately arranged by the experimenter, rather than randomly sampled from a population of possible levels. Inferences in fixed effect models are restricted to the set of conditions that the experimenter has chosen, whereas random effect models provide inferences for a population from which a sample is drawn.

As a practical matter, it is hard to justify designating a parameter as a random effect if the parameter space is not sampled well. Consider environments, for example, since we cannot control the weather, it is tempting to designate environments as random effects, however drawing inferences to a targeted population of environments will be difficult if we sample a small number of environments, say less than 40. Thus, as a practical matter, the genetic improvement component of a breeding program will consider environments as fixed effects (or nuisance parameters), because our main interest is in drawing inferences about the members of a breeding population and their interactions with environments, wheras the product placement component of a breeding program will evaluate a relatively small number of selected genotypes in a large number of environments. Thus, for this phase the models will consider cultivars (lines, hybrids, synthetics, etc.) as fixed effects and environments as random effects.

Mixed Models

Because the inference space of interest for genetic improvement is derived from random samples of genotypes obtained from a conceptually large breeding population, we do not consider genotypes as fixed effects until the genotypes have been selected for a cultivar development program. At the same time it is a rare experimental design that does not include a fixed effect. Often random effects, such as environments are classified as fixed effects in ***mixed models*** so that inferred predictions are determined using computational methods that provide restricted ***maximum likelihood*** methods. More on this topic can be found in the section on **Statistical Inference**.

Installation of R

Introduction and Objectives – Installation of R

- Learn to download and install R and R Studio.
- Learn to start an R analysis project.
- Learn how to upload data that is CSV formatted.

Background

R is a powerful language and environment for statistical computing and creating graphics. The main advantages of R are the fact that R is a free software and that there is a lot of help available. It is quite similar to other programming tools such as SAS (not freeware), but more user-friendly than programming languages such as C++ or FORTRAN. You can use R as it is, but for educational purposes we prefer to use R in combination with the RStudio interface (also free software), which has an organized layout and several extra options.

Directory Of R Commands Used

- getwd()
- setwd()
- ?
- help.search()
- example()
- read.csv()
- rm()
- rm(list=())
- head()
- hist()
- attach()
- boxplot()
- str()
- as.factor()
- aov()
- summary()

References

Up and Running with R (Internet resource)

Exercise

Imagine that you've been recently hired as a data analyst for a brand new seed company and have been asked by your supervisor to conduct an analysis of variance (ANOVA) on yield trial data from 3 synthetic maize populations planted in 3 reps each. Your company does not have funds to purchase commercial statistical software, thus you must either do the analysis by hand or use freely available software. Since you will have to analyze much larger data sets in the near future, you opt to learn how to carry out the ANOVA using the freely available software R and R-Studio.

Install R

To install R on your computer, go to the home website of R and do the following (assuming you work on a Windows computer):

- Click CRAN under Download,Packages in the left bar
- Choose a download site close to you (eg: USA: Iowa State University, Ames, IA)
- Choose Download R for Windows
- Click Base
- Choose Download R 3.1.1 for Windows and choose default answers for all questions (click "next" for all questions)

Install RStudio

After finishing above setup, you should see an icon on your desktop. Clicking on this would start up the standard interface. We recommend, however, using the RStudio interface. To install RStudio, go to the RStudio homepage and do the following (assuming you work on a windows computer):

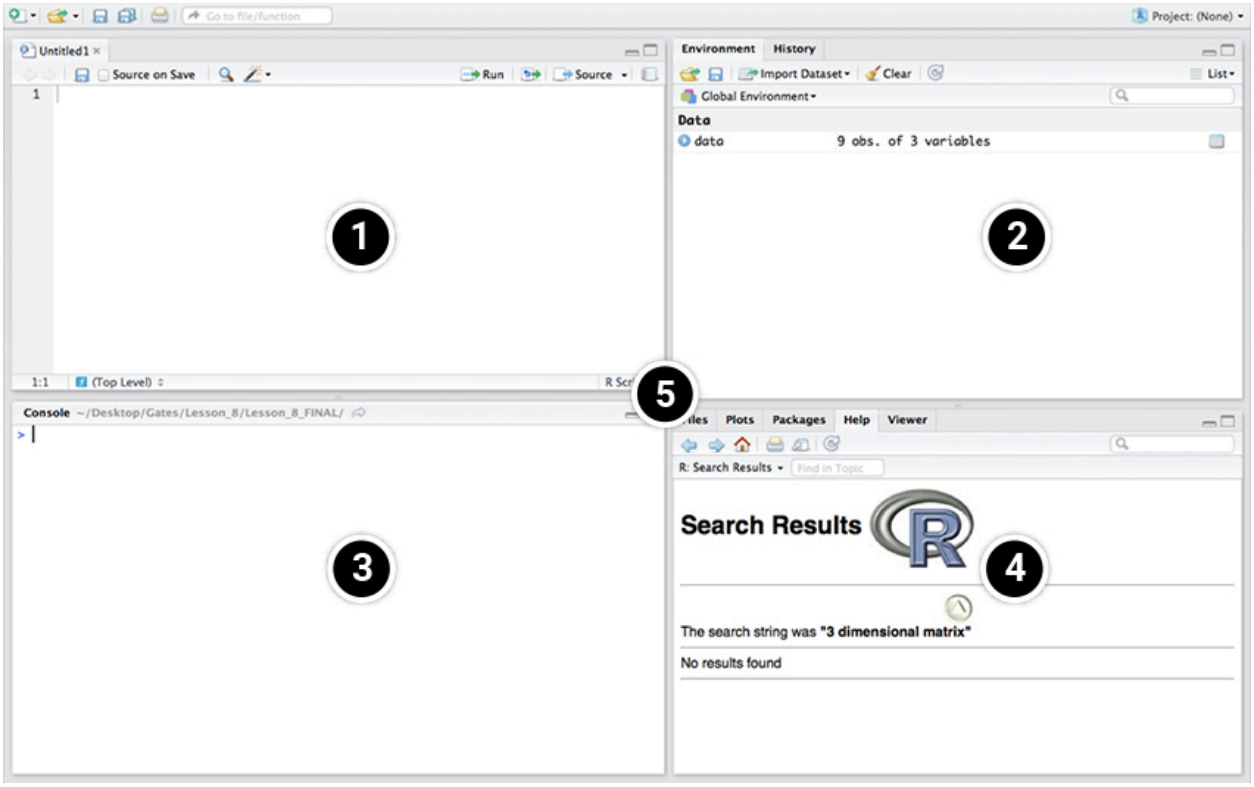
- Click Download RStudio
- Click Desktop
- Click RStudio 0.98.977 – Windows XP/Vista/7/8 under Installers for ALL Platforms to initiate download
- Open the .exe file from your computer's downloads and run it and choose default answers for all questions (click "next" for all questions)

RStudio Layout

1. **Script Window**: In this window, collections of commands (scripts) can be edited and saved. If this window is not present upon opening RStudio, you can open it by clicking File→New

File→R Script. Just typing a command in the Script window and clicking enter will not cause R to run the command; the command has to get entered in the Console window before R executes the command. If you want to run a line from the script window, you can click Run on the toolbar or press CTRL+ENTER to enter the line into the console view.

2. **Environment / History Window**: Under the Environment tab you can see which data and values R has in its memory. The History tab shows what has been entered into the console.
3. **Console window**: Here you can type simple commands after the > prompt and R will then execute your command. This is the most important window, because this is where R actually runs commands.
4. **Files / Plots / Packages / Help**: Here you can open files, view plots (also previous plots), install and load packages or use the help function.
5. You can change the size of each of the windows by dragging the grey bars between the windows.



