The goal of quantitative genetics is the understanding of the causes of variation in traits of interest.

The Nature of Quantitative Variation

One of the first critical observations about the nature of trait variation was made by the great Danish botanist Wilhelm Johannsen (1903, 1909). Starting from a common stock of beans, he produced several inbred lines. For each line, parental seeds of different weights were planted and the mean seed weight in the offspring measured. Johannsen observed that the variation within a pure line was not heritable — the mean seed weights of the offspring were essentially independent of the weights of the parental seed (Figure 1.1). To clarify the distinction between genetic and environmental effects, he coined the terms genotype (to denote genetically identical members of a pure line) and phenotype (the actual observed value for an individual — a compounding of genetic and environmental effects).

![Graph showing mean offspring seed size as a function of parental seed size for some of Johannsen’s pure lines.](image)

**Figure 1.1** Mean offspring seed size as a function of parental seed size for some of Johannsen’s pure lines. The data for the different lines are denoted by different symbols. If there is a heritable component to seed weight within a pure line, a line with positive slope is expected — larger parents should yield larger offspring. However, within each line, mean offspring size is essentially independent of the parental phenotype. (Data from Johannsen 1903.)

Expressed another way, if we grow out a large number of progeny from a single inbred line, and then select the largest beans to plant the following season, we would observe no change (i.e., no response to selection) in seed weight in the next generation because none of the within-line variation is due to heritable genetic differences. Within the pure line, all differences are due to environmental differences. Thus identical genotypes can differ in trait values, reflecting differences in environmental contributions, and this variation is of no use (indeed, it is a confounding factor) in attempts to improve plant performance.
While variation within a pure (i.e., inbred) line is environmental (and hence not heritable), there are consistent differences between lines. This is an example of genetic – heritable – variation. Thus trait (phenotypic) variation arises from differences in both genotypes and environments, and we cannot distinguish the relative contributions of these two sources simply by looking at phenotypic variation. To do so, we (as in Figure 1.1) need to consider variation in relatives.

Further (as we will shortly see) not all differences in genotypes are heritable. If we clone individuals (such as the offspring from a pure line), then the parents and offspring have identical genotypes. However, with sexual reproduction, the genotype of an offspring is a mixture of the genotypes of its parents. One of the goals of quantitative genetics is to work out just how much an offspring resembles a parent in such cases, which in turns informs us as to how efficient selection can be. This leads us into a discussion of additive and non-additive genetic variances that we will begin later in this lecture. For now, let’s consider the nature of genetic variation.

Causes of Genetic Variation: The Multiple-factor Hypothesis

We are used to thinking about the genetics of discrete plant phenotypes, for example tall versus short peas in Mendel’s experiments. In this example, variation (different alleles) at a single gene contributes to a major phenotypic difference. However, for many traits of interest (such as yield), it is often the case that a large number of genes, each of modest effect, collectively contribute to the genetic variation. This is the multiple-factor hypothesis.

Early support for this multiple-factor hypothesis came from H. Nilsson-Ehle (1909), a Swedish geneticist working on various cereal crops. Many of the characters that he examined yielded 3:1 ratios in the F$_2$ generation following the cross of two parental strains, consistent with expectations for a single segregating locus with one allele completely dominant over the other. However, there were some striking exceptions. For example, when red-seeded and white-seeded wheat strains were crossed, the F$_1$ progeny were identical in color (light red), but in some of the F$_2$ crosses, a ratio of 63 red:1 white seeds was observed. Nilsson-Ehle interpreted this to be the result of the segregation of three independent factors, the initial parents being $AABBCC$ and $aabbcc$, all members of the F$_1$ being $AaBbCc$ and hence uniform in color, and the F$_2$ consisting of all possible genotypes, only one of which ($aabbcc$) gives rise to white seed. The probability of obtaining an $aabbcc$ offspring from an $AaBbCc \times AaBbCc$ cross is $(1/4)^3 = 1/64$.

From these results, Nilsson-Ehle arrived at two general conclusions. First, sexual reproduction can produce a huge diversity of genotypes. For example, since a locus with two alleles $A$ and $a$ can produce three genotypes ($AA$, $Aa$, and $aa$), ten diallelic loci can produce $3^{10} \approx 60,000$ genotypes. Second, given this huge potential diversity of genotypes, apparently new types appearing within a population may be the result of rare segregants rather than new mutations.