Working Directory

Your working directory is a folder on your computer from where files can be entered, or read, into R. When you ask R to open a file with a read command, R will look in the working directory folder for the specified file. When you tell R to save a data set or figure which you've created, R will also save the data or figure as a file in the same working directory folder.

Set your working directory to a folder where all of the example data files for this lesson are located.

1. Create a folder on your desktop; for this example the folder will be called `wd`. Then, obtain the default working directory by entering the command `getwd()` into the console window. R returns the default working directory below.

```
> getwd()
```

```
> [1] "C:/Users/<Username>/Documents"
```

2. Next, set the working directory to the folder on your desktop, `wd`, using the `setwd()` command in the Console window:

```
> setwd("C:/Users/<Username>/Desktop/wd")
```

Notice that to set our working directory to a folder on our desktop, we enter everything that was returned by R from the `getwd()` command before the word Documents, change Documents to Desktop, then add a forward slash followed by the name of our folder (`wd`).

Make sure that the slashes are forward slashes and that you don't forget the quotation marks. R is case sensitive, so make sure you write capitals where necessary. Within the RStudio interface you can also go to **Session → Set working directory** to select a folder to be your working directory.

Libraries

R can do many kinds of statistical and data analyses. The analyses methods are organized in so-called packages. With the standard installation, most common packages are installed. To get a list of all installed packages, go to the packages window (lower right in RStudio). If the box in front of the package name is ticked, the package is loaded (activated) and can be used. You can also type library() in the console window to view the loaded packages.

There are many more packages available on the R-website. If you want to install and use a package (for example, the package called "geometry") you should:

1. Install the package: click on the "packages" tab at the top of the lower-right window in RStudio. Click "install", and in the text box under the heading "packages", type "geometry". You can also simply enter install.packages("geometry") in the console window to install the package.
2. Load the package: under the "packages" tab at the top of the lower-right window in

RStudio, check the boxes of the packages you wish to load (i.e. "geometry"). You can also simply type library("geometry") in the console window to load the package.

Getting Help in R

If you know the name of the function you want help with, you can just type a question mark followed by the name of the function in the console window. For example, to get help on **aov**, just enter:

```
> ?aov
```

Sometimes you don't know the exact name of a function, but you know the subject on which you want help (i.e., Analysis of Variance). The simplest way to get help in R is to click the "Help" tab on the toolbar at the top of the bottom-right window in RStudio, then enter the subject or function that you want help within the search box at the right. This will return a list of help pages pertaining to your query.

Another way to obtain the same list of help pages is by entering the help.search command in the Console. The subject or function which you'd like information about is put inside of brackets and quotation marks, directly following the help.search command. For example, to obtain information about Analysis of Variance, enter into the console:

```
> help.search("Analysis of Variance")
```

If you'd like to see an example of how a function is used, enter "example" followed by the function that you'd like to see an example of (within quotation marks and brackets). For instance, if we wanted to see an example of how the aov function can be used, we can enter into the console:

```
> example("aov")
```

An example is returned in the console window.

Reading the CSV File

Now, we want to read the CSV file from our working directory into RStudio. At this point, we learn an important operator: `<-`. This operator is used to name data that is being read into the R data frame. The name you give to the file goes on the left side of this operator, while the command read.csv goes to its right. The name of the CSV file from your working directory is entered in the parenthesis and within quotations after the read.csv command. The command **header** = **T** is used in the function to tell R that the first row of the data file contains column names, and not data.

Read the file into R by entering into the **Console:**

data <- read.csv("Review Models Install R ALA data.csv", header = T)

Tip: If you are working out of the **Console** and received an error message because you typed something incorrectly, just press the ↑ key to bring up the line which you previously entered. You can then make corrections on the line of code without having to retype the entire line in the console window again. This can be an extremely useful and time saving tool when learning to use a new function. Try it out.

If the data was successfully read into R, you will see the name that you assigned the data in the **Workspace/History** window (top-right).

Examining the Data

Let us look at the first few rows of the data. We can do this by entering the command `head(data)` in the console. If we want to look at a specific number of rows, let us say just the first 3 rows, we can enter `head(data, n=3)` in the Console. Try both ways.

First, enter into the console:

```
> head(data)
  Pop  Rep  Yield
1   30    1  137.1
2   30    2  124.4
3   30    3  145.9
4   40    1  166.1
5   40    2  147.4
6   40    3  142.7
```

Now, try looking at only the first 3 rows:

```
> head(data)
  Pop  Rep  Yield
1   30    1  137.1
2   30    2  124.4
```

```
  3   30    3  145.9
```

Viewing and Removing Datasets

Now, let us say we are finished using this dataset and want to remove it from the R data frame. To accomplish this, we can use the rm command followed by the name of what we want removed in parenthesis. Let us remove the data from the R data frame. Enter into the console `rm(data)`.

```
rm(data)
```

The dataset data should no longer be present in the Workspace/History window.

What if we have many things entered in the R data frame and want to remove them all? There are two ways that we can do this. To demonstrate how, let us first enter 3 variables (x,y, and z) into the R data frame. Set x equal to 1, y equal to 2, and z equal to 3.

```
x<-1
```

```
y<-2
```

```
z<-3
```

Clicking on 'clear' in the History/Environment window (top right) will clear everything in the R data frame. Another way to remove all data from the R data frame is to enter in the console:

```
rm(list=ls())
```

Try both ways.

EDA with R

Objectives

- Students will conduct exploratory data analyses (EDA) on data from a simple Completely Randomized Design (CRD).
- Assess whether students know how to interpret results from EDA.
- Students will conduct an Analysis of Variance (ANOVA) on data from a simple CRD.

Directory Of R Commands Used

- `getwd()`

- `setwd()`
- `read.csv()`
- `rm()`
- `rm(list=())`
- `hist()`
- `attach()`
- `boxplot()`
- `str()`
- `as.factor()`
- `aov()`
- `summary()`

Set Working Directory

Before you can conduct any analysis on data from a text file or spreadsheet, you must first enter, or read, the data file into the R data frame. For this activity, our data is in the form of an Excel comma separated values (or CSV) file; a commonly used file type for inputting and exporting data from R.

Make sure that the data file for this exercise is in the working directory folder on your desktop.

Note: We previously discussed how to set the working directory to a folder named on your desktop. For this activity, we will repeat the steps of setting the working directory to reinforce the concept.

In the **Console** window, enter `getwd()`. R will return the current working directory below the command you entered:

```
getwd()
```
[1] "C:/Users/<Username>/Documents"

Set the working directory to the folder on your desktop by entering `setwd()`. For a folder named 'wd' on our desktop, we enter:

```
setwd("C:/Users/<Username>/Desktop/wd")
```

Please note that the working directory can be in any other folder as well, but the data file has to be in that specific folder.

Reading the CSV File

Now, we want to read the CSV file from our working directory into RStudio. At this point, we learn an important operator: `<-`. This operator is used to name data that is being read into the R data frame. The name you give to the file goes on the left side of this operator, while the command read.csv goes to its right. The name of the CSV file from your working directory, in this case CRD.1.data.csv, is entered in the parenthesis and within quotations after the read.csv command. The command **header** = **T** is used in the function to tell R that the first row of the data file contains column names, and not data.

```
data <- read.csv("CRD.1.data.csv", header = T)
```

Tip: If you are working out of the **Console** and received an error message because you typed something incorrectly, just press the ↑ key to bring up the line which you previously entered. You can then make corrections on the line of code without having to retype the entire line in the console window again. This can be an extremely useful and time saving tool when learning to use a new function. Try it out.

If the data was successfully read into R, you will see the name that you assigned the data in the **Workspace/History** window (top-right).