Subsequent studies quickly confirmed these ideas. East (1911, 1916) and Emerson (1910; Emerson and East 1913) examined quantitative variation in a large number of plants. Typically, strains differing widely in some character were crossed and the variance of the resulting F$_1$ and F$_2$ generations recorded. In most of these crosses, especially when the parental populations were formed by repeated self-fertilizations, an outbreak of variation was seen in the F$_2$ (Figure 1.2). Such outbreaks of variation, resulting from the segregation of multiple genotypes from the F$_1$ heterozygotes, are consistent with the multiple-factor hypothesis. For example, if the two parents being crossed are inbred lines with genotypes $AABBCC$ and $aabbcc$, then the resulting F$_1$ also has a single genotype $AaBbCc$. However, in the F$_2$, all of the F$_1$ heterozygotes can segregate, so that an $Aa$ parent can have $AA$, $Aa$, or $aa$ offspring. Thus, the F$_2$ consists of a large collection of different genotypes, and hence is more variable (having both genetic and environmental contributions to variation) than either of the parental or F$_1$ lines (which have only environmental contributions to the variance).

Lecture 1, pg. 2
Figure 1.2  The distribution of ear size in the F₁ and F₂ generations formed by crossing two inbred lines of corn differing in ear length. The observed number of ears is given below each size class. The variation seen in the P₁, P₂ and F₁ populations is due entirely to environmental factors, as all individuals in each population have the same genotype. These three populations show roughly similar amounts of variation. In contrast, the F₂ generation shows considerably more variation, reflecting the diversity of genotypes in this population generated by segregation of genes in the F₁ parents. (Data from East 1911.)

Variance, Covariance and Regressions

Given the above discussion, it should come as no surprise that quantitative genetics is based on measures of variability and association, and here we remind the reader of a few basic facts for standard statistic measures of variation (the variance) and association (covariances, correlations, and regressions).

Lecture 1, pg. 3
The Variance:

The standard measure of variation is the variance,

$$Var(x) = E[(x - \mu_x)^2]$$  \hspace{1cm} (1.1a)

Here $E()$ denotes the expected value or population mean of the quantity of interest, so that the variance is the average value of the squared deviation of a random variable about its mean ($\mu_x$). $Var(x)$ is a measure of uncertainty – the larger the variance, the more spread of a variable about its mean. Note that we can also write the variance as

$$Var(x) = E[x^2] - \mu_x^2$$  \hspace{1cm} (1.1b)

Note that if the mean is zero, then $Var(x) = E[x^2]$. The variance of $x$ is also written as $\sigma^2(x)$ and we often unconsciously switch between these different notations.

The Covariance:

One of the most useful measures in quantitative genetics is the covariance between two variables, which is a (linear) measure of association. Formally, the covariance, $Cov(x, y)$, of two random variables $x$ and $y$ is defined by

$$Cov(x, y) = E[(x - \mu_x) \ast (y - \mu_y)]$$
$$= E(xy) - \mu_x\mu_y$$
$$= \text{mean of the product} - \text{product of the means}$$  \hspace{1cm} (1.2)

The covariance is also written as $\sigma(x, y)$ and $\sigma_{xy}$. As the figure (below) shows, if $x$ and $y$ are positively associated, then $Cov(x, y) > 0$, while if they are negatively associated, then $Cov(x, y) < 0$. Note that the covariance is a measure of the linear association between two variables — even though $x$ perfectly predicts $y$ is the far right panel, there is no linear trend, so that $Cov(x, y) = 0$. While $Cov(x, y) = 0$ when $x$ and $y$ are independent, the converse is NOT true, as $Cov(x, y) = 0$ does not necessarily imply that $x$ and $y$ are independent (again, as evidenced by the last panel).