Exploring the Data

Let us do some preliminary exploring of the data.

Read the data set into the R data frame.

```
>  data <- read.csv("CRD.1.data.csv", header = T)
```

First, let us look at a histogram of the yield data to see if they follow a normal distribution. We can accomplish this using the hist command.

Enter into the console:

```
>  hist(data$Yield, col="blue", main= "Histogram of Yield of 3 Synthetic Maize Populations",
   xlab="Yield (bushels/acre)", ylab="Frequency")
```

R returns the histogram in the Files/Plots/Packages/Help window (bottom-right).

Histogram

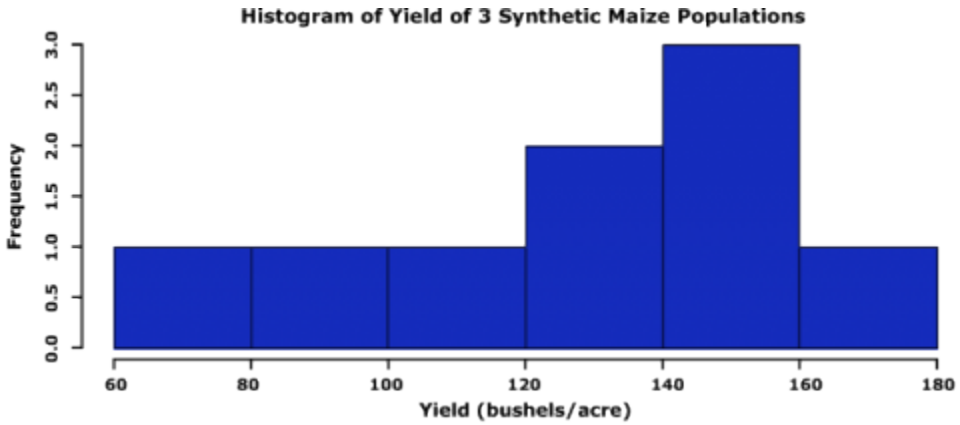


Fig. 8

Let us go through the command we just entered: `data$Yield` specifies that we want to plot the values from the column Yield in the data, `col="blue"` indicates which color the histogram should be, the entry in quotations after `main=` indicates the title that you'd like to give the histogram, the entries after `xlab=` and `ylab=` indicate how the x and y axes of the histogram should be labeled. The histogram appears in the bottom-right window in RStudio.

The histogram can be saved to your current working directory by clicking 'export' on the toolbar at the top of the lower-right window, then clicking "save plot as PNG" or "save plot as a PDF". You may then select the size dimensions you would like applied to the saved histogram.

Boxplots

Let us now look at some boxplots of yield by population for this data. First, enter into the **Console** window `attach(data)`. The **attach** command specifies to R which data set we want to work with, and simplifies some of the coding by allowing us just to use the names of columns in the data, i.e. `Yield` vs. `data$Yield`. After we enter the **attach** command, we'll enter the **boxplot** command.

```
> attach(data)
```

```
> boxplot(Yield~Pop, col="red", main="Yield by Population", xlab="Synthetic Population",
  ylab="Yield")
```

R returns the boxplot in the bottom-right window.

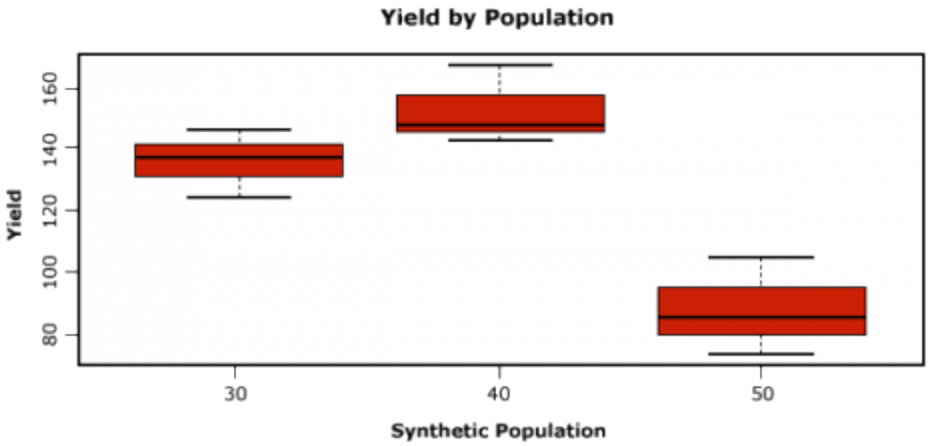


Fig. 9

Let us go through the boxplot command: `Yield~Pop` indicates that we want boxplots of the yield data for each of the 3 populations in our data, `col=` indicates the color that we want our boxplots to be, `main=` indicates the title we want to give the boxplots, and `xlab=` and `ylab=` indicate what we want the x and y axes labeled as.

Note: Yield is capitalized in our data file, thus it MUST also be capitalized in the **boxplot** command.

Mean and Coefficient of Variance

The coefficient of variance can be calculated for each population in the data set. Looking at the data, we can see that lines 1 to 3 pertain to population 30. We know that the coefficient of variation for a sample is the mean of the sample divided by the standard deviation of the sample. By using the command `mean()`, we can calculate the mean for a sample. Remember that to specify a column from a data frame, we use the `$` operator. If we want to calculate the mean of population 30 from the data (rows 1 to 3 in the data), we can enter

```
> mean(data$Yield[1:3])
```

To calculate the standard deviation of the yield for population 30, enter

```
> sd(data$Yield[1:3])
```

The coefficient of variance is therefore calculated by entering

```
> mean(data$Yield[1:3])/sd(data$Yield[1:3])
```

One-Factor ANOVA of a CRD

Now that we've gained some intuition about how the data behave, let us carry out an ANOVA with one factor (Pop) on the data. We first need to specify to R that we want Population to be a factor. Enter into the **Console**

```
> Pop<-as.factor(Pop)
```

Let us go through the command above: `as.factor(data$Pop)` specifies that we want the `Pop` column in dataset `data` to be a factor, which we've called `Pop`.

Now that we have population as a factor, we're ready to conduct the ANOVA. The model that we are using for this one-factor ANOVA is Yield=Population.

In the **Console**, enter

```
> mean(data$Yield[1:3])/sd(data$Yield[1:3])
```

Interpret the Results

Let us look at the ANOVA table. Enter out in the **Console** window.

```
> out
```

In this ANOVA table, the error row is labelled *Residuals*. In the second and subsequent columns you see the degrees of freedom for *Pop* and *Residuals* (2 and 6), the treatment and error sums of squares (6440 and 1011), the treatment mean square of 3220, the error variance = 169, the F ratio and the P value (19.1 and 0.0025). The double asterisks (**) next to the P value indicate that the difference between the yield means of the three populations is significant at 0.1% (i.e. we reject the null hypothesis that the yield means of each population are statistically equivalent). Notice that R does not print the bottom row of the ANOVA table showing the total sum of squares and total degrees of freedom.

Hypothesis Tests

Objectives

Demonstrate ability to interpret types of errors that can be made from testing various kinds of hypotheses.

Statistical Inference

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots and simple descriptive statistics such as the range (maximum, minimum), quartiles, and the sample average, median, mode. These exploratory data analysis (EDA) techniques should always be used prior to **estimation** and **hypothesis testing**. However, prior to conducting EDA, the phenotype should be **modeled** using the parameters in the experimental and sampling designs.

Null and Alternative Hypotheses

Hypotheses are questions about parameters in models. For example, "Is the average value for a trait different than zero?" is a question about whether the parameter μ is non-zero. Formally, the proposition $H_0 : \mu = 0$, is called the null hypothesis, while a proposition $H_\alpha : \mu \neq 0$ is called an alternative hypothesis.

A test statistic is used to quantify the plausibility of the data if the null hypothesis is true. For this simple hypothesis the value of the test statistic should be close to zero if the null hypothesis is true and far from zero if the alternative hypothesis is true. Notice that in all linear models there is a parameter, ε, included to indicate that there is some random variability in the data that cannot be ascribed to the other parameters in the model. It is entirely possible that the variability in the data is due entirely to ε and that an estimate of μ that is not zero is due to this random variability.

Inferential Errors from Hypothesis Testing

How often will the estimate of μ be different from zero when H_o is true? We can answer this question by rerunning an experiment in which we know μ = 0 a million times, generate a histogram of the resulting distribution and then see how often (relative to 1 million) an estimated mean that is equal to or more extreme than our experimental estimate occurs. This is the frequency associated with finding our estimated value or a more extreme value when H_o is true.

The good news is that we don't have to conduct a million such experiments because someone else has already determined the distribution when μ = 0, is true. The frequency value associated with a test statistic as extreme or more extreme than the one observed is often referred to as a 'p' value. The smaller the p value, the more comfortable we should be in rejecting the null hypothesis in favor of an alternative hypothesis. Keep in mind that we can be wrong with making such a decision. In fact we are admitting that such a decision will be incorrect at a frequency of p.

Error Types

Consider another simple example where we hypothesize that two genotypes have the same mean for some trait of interest. The difference between two genotypes is tested by Equation 19:

$$\delta_{ij} = g_i - g_j$$

Equation 19 Formula for testing the difference between two genotypes.

where:
$g_i \ and \ g_j$ = the i^{th} and j^{th} true genotypic effects on the trait of interest.

Whether or not a decision based on observed data is correct depends on the true value of the difference between the means (Table 4).

Table 4 Possible outcomes in testing the hypothesis that $\delta_{ij} = 0$. Columns indicate the three possible true states. Rows indicate the three possible decisions made on the basis of estimates from measured data.

Decision based on empirical data	True Situation		
	$\delta_{ij} < 0$	$\delta_{ij} = 0$	$\delta_{ij} > 0$
1. $\delta_{ij} < 0$	Correct decision	Type I error	Type III error
2. $\delta_{ij} = 0$	Type II error	Correct decision	Type II error
3. $\delta_{ij} >$	Type III error	Type I error	Correct decision

A Type I error is committed if the null hypothesis is rejected when it is true ($\delta_{ij}=0$ and the hypothesis of equality is rejected). A Type II error is committed if the null hypothesis is accepted when it is really false ($\delta_{ij}\neq 0$ and the hypothesis of equality is not rejected). Type I error is also called "false positive", and Type II error is also known as a "false negative." A Type III error occurs if the first decision is made when the third decision should have been made. This error also occurs if the third decision was made when the first decision was correct. Type III errors are sometimes called reverse decisions.

Significance Levels

The probability (or frequency) of a Type I error is the level of significance, denoted by α.
The choice of α can be any desirable value between 0 and 1.
For example, if a test is carried out at the 5% level, α is 0.05.

If you carry out tests at 5% level you will reject 5% of the hypotheses you test when they are really

true.
The rejection rate can be reduced by choosing a lower level of α
However, the choice of α will affect the frequency of Type II and Type III errors.

A Type III error rate, γ, is the frequency of incorrect reverse decisions and is always less than α/2 even for the smallest magnitudes of the standardized true difference, δ_{ij}/σ_d where σ_d is the parameter value of the standard error of the mean difference. Representative values of γ are shown below in Table 5.

Table 5 Type III error rates, y, when the df associated with a t-test is 40.

Standardized true difference δ_{ij}/σ_d	**Significance Level (α)**			
	0.05	0.10	0.20	0.40
0.3	0.0127	0.0271	0.0584	0.1283
0.9	0.002	00068	0.0167	0.0438
1.5	0.0005	0.0014	0.0039	0.019
2.1	0.0001	0.0002	0.0008	0.0026
2.7	0.0000	0.0000	0.0001	0.0005

Power of the Test

Lastly, consider the error that is committed if the null hypothesis is not rejected when it is truly false. This is also known as a Type II error, and the probability of this type of error is denoted by β. It is the frequency of failure to detect real differences and is also affected by both the choice of α and the magnitude of the standardized true difference (Table 6).

Table 6 Type II error rates, ß, or the frequencies of failure to detect differences when the test of significance is based on 40 df.

Standardized true difference δ_{ij}/σ_d	**Significance Level α**			
	0.05	0.10	0.20	0.40
0.3	0.941	0.886	0.781	0.579
0.9	0.863	0.774	0.639	0.437
1.5	0.697	0.571	0.419	0.248
2.1	0.469	0.340	0.214	0.107
2.7	0.251	0.158	0.085	0.035

Notice that $\alpha + \beta \neq 1.0$. The power of the test is = 1 – β and is denoted π, thus $\beta + \pi = 1.0$. The power of a test is the probability of rejecting the null hypothesis when it is false. It can be increased

by decreasing either the value of α or decreasing the value of σ_d by increasing the number of replications per treatment or by improving the experimental design.

Analysis of Variance

Objectives

Students will demonstrate the ability to conduct and interpret Analysis of Variance.

Statistical Inference

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots, and simple descriptive statistics such as the range (maximum, minimum), quartiles, and the sample average, median, mode. These exploratory data analysis (EDA) techniques should always be used prior to **estimation** and **hypothesis testing**. However, prior to conducting EDA, the phenotype should be **modeled** using the parameters in the experimental and sampling designs.

Background

The AOV has been the primary tool for testing hypotheses about parameters in linear models. The AOV was originally developed and introduced for analyses of quantitative genetic questions by R.A. Fisher (1925). Since its introduction, the assumptions underlying the AOV have guided development of sophisticated experimental designs, and with increasing computational capabilities the AOV has evolved to provide estimates of variance components from these designs. While the breadth and depth of experimental design and analyses of linear models are beyond the scope of this class, it is worth recalling the salient features of experimental design and their impact on inferences from the AOV.

Experimental Designs consist of **design structures**, **treatment structures**, and allocation of these structures to **experimental units**. Typical design structures utilized by plant breeders include Randomized Complete Block, Lattice Incomplete Block and Augmented Designs. The primary treatment designs of interest of plant breeders involve allocation of genotypes to experimental units. This is accomplished primarily through mating, although with the emergence of biotechnologies, such as protoplast fusion, tissue culture and various transgenic technologies, there are many ways to allocate treatments (genotypes) to experimental units. Would you consider treatments from these technologies as fixed or random effects? Why? Experimental units can be split in both time and space, resulting in the ability to apply treatment and design structures to different sized experimental units.

Design Principles

Design principles in allocation of treatment and design structures to experimental units include **Randomization**, **Replication** and **Blocking**. These are principles rather than rigid rules. As such, they provide flexibility in designing experiments to draw inferences about biological questions. Assuming that these principles are applied appropriately, experimental data can be used for obtaining unbiased estimates of treatment effects, variances, covariances, and even predict breeding values.

Completely Random Design

Let us imagine that we have two plant introduction accessions. We wish to evaluate whether these two accessions are unique with respect to yield.

Assume that we have 10 plots available for purposes of testing the null hypothesis that there is no difference in their yield. Also, assume that we have enough seed to plant 200 seeds in each plot.

Let us next assume that the 10 plots consist of two-row plots that are arranged in a 5×2 grid consisting of five ranges with 2 plots per range. We can randomly assign seed from each accession to the 10 plots. This would represent a Completely Random Design (CRD). Can you explain why?