The correlation, $r(x, y)$ [the notation $\rho(x, y)$ and $\rho_{xy}$ is also used] is a scaled measure of the covariance, where

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}}$$  \hspace{1cm} (1.3)

Since the range of correlation is restricted to between $-1$ and $+1$, it provides a standard metric for comparing the amount of association between pairs of variables that show different levels of variation. For example, a covariance of 10 implies a relatively small association if both variables have a variance of 100 ($r = 10/100 = 0.1$), but complete association if both variables have a variance of 10 ($r = 10/10 = 1$).
Covariance and Regressions:

There is a very close connection between the regression of one variable on another and the covariance between the two variables. The slope $b_{y|x}$ for the best linear fit of $y$ given an observed value of $x$ is given by

$$b_{y|x} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$  (1.4a)

The (best linear) predicted value $\hat{y}$ for $y$ given we know $x$ is

$$\hat{y} = \bar{y} + b_{y|x} (x - \bar{x})$$  (1.4b)

Correlations and regression slopes are related as follows:

$$r(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = b_{y|x} \sqrt{\frac{\text{Var}(x)}{\text{Var}(y)}}$$  (1.4c)

Thus, if the variances of $x$ and $y$ are the same, then $r(x, y) = b_{y|x} = b_{x|y}$. Finally, the residual variance of a regression is the noise in the difference between the predicted and actual values,

$$\sigma^2_e = \sigma^2(\hat{y} - y) = (1 - r^2)\sigma_Y^2$$  (1.4d)

The greater the correlation $r$ between $x$ and $y$, the smaller the residual variance (i.e., the smaller the uncertainty, in predicting the value of $y$ given $x$). Put another way, $r^2$ is the fraction of the total variation in $y$ explained given we know $x$.

Useful Properties of Variances and Covariances:

- The covariance function is symmetric, $\text{Cov}(x, y) = \text{Cov}(y, x)$
- The covariance of a variable with itself is the variance, e.g., $\text{Cov}(x, x) = \text{Var}(x)$
- If $a$ is a constant, then $\text{Cov}(ax, y) = a \cdot \text{Cov}(x, y)$
- $\text{Var}(ax) = a^2\text{Var}(x)$. This follows since $\text{Var}(ax) = \text{Cov}(ax, ax) = a^2\text{Cov}(x, x) = a^2\text{Var}(x)$
- $\text{Cov}(x + y, z) = \text{Cov}(x, z) + \text{Cov}(y, z)$, i.e., the covariance of a sum is the sum of covariances.

More generally,

$$\text{Cov}\left(\sum_{i=1}^{n} x_i, \sum_{j=1}^{m} y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(x_i, y_j)$$

- $\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y) + 2\text{Cov}(x, y)$. Hence, the variance of a sum, $\text{Var}(x + y)$, equals the sum of the variances, $\text{Var}(x) + \text{Var}(y)$, only when the variables have a covariance of zero.

Contribution of a Locus to the Phenotypic Value of a Trait

We now turn to the underlying theory for the analysis of complex traits. The basic model for quantitative genetics is that the phenotypic value $P$ of a trait is the sum of a genetic value $G$ plus an environmental value $E$,

$$P = G + E$$  (1.5a)

The genetic value $G$ represents the average phenotypic value for that particular genotype if we were able to replicate it over the distribution (or universe) of environmental values that the population
is expected to experience. However, not all genotypes act the same in the same environments. For example, one line may be high-performing over all environments while another may only be a high performer in wet environments. These are examples of **genotype-environment interaction**, and our basic model becomes

\[ P = G + E + G \times E \]  \hspace{1cm} (1.5b)

The genotypic value \( G \) is usually the result of a number of loci that influence the trait. However, we will start by first considering the contribution of a single locus, whose alleles are \( Q_1 \) and \( Q_2 \). We need a parameterization to assign genotypic values to each of the three genotypes, and there are several slightly different notations used in the literature:

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>( Q_1Q_1 )</th>
<th>( Q_1Q_2 )</th>
<th>( Q_2Q_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Trait Value:</td>
<td>( C )</td>
<td>( C + a(1 + k) )</td>
<td>( C + 2a )</td>
</tr>
<tr>
<td></td>
<td>( C - a )</td>
<td>( C + d )</td>
<td>( C + a )</td>
</tr>
</tbody>
</table>

Here \( C \) is some background value, which we usually set equal to zero. What matters here is the difference \( 2a \) between the two homozygotes,

\[ a = \frac{[G(Q_2Q_2) - G(Q_1Q_1)]}{2} \]  \hspace{1cm} (1.6a)

and the relative position of the heterozygote compared to the average of the homozygotes,

\[ d = \frac{G(Q_1Q_2) - \frac{G(Q_1Q_1) + G(Q_2Q_2)}{2}}{2} \]  \hspace{1cm} (1.6b)