Fixed and Random Effects

Prior to execution of the experiment, we want to model the phenotypic data using a linear function. In this case we would model the phenotypic data using Equation 20:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Equation 20 Linear model for phenotype.

where:
Y_{ij} = the yield of plot i, j, where i = 1, 2 for accession and j = 1, ..., 5 for replicate,
μ_i = represents the mean of accession i,
ε_{ij} = the error, ~ i.i.d. N(0, σ^2).

It is important to get in the habit of recognizing whether the parameters of the model are considered random or fixed effects.

In this first model, since we selected the two accessions, rather than sampled them from some population, we should consider them to be fixed effects. The parameter $\varepsilon_{i,j}$ representing the residual or error in the model is based on a sample of plots to which experimental units are assigned, so $\varepsilon_{i,j}$ is considered a random effect.

AOV Based on Yield

Next, let us say that we evaluate the plots for yield (bushels per acre) as well as stand counts (plants per plot) at the time of harvest. The resulting data might look something like in Table7.

Table 7 Table 1 Yield (t per ha) as well as stand counts (plants per plot).

n/a	PI accession 1		PI accession 2	
Block	(t/ha)	(plants/plot)	(t/ha)	(plants/plot)
1	1.69	91	1.88	102
2	1.95	122	1.82	89
3	2.20	143	2.01	139
4	2.13	145	2.01	147
5	1.76	110	1.95	112

If we conduct an AOV based on yield using the model for a CRD, we will generate a table that looks something like Table 8.

Table 8 ANOVA outline.

Source	df	MS	F	Prob
Accession	1	n/a	n/a	n/a
Residual	8	n/a	n/a	n/a

Blocking Ranges

Suppose that we suspect a gradient for some soil factor (moisture, organic matter, fertility, etc.) across the ranges. In order to remove the effect of the gradient on our comparisons between the two accessions, we should probably 'block' each range as a factor in our model.

Let us further assume that we block the accession 'treatments' into five blocks consisting of two plots each. If we randomly group pairs of the accessions into 5 sets, next randomly assign each set to a range, and third, randomly assign each accession within a set to the plots within ranges, we will have a randomized complete block design (RCBD) that can be modeled as $Y_{ij} = b_j + \mu_i + \epsilon_{ij}$ where the definition of parameters is the same as the CRD model, but with the added term for a blocking factor.

Mixed Linear Model

In this second model, the accessions are selected so we should consider them to be fixed effect parameters. Although the block parameter represents a sample of many possible blocks in the field trial, there are only a few blocks that represent a "nuisance" source of variability, so we can treat them as a fixed effect, while the parameter ε_{ij} represents the residual or error in the model which is based on a sample of plots to which experimental units are assigned.

Thus ε_{ij} is considered random effects where $\varepsilon_{ij} \sim$ i.i.d. $N(0, \sigma_e^2)$, and the model is considered a mixed linear model.

Table 9 ANOVA outline for mixed model.

Source	df	MS	F	Prob
Block	4	n/a	n/a	n/a
Accession	1	n/a	n/a	n/a
Residual	4	n/a	n/a	n/a

Regression and Prediction

Objectives

Demonstrate ability to conduct and interpret regression analyses.

Statistical Inference

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots, and simple descriptive statistics such as the range (maximum, minimum), quartiles, and the sample average, median, mode. These exploratory data analysis (EDA) techniques should always be used prior to **estimation** and **hypothesis testing**. However, prior to conducting EDA, the phenotype should be **modeled** using the parameters in the experimental and sampling designs.

Linear Regression

Historically, linear (and non-linear) regression has not been utilized extensively by plant breeders, although it provides the conceptual foundation for understanding additive genetic models and analysis of covariance. Recently, with the emergence of molecular marker technologies, the importance of linear regression has manifested itself in the development of predictive methods

such as Genomic Prediction. Linear regression is an approach to modeling the relationship between a scalar dependent variable *Y* (e.g., harvestable grain yield per unit of land) and one or more explanatory variables (e.g., breeding values of lines involved in crosses) denoted by *X*.

Basic Assumptions

In linear regression, the phenotype is modeled using a linear function. There are four basic assumptions made about the relationship between a response variable *Y* and an explanatory variable *X*.

1. All *Y* values are from independent experimental or sample units.
2. For each value of *X*, the possible *Y* values are distributed as normal random variables.
3. The normal distribution for *Y* values corresponding to a particular value of *X* has a mean μ{Y|X} that lies on a line (Equation 21):

$$\mu\{Y|X\} = \beta_0 + \beta_1 X$$

Equation 21 Formula for estimating the mean.

where:
β_0 = the intercept and represents the mean of the *Y* values when *X* = 0,
β_1 = the slope of the line, that is, the change in the mean of *Y* per unit increase in *X*.

4. The distribution of *Y* values corresponding to a particular value of *X* has standard deviation $\sigma\{Y|X\}$. The standard deviation is usually assumed to be the same for all values of *X* so that we may write $\sigma\{Y|X\}=\sigma$. Violation of the last assumption is typical in plant breeding data and the development of methods to account for unequal variances is an area of important research.

Simple Linear Regression

Suppose we have *n* observations of a response variable Y and an explanatory variable X: (X_1,Y_1), . . . , (X_n,Y_n). The model can be rewritten as in Equation 22:

$$Y_i = \beta_0 + \beta_1 X_1 + e_i$$

Equation 22 Linear model for phenotype.

where:
\small All terms are as defined previously,

for i = 1, . . . , n experimental units. e_i , . . . , e_n are assumed to be independent normal random variables with mean 0 and standard deviation σ{Y|X}=σ. Thus, least-squares estimates of the Y_i values are obtained using Equation 23:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_i x_i$$

Equation 23 Least squares estimates model for Y values.

where:
$\small All terms$ are as defined previously.

The residual e_i (e_1, . . . , e_n) can be calculated as represented in Equation 24:

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_i x_i)$$

Equation 24 Formula for estimating residuals.

where:
$\small All terms$ are as defined previously.

Parameter Estimates

The estimators for parameters β_0, β_i , and **σ** are

$$\hat{\beta}_i = \frac{\sum_{i-1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1(\bar{x})$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n-2}}$$

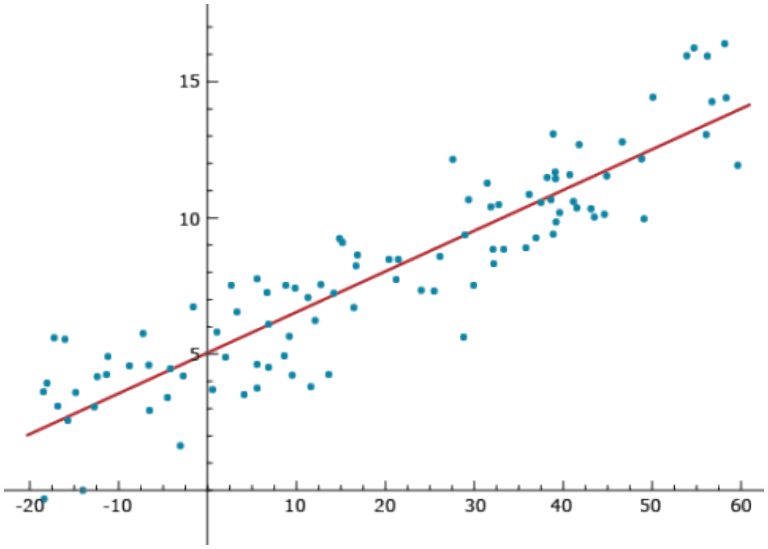


Fig. 10 Estimator plot.

Prediction

Notice that Equation 23, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_i x_i$, provides a predicted value of Y_i. Imagine that the xi (i=1, …, n) values are a genotypic index for cultivar/individual i, such as the sum of all allelic values (+1 or -1) at quantitative trait loci throughout the genome. Some cultivars could have 60 positive allelic values and no negative allelic values, while other cultivars could have a genotypic index of -20 (see Figure 10). If the positive genotypic index values are associated with high phenotypic values, such as in the figure, then we will have a strong positive linear relationship between the genotypic index and the phenotypes. A strong linear relationship can enable the plant breeder to predict phenotypes without having to spend resources on growing cultivars. The stronger the linear relationship is between the genotypic index and the phenotype (less variability around the line), the better the ability to predict. This concept represents the foundation for what is widely referred to as Genomic Prediction.

There are a number of details about how allelic values are estimated and combined into genotypic indices. The foundational concepts that address these details are covered in the Introduction to Quantitative Genetics section.

Analysis of Covariance

Objective: Demonstrate ability to conduct and interpret Analysis of Covariance

Statistical Inference

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots, and simple descriptive statistics such as the range (maximum, minimum), quartiles, and the sample average, median, mode. These exploratory data analysis (EDA) techniques should always be used prior to **estimation** and **hypothesis testing**. However, prior to conducting EDA, the phenotype should be **modeled** using the parameters in the experimental and sampling designs.

AOC is typically applied when there is a need to adjust results for variables that cannot be controlled by the experimenter. For example, imagine that we have two plant introduction accessions and we wish to evaluate whether these two accessions are unique with respect to yield. Also, imagine that germination rates for each accession is different but unknown, especially under field conditions in a new environment. We could decide to overplant each plot and reduce the number of plants per plot to a constant number equal to a stand count that is typical under current Agronomic practices. However, such an approach will be labor-intensive and not as informative as simply adjusting plot yields for stand counts.

Example

Assume that we have 10 plots available for purposes of testing the null hypothesis that there is no difference in their yield. Also, assume that we have enough seed to plant 200 seeds in each plot, although current agronomic practices are more closely aligned with stands of about 125 plants per plot. Let us next assume that the 10 plots consist of two-row plots that are arranged in a 5×2 grid consisting of five ranges with 2 plots per range. Suppose that we suspect a gradient of some soil factor (moisture, organic matter, fertility, etc.) across the ranges. In order to remove the effect of the gradient on our comparisons between the two accessions we should probably 'block' each range as a factor in our model. If we randomly group pairs of the accessions into 5 sets, next randomly assign each set to a range and third randomly assign each accession within a set to the plots within ranges, we will have a RCBD. At the time of harvest, we evaluate the plots for yield (bushels per acre) as well as stand counts (plants per plot). The resulting data are arranged in Table 10.

Table 10 Data from RCBD plot.

n/a	PI accession 1		PI accession 2	
Block	**(t/ha)**	**(plants/plot)**	**(t/ha)**	**(plats/plot)**
1	1.69	91	1.88	102
2	1.95	122	1.82	89
3	2.20	143	2.01	139
4	2.13	145	2.01	147
5	1.76	110	1.95	112

Model Equation

If we model the yield data as $Y_{ij} = b_j + \mu_i + \varepsilon_{ij}$, where Y_{ij} is the yield of plot ij , μ_i represents the mean of accession i, b_j represents the j^{th} block in which each pair of accessions are grown and ε_{ij} ~ i.i.d. $N(0,\sigma^2)$, the resulting analysis revealed that the variability between accessions is not much greater than the residual variability. We might interpret this to mean that there is no difference in yield between the two accessions. However, our real interest is in whether there is a difference between the accessions at the same stand counts. A more appropriate model for the question of interest is as in Equation 25:

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + b_j + \varepsilon_{ij}$$

Equation 25 Formula for calculating phenotype in a plot.

where:
α_i = intercept for accession i,
β_i = slope of accession i,
X_{ij} = the j^{th} stand count in accession i (i = 1, 2,)
b_j = random effect parameter.

The model has two intercepts, denoted α_i (i = 1, 2) for each of the accessions, and two slopes denoted β_i (i = 1, 2), for each of the accessions. X_{ij} is the j^{th} stand count in accession i (i = 1, 2). The model also has random effects parameters denoted by b_j and $\varepsilon_{i,j}$ where b_j~ i.i.d. $N(0,\sigma_b^2)$ and $\varepsilon_{i,j}$ ~ i.i.d. $N(0, \sigma_e^2)$. The resulting analyses of variability associated with each of the parameters is known as Analysis of Covariance, and can be thought of as an approach that takes advantage of both regression and ANOVA, i.e., an AOC model includes parameters representing both regression and factor variables. The result of the estimation procedure will enable us to evaluate whether the accessions are equal at stand counts of interest. In other words it will be possible

to adjust yield values to various stand counts of interest. As a matter of ethics in science, the variable stand count of interest needs to be modeled prior to conducting the field trial.

Computational Considerations

Key Concepts

As long as data are balanced all computational algorithms will provide the same estimates of variance components.

- When data are not balanced, either by design or accident, simple algorithms implemented in many widely used software packages (EXCEL, JMP for exampes) will not provide correct estimates of variance components.

Statistical Inference

The purpose of statistical inference is to interpret the data we obtain from sampling or designed experiments. Preliminary insights come from graphical data summaries such as bar charts, histograms, box plots, stem-leaf plots and simple descriptive statistics such as the range (maximum, minimum), quartiles, and the sample average, median, mode. These exploratory data analysis (EDA) techniques should always be used prior to **estimation** and **hypothesis testing**. However, prior to conducting EDA, the phenotype should be **modeled** using the parameters in the experimental and sampling designs.

Computational Methods

Most plant breeding data are obtained using a limited number of field plot designs consisting of lines (cultivars, hybrids, synthetics, etc.), environments and occasionally complete blocks, but usually incomplete blocks, within environments. Further, numbers of observations per source of variation are seldom balanced; even if designed for balance, some plots are lost during the course of a growing season. Thus, while the algorithm for obtaining EMS (described in the section Statistical Inference: Analysis of Variance) is useful for learning basic concepts, it is of little practical use for most plant breeding projects. Just as the estimates of means need to be adjusted through use of lsmeans, advanced computational methods are needed to obtain accurate estimates of variance components of the linear model when data are obtained from unbalanced conditions. The computational methods are affected by fixed effects, random effects or a mixture of both types of effects. There are three primary computation methods for estimating variance

components: Method of Moments (MM), Maximum Likelihood (ML), and Restricted Maximum Likelihood (REML).

Regression, Anova, and AOC

Computation of the MM estimators of variance components is essentially a matter of equating observed mean squares, calculated using the sums of squared deviations and cross products, with the expected mean squares, as demonstrated by Lorenzen and Anderson (1993, Design of Experiments: A No-Name Approach. p 71-72). These are appropriate if the data are balanced. Most advanced statistical software packages, e.g., SAS and R, calculate the sums of squares and cross products for the MM using the MIVQUE(0) algorithm (Minimum Variance Quadratic Unbiased Estimator, with no weighting for random effects).

Computation of ML and REML are derived from MIVQUE(0); both use MIVQUE(0) estimates as starting points in an iterative algorithm that maximizes the likelihood function, assuming that the random effects are distributed as random normal variables. The difference between ML and REML is that the likelihood function in REML is maximized only for the random effects, i.e., the fixed effects are removed from the likelihood function. For a model consisting of only random effects, both ML and REML will provide the same results. Indeed, for completely balanced data from random effects models, all three computational methods provide the same results. When dealing with unbalanced data or mixed effect models, REML has been shown to be the best computational method.