If the genotypic value of the heterozygote is exactly intermediate, \( d = k = 0 \) and the alleles are said to be **additive**. If \( d = a \) (or equivalently \( k = 1 \)), then allele \( Q_2 \) is completely dominant to \( Q_1 \) (i.e., \( Q_1 \) is completely recessive). Conversely, if \( d = -a \) (or equivalently \( k = -1 \)) then \( Q_1 \) is dominant to \( Q_2 \). Finally if \( d > a \) (or equivalently \( k > 1 \)) the locus shows **overdominance** with the heterozygote having a larger value than either homozygote. Thus \( d \) (and equivalently \( k \)) measures the amount of dominance at this locus. Note that \( d \) and \( k \) are related by

\[ ak = d, \hspace{1cm} \text{or} \hspace{1cm} k = \frac{d}{a} \]  \hspace{1cm} (1.6c)

The reason for using both \( d \) and \( k \) is that some expressions are simpler using one parameterization over the other.

**Fisher’s Decomposition of the Genotypic Value**

Quantitative genetics as a field dates back to R. A. Fisher’s brilliant (and essentially unreadable) 1918 paper, in which he not only laid out the field of quantitative genetics, but also introduced the term variance and developed the important statistical tool of analysis of variance (ANOVA). Not surprisingly, his paper was initially rejected.

Fisher had two fundamental insights. First, that parents do not pass on their entire genotypic value to their offspring, but rather pass along only one of the two possible alleles at each locus. Hence, only part of \( G \) is passed on and thus we decompose \( G \) into component that can be passed along and those that cannot. Fisher’s second great insight was that phenotypic correlations among known relatives can be used to estimate the variances of the components of \( G \). We examine these in turn.

Fisher suggested that the genotypic value \( G_{ij} \) associated with the \( Q_iQ_j \) genotype can be written in terms of the average effects \( \alpha \) for each allele and a dominance deviation \( \delta \) giving the deviation of the actual value for this genotype from the value predicted by the average contribution of each of the single alleles,

\[ G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij} \]  \hspace{1cm} (1.7)

Lecture 1, pg. 6
The motivation for this decomposition is sexual reproduction. A parent of genotype $G_{ij}$ passes along either allele $A_i$ or $A_j$ (and hence a value of $\alpha_i$ or $\alpha_j$) to its offspring.

The predicted genotypic value is $\hat{G}_{ij} = \mu_G + \alpha_i + \alpha_j$, where $\mu_G$ is simply the average genotypic value,

$$\mu_G = \sum G_{ij} \cdot \text{freq}(Q_iQ_j)$$

Note that since we assumed the environmental values have mean zero, $\mu_G = \mu_P$, the mean phenotypic value. Likewise $G_{ij} - \hat{G}_{ij} = \delta_{ij}$, so that $\delta$ is the residual error, the difference between the actual value and that predicted from the regression. Since $\alpha$ and $\delta$ represent deviations from the overall mean, they have expected values of zero.

You might notice that Equation 1.7 looks like a regression. Indeed it is. Suppose we have only two alleles, $Q_1$ and $Q_2$. In this case we can re-express Equation 1.7 as

$$G_{ij} = a + bN + e$$

where $N$ is the number of copies of allele $Q_2$, and

$$a = \mu_G + 2\alpha_1 \quad b = \alpha_2 - \alpha_1 \quad e = \delta_{ij}$$

Note that

$$2\alpha_1 + (\alpha_2 - \alpha_1)N = \begin{cases} 2\alpha_1 & \text{for } N = 0, \text{ e.g., } Q_1Q_1 \\ \alpha_1 + \alpha_2 & \text{for } N = 1, \text{ e.g., } Q_1Q_2 \\ 2\alpha_2 & \text{for } N = 2, \text{ e.g., } Q_2Q_2 \end{cases}$$

Thus we have a regression, where $N$ (the number of copies of allele $Q_2$) is the dependent variable, the genotypic value $G$ the dependent variable, $(\alpha_2 - \alpha_1)$ is the regression slope, and the $\delta_{ij}$ are the residuals of the actual values from the predicted values. Recall from the standard theory of least-squares regression that the correlation between the predicted value of a regression ($\mu_G + \alpha_i + \alpha_j$) and the residual error ($\delta_{ij}$) is zero, so that $\sigma(\alpha_i, \delta_j) = \sigma(\alpha_k, \delta_j) = 0$.