Further Considerations

As a practical matter, if your data is missing less than 10% of the experimental units within any environment, the MM approach will provide estimates that are almost as good as REML. Otherwise, the estimates should be obtained with mixed model equations (MME) and a REML algorithm. We encourage the use of R or SAS software for conducting data analyses. R is free, while the SAS license fees pay for more rigorous quality assurance.

Matrix Algebra

A **matrix** is a collection of numerical values arranged in rows and columns. Herein, the elements of a matrix are enclosed in brackets. For example,

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

is a matrix with 4 elements arranged in 2 rows and two columns.

Matrices with more than two or more rows and columns are denoted with upper case bold letters. Vectors are a special type of matrix with only one row or one column. For example,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \text{ or } y = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}.$$

Special Kinds of Matrices

Vector matrices are denoted with lower case bold italicized letters. A matrix consisting of only one row and one column is referred to as a **scalar matrix**. A **square matrix** has the same number of rows and columns. A **diagonal matrix** is a square matrix with off-diagonal elements equal to 0. An **identity matrix** is a diagonal matrix with diagonal elements = 1. The identity matrix is almost always denoted **I**.

Operations

Matrices must be conformable, i.e., matrix operations have requirements on the numbers of rows and columns.

It is possible to add or subtract two matrices, but only if they have the same numbers of rows and columns. For example,

$$C = A - B = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \begin{bmatrix} a_{11} - b_{11} & a_{12} - b_{12} & a_{13} - b_{13} \\ a_{21} - b_{21} & a_{22} - b_{22} & a_{23} - b_{23} \\ a_{31} - b_{31} & a_{32} - b_{32} & a_{33} - b_{33} \end{bmatrix}.$$

It is possible to multiply a matrix by a scalar value (say 'v') by simply multiplying all elements of the matrix by the scalar value, v. Thus,

$$D = vA = Av = D = \begin{bmatrix} va_{11} & va_{12} & va_{13} \\ va_{21} & va_{22} & va_{23} \\ va_{31} & va_{32} & va_{33} \end{bmatrix}.$$

Multiplying Vectors

It is possible to multiply two vectors, but only if 1) one of the vectors is a row vector, 2) the second is a column vector, 3) the row vector has as many elements as the column vector. For example,

$$\begin{bmatrix} 1 & 3 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \textit{ is a legal operation, whereas } \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \textit{ is not.}$$

The operation of vector multiplication in the first instance indicates that we have a 1×3 matrix multiplied by a 3×1 matrix. The way we carry out the vector multiplication is to multiply the elements from each matrix in a pairwise manner, then sum the results of all 3 pairs:

$$\begin{bmatrix} 1 & 3 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} = 1 \times 2 + 3 \times 4 + 5 \times 6 = 44.$$

Multiplying Vectors In Reverse

We could also apply the rule of multiplying and summing pairs of elements to the reverse arrangement of these two vectors:

$$\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 4 & 12 & 20 \\ 6 & 18 & 30 \end{bmatrix}.$$

Notice that the order of arrangement of vectors matters. Likewise, the arrangement of matrices that are to be multiplied matters. Virtually all types of matrix multiplication involve the multiplication of a row vector by a column vector. In essence, we partition each matrix into a set of row and column vectors, then apply the rules of vector multiplication.

Matrix Multiplication

Let us consider **C**=**AB**. $\mathbf{c}_{ij} = \mathbf{a}_i.\mathbf{b}_j$, where $\mathbf{a}_i$ is the i^{th} row vector of **A** and $\mathbf{b}_j$ is the j^{th} column vector of **B**. For example,

$$Given:$$

$$A = \begin{bmatrix} 2 & 8 & -1 \\ 3 & 6 & 4 \end{bmatrix} \; and \; B = \begin{bmatrix} 1 & 7 \\ 9 & -2 \\ 6 & 3 \end{bmatrix}$$

$$then$$

$$c11 = a1 \cdot b1 = \begin{bmatrix} 2 & 8 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 9 \\ 6 \end{bmatrix} = 2x1 + 8x9 - 1x6 = 68$$

$$and$$

$$c12 = a1 \cdot b2 = 1, \; c21 = a2 \cdot b1 = 81, \; c21 = a2 \cdot b1 = 21$$

$$and$$

$$AB = \begin{bmatrix} 68 & 1 \\ 81 & 21 \end{bmatrix} = C$$

.

Notice that matrix multiplication requires that the first matrix must have as many columns as the second matrix has rows. Thus, **AB** is usually not equal to **BA**. Indeed, while **AB** may be possible, **BA** may not. Lastly verify for yourself that **IA**, **IB** and **Ix** = **A**, **B** and **x** respectively.

Additional Important Operations

The transpose of a matrix, denoted as **A'** (or $\mathbf{A^t}$ or $\mathbf{A^T}$) is a useful operation in which the first row of a matrix becomes the first column of its transpose, while the second, third, … etc rows become the second, third, … etc columns of its transpose. For example,

$$A = \begin{bmatrix} 2 & 8 & -1 \\ 3 & 6 & 4 \end{bmatrix}, \quad A' = \begin{bmatrix} 2 & 3 \\ 8 & 6 \\ -1 & 4 \end{bmatrix}.$$

The inverse of a matrix is best understood by recalling that in scalar algebra the inverse of a number multiplied by the number will be = 1. Thus the inverse of x is x^{-1}. In matrix algebra the inverse of a matrix is a matrix when multiplied by the original matrix is **I**. That is $\mathbf{AA^{-1}}$ = $\mathbf{A^{-1}A}$= **I**. Only square matrices will have an inverse, although not all square matrices will have an inverse. Bernardo describes how to calculate the inverse of a simple 2×2 matrix and it is possible to calculate inverse matrices consisting of 3×3 elements, but calculations of inverses of larger matrices are better left to software.

References

Bernardo, R. 1996. Best linear unbiased prediction of maize single-cross performance. Crop Sci 36:50–56.

Bernardo, R. 2002. Breeding for quantitative traits in plants. Stemma Press.

Comstock, R. E. 1978. Quantitative genetics in maize breeding. In: Walden DB (ed) Maize breeding and genetics. Wiley, New York, p 191–206.

Christensen, R. 1997. Log-Linear Models and Logistic Regression (2nd ed.) New York: Springer-Verlag.

Crossa, J., R. C. Yang, P. and Cornelius. L. 2004. Studying crossover genotype × environment interaction using linear-bilinear models and mixed models. J. Agric. Biol. Environ. Stat. 9 (3):362–80.

Fehr, W. R. 1991. Principles of cultivar development vol. 1: Theory and technique. MacMillan Publishing Company, USA.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399–433.

Fisher, R. A. 1925. Theory of Statistical Estimation. Oliver & Boyd, Edinburgh.

Fisher, R. A. 1928. Statistical methods for research workers. Scotland: Oliver and Boyd.

Fisher, R. A. 1935. The Design of Experiments (8th ed., 1966), New York: Hafner Press.

Hayes, H. K., and F. R. Immer, 1942. Methods Of Plant Breeding. McGraw-Hill publications in the agricultural sciences.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Lorenzen, T., and V. Anderson. 1993. Design of Experiments: A No-Name Approach. p 71-72)

Lush, J. L. 1948.The genetics of populations. Mimeographed notes, Iowa State College, Ames, Iowa.

McCullagh, P,. and, J. A. Nelder. 1989. Generalized Linear Models, volume 37 of Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2nd edition

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic values using genome-wide dense marker maps. Genetics 157: 1819–1829.

Piepho, H. P. 2009. Data Transformation in Statistical Analysis of Field Trials with Changing Treatment Variance. Agron. J. 101:865-869.