To obtain the $\alpha$, $\mu_G$ and $\delta$ values, we use the notation of

<table>
<thead>
<tr>
<th>Genotypes:</th>
<th>$Q_1Q_1$</th>
<th>$Q_1Q_2$</th>
<th>$Q_2Q_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Trait Value:</td>
<td>0</td>
<td>$a(1+k)$</td>
<td>$2a$</td>
</tr>
<tr>
<td>frequency (HW):</td>
<td>$p_1^2$</td>
<td>$2p_1p_2$</td>
<td>$p_2^2$</td>
</tr>
</tbody>
</table>

A little algebra gives

$$\mu_G = 2p_1p_2a(1+k) + 2p_2^2a = 2p_2a(1 + p_1k)$$

(1.10a)

Recall that the slope of a regression is simply the covariance divided by the variance of the predictor variable, giving

$$\alpha_2 - \alpha_1 = \frac{\sigma(G, N_2)}{\sigma^2(N_2)} = a \left[ 1 + k \left( p_1 - p_2 \right) \right]$$

(1.10b)

See Lynch and Walsh, Chapter 4 for the algebraic details leading to Equation 1.10b. Since we have chosen the $\alpha$ to have mean value zero, it follows that

$$E[\alpha] = p_1\alpha_1 + p_2\alpha_2 = 0$$

When coupled with Equation 1.10b this implies (again, see L & W Chapter 4)

$$\alpha_2 = p_1a \left[ 1 + k \left( p_1 - p_2 \right) \right]$$

(1.10c)

$$\alpha_1 = -p_2a \left[ 1 + k \left( p_1 - p_2 \right) \right]$$

(1.10d)
Finally, the dominance deviations follow since
\[
\delta_{ij} = G_{ij} - \mu_G - \alpha_i - \alpha_j
\] (1.10e)

Note the important point that both \( \alpha \) and \( \delta \) are functions of allele frequencies and hence change as the allele (and/or genotype) frequencies change. While the \( G_{ij} \) values remain constant, their weights are functions of the genotype (and hence allele) frequencies. As these change, the regression coefficients change.

**Average Effects, Additive Genetic and Breeding Values**

Why do we go to all of this trouble to decompose the genotypic value into additive effects \( \alpha \) and the dominance deviations? Recall our comments that parents only pass along part of their genotypic value to their offspring. In particular, if a parent is \( A_iA_j \), then it only passes along either an \( A_i \) or \( A_j \) allele to its offspring, not both. Thus, it does not pass along the genotype value \( G_{ij} \), but rather a part of that. Which part? That part due to the effects of single alleles, \( \alpha_i + \alpha_j \). In particular, recall that \( \alpha_i \) is the average value (relative to the mean) for an individual with a copy of allele \( A_i \). A \( A_iA_k \) parent passes along allele \( A_i \) half the time, for an average contribution of \( \alpha_i \), and half the time passes along allele \( A_k \) for an average contribution of \( \alpha_k \). Thus, if we mate this parent at random, the average value of their offspring is

\[
\mu + \frac{\alpha_i + \alpha_k}{2}
\]

The additive genetic value \( A \) of an individual is simply the sum of the average effects \( \alpha \) summed over all loci. Thus, the expected mean value of offspring from a parent is just \( \mu + A/2 \). Likewise, the mean value of offspring from two parents is simply the population mean plus the average of their two additive genetic values. Since \( A \) predicts the mean for breeding proposal (under random mating), it is also called the breeding value (especially in animal breeding). While the breeding value \( A \) of an individual seems like a rather abstract quantity, it is actually straightforward to directly estimate. Suppose a male (sire) dairy cow is crossed to a large number of unrelated females, and we measure milk yield in his daughters. Suppose the population mean for milk yield is \( \mu = 100 \), while his offspring have mean 120. That sire’s breeding value is just \( 120 = \mu + A/2 \), or \( A = 2 \cdot (120 - 100) = 40 \). Thus we not only estimated his breeding value, we did so for a trait the male does not even express!