How to cite this module: Beavis, W. and A. A. Mahama 2023. Plant Breeding Basics. In W. P. Suza, & K. R. Lamkey (Eds.), *Quantitative Genetics for Plant Breeding*. Iowa State University Digital Press.

Applied Learning Activities

The following downloadable Applied Learning Activities (ALAs) and associated files are aligned with the chapters linked below:

Chapter 1

- Disequilibrium [PDF]
- Ideal Population_HWE [PDF]
- Fate of a Rare Allele [XLSX]

Chapter 2

- Gametic and Linkage D_Likelihood [PDF]
- Gametic and Linkage D [PDF]
- Gamete and LD Expected frequencies with linkage and selfing [XLSX]

Chapter 3

- Full-sib Mating [PDF]
- Introgression [PDF]
- Self-pollination [PDF]
- Example – Self-pollination with Coefficient of Inbreeding [XLSX]

Chapter 4

- Relationship-Coefficient_Solve-Example [PDF]
  - Eval Dist Metrics [XLSX
- Cluster Analysis [PDF]
  - ALA 4.6_DS [CSV]
  - Cluster Analysis [TXT] (For use as .R file)

Chapter 5

- Breeding Values ALA [PDF]
  - ALA5.2 DS [XLSX]
- Epistasis [PDF]
- GV and Population Means_Solved-example [PDF]

Chapter 6

- Genetic Variance Components [PDF]
- Quantitative Genetic Models ds6 [CSV]
- Genetic Covariance and Heritability [PDF]
- Genetic Covariances and Heritability ds7 [CSV]

Chapter 7

- Evaluating new breeding germplasm ALA [PDF]
  - ALA7.1_ds [CSV]
  - ALA7.1 [TXT] (For use as .R file)
- Estimating heritability of testcross population ALA [PDF]
  - ALA7.2_ds [CSV]
  - ALA7.2 [TXT] (For use as .R file)
- Deriving Expectation of Mean Squares [PDF]

Chapter 8

- GCA and SCA_North Carolina Design II [PDF]
  - QG_Mod8_ALA8.3_ds1 [CSV]
- General and Specific Combining Ability [PDF]
  - QG_Mod8_ALA8.1_ds1 [CSV]
  - QG_Mod8_ALA8.1_ds2 [CSV]
  - QG_Mod8_ALA8.1 R [TXT] (For use as .R file)
- Using the North Carolina II Design [PDF]
- QG_Mod8_ALA8.2 [TXT] (For use as .R file)

Chapter 9

- Estimate heritability [PDF]
- Estimates of heritability_unbalanced data set [PDF]
  - QG_Mod9_ALA9.2_ds [CSV]
  - QG_Mod9_ALA9_2 [TXT] (For use as .R file)
- Response to Selection and Genetic Gain [PDF]

Chapter 10

- Types of GxE – Cluster Analysis ALA [PDF]
- Types of GxE – Partition GxE ALA [PDF]
  - Multi Environment Trials Types of GxE ds5 [CSV]
- QG_Mod10_ALAs_ds [CSV]
- QG_Mod10_ALA10_1 [TXT] (For use as .R file)
- QG_Mod10_ALA10_2 [TXT] (For use as .R file)
- QG_Mod10_ALA10.2_env_pheno_ds [CSV]
- QG_Mod10_ALA10.2_env_pheno_stdize_ds [CSV]
- QG_Mod10_ALA10.2_env_gxe_ds [CSV]
- QG_Mod10_ALA10.2_env_gxe_stdize_ds [CSV]
- QG_Mod10_ALA10_3 [TXT] (For use as .R file)

Chapter 11

- Multiple Trait Selection ALA [PDF]
  - QG_Mod11_ALA11.2 [CSV]
  - QG_Mod11_ALA11.2 [XLSX]
- QG_Mod11_ALA11_2 [TXT] (For use as .R file)
- Matrix Algebra ALA [PDF]
  - QG_Mod11_ALA11.3_Review_Matrix_Algebra [PDF]
- Matrix Algebra – Smith-Hazel Index [PDF]
- QG_Mod11_ALA11.4 [CSV]
- QG_Mod11_ls mean s t1 and t2 [CSV]
  - QG_Mod11 ALA11.4 ls mean s t1 [CSV]
  - QG_Mod11 ALA11.4 ls mean s t2 [CSV]

Chapter 12

- Linear Mixed Models – Samples of lines ALA [PDF]
- Linear Mixed Models – Reduce plots by half ALA [PDF]
  - MET ds5 [CSV]
- MET ALA12_2 [TXT] (For use as .R file)
- MET ALA12_3 [TXT] (For use as .R file)
- MET ALA12_4 [TXT] (For use as .R file)

Chapter 13

- Simulation modeling ALA [PDF]
  - QG_Mod13_ALA13.1_ds [CSV]
  - QG_Mod13_ALA13.1 [TXT] (For use as .R file)
- QG_ALA selection GBLUP [PDF]
  - DS8 RILs [XLSX]
- Selection Heritability and Genetic Gain Simulation Comparison ALA [PDF]

Plant Breeding Basics

- Install R ALA [DOC]
- RawDataEarPhenotypes_0 [XLSX]
- Review EDA with R ALA P[PDF]
  - Review EDA with R ds3 [CSV]
- Review Exploratory Data Analysis ALA [PDF]
  - Review Exploratory Data Analysis ds2 [CSV]
- Review Statistical Inference Analysis of Variance ALA [PDF]
- Review Statistical Inference Regression and Prediction ALA [PDF]
  - Review Statistical Inference Regression and Prediction ds4 [CSV]
- Review Trait Measures ALA [PDF]
- Review Types of Models Data Management ALA [PDF]
  - Review Types of Models Data Management ds1 [XLSX]
- Review Types of Models Field Plot Design ALA [PDF]

Contributors

Editors

Walter Suza

Suza is an Adjunct Associate Professor at Iowa State University. He teaches courses on Genetics and Crop Physiology in the Department of Agronomy. In addition to co-developing courses for the ISU Distance MS in Plant Breeding Program, Suza also served as the director of Plant Breeding e-Learning in Africa Program (PBEA) for 8 years. With PBEA, Suza helped provide access to open educational resources on topics related to the genetic improvement of crops. His research is on the metabolism and physiology of plant sterols. Suza holds a Ph.D. in the plant sciences area (with emphasis in molecular physiology) from the University of Nebraska-Lincoln.

Kendall Lamkey

Lamkey is the Associate Dean for Facilities and Operations for the College of Agriculture and Life Sciences at Iowa State University. He works in collaboration with the dean, associate deans, department chairs, college-level centers, and other unit leaders to ensure that operations directly advance the mission of the college and that resources are deployed wisely and efficiently. Previously, he served as the chair for the Department of Agronomy at Iowa State University, where, in addition to advocating for research and the PBEA program, he oversaw the Agronomy Department's educational direction, its faculty, and Agronomy Extension and Outreach. Dr. Lamkey is a corn breeder and quantitative geneticist and conducts research on the quantitative genetics of selection response, inbreeding depression, and heterosis. He holds a Ph.D. in plant breeding from Iowa State University and a master's in plant breeding from the University of Illinois. Lamkey is a fellow of the American Society of Agronomy and the Crop Science Society of America and has served as an associate editor, technical editor, and editor for *Crop Science.*

Chapter Authors

William Beavis, Ursula Frei, M. L. Harbur, Reka Howard, Kendra Meade, Laura Merrick, Ken Moore, Ron Mowers, and Dennis Todey

Contributors

Anthony A. Mahama, Gretchen Anderson, Todd Hartnell, Andy Rohrback, Tyler Price, Glenn Wiedenhoeft, and Abbey K. Elder