If the population has no variance in breeding values (\( A = 0 \) for all members of the population), then the mean value of offspring from any parents is simply the population mean — parental values do not predict offspring values and hence we do not expect a response to selection. Thus the variance in these breeding values — the additive-genetic variance — is of special interest.

**Genetic Variances**

Recall that the genotypic value is expressed as
\[
G_{ij} = \mu_g + (\alpha_i + \alpha_j) + \delta_{ij}
\]

The term \( \mu_g + (\alpha_i + \alpha_j) \) corresponds to the regression (best linear) estimate of \( G \), while \( \delta \) corresponds to a residual. Recall from regression theory that the estimated value and its residual are uncorrelated, and hence \( \alpha \) and \( \delta \) are uncorrelated. Since \( \mu_G \) is a constant (and hence contributes nothing to the variance) and \( \alpha \) and \( \delta \) are uncorrelated,
\[
\sigma^2(G) = \sigma^2(\mu_g + (\alpha_i + \alpha_j) + \delta_{ij}) = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij})
\] (1.13)

Lecture 1, pg. 8
Equation 1.13 is the contribution from a single locus. Assuming linkage equilibrium, we can sum over loci,

$$\sigma^2(G) = \sum_{k=1}^{n} \sigma^2(\alpha^{(k)}_i + \alpha^{(k)}_j) + \sum_{k=1}^{n} \sigma^2(\delta_{ij})$$

This is usually written more compactly as

$$\sigma^2_G = \sigma^2_A + \sigma^2_D$$  \hspace{1cm} (1.14)

where $\sigma^2_A$ is the additive genetic variance and represents the variance in breeding values in the population, while $\sigma^2_D$ denotes the dominance genetic variance and is the variance in dominance deviations.

Suppose the locus of concern has $m$ alleles. Since (by construction) the average values of $\alpha$ and $\delta$ for a given locus have expected values of zero, the contribution from that locus to the additive and dominance variances is just

$$\sigma^2_A = E[\alpha^2] = 2 \sum_{i=1}^{m} \alpha_i^2 p_i, \quad \text{and} \quad \sigma^2_D = E[\delta^2] = \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_{ij}^2 p_i p_j$$  \hspace{1cm} (1.15)

For one locus with two alleles, these become

$$\sigma^2_A = 2p_1 p_2 a^2[1 + k (p_1 - p_2)]^2 = 2p_1 p_2 [a + d (p_1 - p_2)]^2$$  \hspace{1cm} (1.16a)
and

\[ \sigma_D^2 = (2p_1 p_2 ak)^2 = (2p_1 p_2 d)^2 \]  
(1.16b)

The additive (dashed line), dominance (dotted line) and total (\(\sigma_G^2 = \sigma_A^2 + \sigma_D^2\), solid line) variance are plotted below for several different dominance relationships.

Note (from both the figures and from Equation 1.16) that there is plenty of additive variance even in the face of complete dominance. Indeed, dominance (in the form of the dominance coefficient \(k\)) enters the expression for the additive variance. This is not surprising as the \(\alpha\) arise from the best-fitting line, which will incorporate some of the departures from additivity. Conversely, note that the dominance variance is zero if there is no dominance (\(\sigma_D^2 = 0\) if \(k = 0\)). Further note that \(\sigma_D^2\) is symmetric in allele frequency, as \(p_1 p_2 = p_1(1 - p_1)\) is symmetric about 1/2 over (0,1).

**Epistasis**

Epistasis, nonadditive interactions between alleles at different loci, occurs when the single-locus genotypic values do not add to give two-locus genotypic value. For example, suppose that the average value of a \(AA\) genotype is 5, while an \(BB\) genotype is 9. Unless the average value of the \(AABB\) genotype is \(5 + 4 = 9\), epistasis is present in that the single-locus genotypes do not predict the genotypic values for two (or more) loci. Note that we can have strong dominance within each locus and no epistasis between loci. Likewise we can have no dominance within each locus but strong epistasis between loci.

The decomposition of the genotype when epistasis is present is a straight-forward extension of the no-epistasis version. For two loci, the genotypic value is decomposed as

\[
G_{ijkl} = \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kl})
+ (\alpha \alpha_{ik} + \alpha \alpha_{jl} + \alpha \alpha_{jk} + \alpha \alpha_{il})
+ (\alpha \delta_{ikl} + \alpha \delta_{jkl} + \alpha \delta_{kij} + \alpha \delta_{lij})
+ (\delta \delta_{ijkl})
= \mu_G + A + D + AA + AD + DD
\]  
(1.17)

Here the breeding value \(A\) is the average effects of single alleles averaged over genotypes, the dominance deviation \(D\) the interaction between alleles at the same locus (the deviation of the single locus genotypes from the average values of their two alleles), while AA, AD and DD represent the (two-locus) epistatic terms. AA is the additive-by-additive interaction, and represents interactions between a single allele at one locus with a single allele at another. AD is the additive-by-dominance interaction, representing the interaction of single alleles at one locus with the genotype at the other locus (e.g., \(A_i\) and \(B_jB_k\)), and the dominance-by-dominance interaction DD is any residual interaction between the genotype at one locus with the genotype at another. As might be expected, the terms in Equation 1.17 are uncorrelated, so that we can write the genetic variance as

\[
\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2
\]  
(1.18)

More generally, with \(k\) loci, we can include terms up to (and including) \(k\)-way interactions. These have the general form of \(A^n D^m\) which (for \(n + m \leq k\)) is the interaction between the \(\alpha\) effects at \(n\) individual loci with the dominance interaction as \(m\) other loci. For example, with three loci, the potential epistatic terms are

\[
\sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 + \sigma_{AAA}^2 + \sigma_{AAD}^2 + \sigma_{ADD}^2 + \sigma_{DDD}^2
\]  
(1.19)
Resemblance Among Relatives

Fisher’s second key insight was that we can use resemblances between relatives to estimate these genetic variances, in particular to estimate \( \text{Var}(A) \). Note that relatives resemble each other for quantitative traits more than they do unrelated members of the population for two potential reasons:

- Relatives share genes. The closer the relationship, the higher the proportion of shared genes
- Relatives share the same environment

The Genetic Covariance Between Relatives

The genetic covariance \( \text{Cov}(G_x, G_y) = \text{covariance of the genotypic values} \ (G_x, G_y) \) of the related individuals \( x \) and \( y \).

We will first show how the genetic covariances between parent and offspring, full sibs, and half sibs depend on the genetic variances \( V_A \) and \( V_D \) (note that \( V_A \), \( \text{Var}(A) \), and \( \sigma^2_A \) are all used interchangeably for the additive variance, with similar notation for the dominance variance). We will then discuss how these covariances are estimated in practice.

Genetic covariances arise because two related individuals are more likely to share alleles than are two unrelated individuals. Sharing alleles means having alleles that are identical by descent (IBD): namely that both copies of an allele can be traced back to a single copy in a recent common ancestor. Alleles can also be identical in state but not identical by descent. For example, both alleles in an \( A_1A_1 \) individual are the same type (identical in state), but they are only identical by descent if both copies trace back to (descend from) a single copy in a recent ancestor.

For example, consider the offspring of two parents and label the four allelic copies in the parents by 1 - 4, independent of whether or not any are identical in state.

Parents: \( A_1A_2 \times A_3A_4 \)

Offspring: \( o_1 = A_1A_3 \quad o_2 = A_1A_4 \quad o_3 = A_1A_3 \quad o_4 = A_2A_4 \)

Here, \( o_1 \) and \( o_2 \) share one allele IBD, \( o_1 \) and \( o_3 \) share two alleles IBD, \( o_1 \) and \( o_4 \) share no alleles IBD.

Offspring and one parent.

What is the covariance of genotypic values between an offspring \( (G_o) \) and its parent \( (G_p) \)? Denoting the two parental alleles at a given locus by \( A_1A_2 \), since a parent and its offspring share exactly one allele, one allele in the offspring came from the parent (say \( A_1 \)), while the other offspring allele (denoted \( A_3 \)) came from the other parent. To consider the genetic contributions from a parent to its offspring, write the genotypic value of the parent as \( G_p = A + D \). We can further decompose this by considering the contribution from each parental allele to the overall breeding value, with \( A = \alpha_1 + \alpha_2 \), and we can write the genotypic value of the parent as \( G_p = \alpha_1 + \alpha_2 + \delta_{12} \) where \( \delta_{12} \) denotes the dominance deviation for an \( A_1A_2 \) genotype. Likewise, the genotypic value of its offspring is \( G_o = \alpha_1 + \alpha_3 + \delta_{13} \), giving

\[
\text{Cov}(G_o, G_p) = \text{Cov}(\alpha_1 + \alpha_2 + \delta_{12}, \alpha_1 + \alpha_3 + \delta_{13})
\]

We can use the rules of covariances to expand this into nine covariance terms,

\[
\text{Cov}(G_o, G_p) = \text{Cov}(\alpha_1, \alpha_1) + \text{Cov}(\alpha_1, \alpha_3) + \text{Cov}(\alpha_1, \delta_{13})
+ \text{Cov}(\alpha_2, \alpha_1) + \text{Cov}(\alpha_2, \alpha_3) + \text{Cov}(\alpha_2, \delta_{13})
+ \text{Cov}(\delta_{12}, \alpha_1) + \text{Cov}(\delta_{12}, \alpha_3) + \text{Cov}(\delta_{12}, \delta_{13})
\]

Lecture 1, pg. 11
By the way have (intentionally) constructed $\alpha$ and $\delta$, they are uncorrelated. Further,

\[
Cov(\alpha_x, \alpha_y) = \begin{cases} 
0 & \text{if } x \neq y, \text{ i.e., not IBD} \\
Var(A)/2 & \text{if } x = y, \text{ i.e., IBD}
\end{cases} \tag{1.20a}
\]

The last identity follows since $Var(A) = Var(\alpha_1 + \alpha_2) = 2Var(\alpha_1)$, so that

\[Var(\alpha_1) = Cov(\alpha_1, \alpha_1) = Var(A)/2\]

Hence, when individuals share one allele IBD, they share half the additive genetic variance. Likewise,

\[
Cov(\delta_{xy}, \delta_{wz}) = \begin{cases} 
0 & \text{if } xy \neq wz, \text{ i.e., both alleles are not IBD} \\
Var(D) & \text{if } xy = wz, \text{ both alleles are IBD}
\end{cases} \tag{1.20b}
\]

Two individuals only share the dominance variance when they share both alleles. Using the above identities (1.20a and b), eight of the above nine covariances are zero, leaving

\[Cov(G_o, G_p) = Cov(\alpha_1, \alpha_1) = Var(A)/2\]  

(1.21)

**Half-sibs.**

Here, one parent is shared, the other is drawn at random from the population. We will call the seed parent the mother or female, and the pollen parent the father or male. In plant breeding, a half-sib family typically occurs by collecting the randomly-pollinated seed from a single plant.

The genetic covariance between half-sibs is the covariance of the genetic values between $o_1$ and $o_2$. To compute this, first note that $o_1$ and $o_2$ share either one allele IBD (from the mother) or no alleles IBD (since the fathers are assumed unrelated, these sibs cannot share both alleles IBD as they share no paternal alleles IBD). The probability that $o_1$ and $o_2$ both receive the same allele from their mother is one-half (because whichever allele the female passes to $o_1$, the probability that she passes the same allele to $o_2$ is one-half). In this case, the two offspring have one allele IBD, and the contribution to the genetic covariance when this occurs is $Cov(\alpha_1, \alpha_1) = Var(A)/2$. When $o_1$ and $o_2$ share no alleles IBD, they have no genetic covariance.

Summarizing:

<table>
<thead>
<tr>
<th>Case</th>
<th>Probability</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o_1$ and $o_2$ have 0 alleles IBD</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>$o_1$ and $o_2$ have 1 allele IBD</td>
<td>1/2</td>
<td>$Var(A)/2$</td>
</tr>
</tbody>
</table>

giving the genetic covariance between half sibs as

\[Cov(G_{o_1}, G_{o_2}) = Var(A)/4\]  

(1.22)

**Full-Sibs.**

Both parents are in common,

\[\begin{array}{c}
\varnothing \\
O_1 \quad O_2
\end{array}\]

Lecture 1, pg. 12
What is the covariance of genotypic values of two full sibs?

Three cases are possible when considering pairs of full sibs: they can share either 0, 1, or 2 alleles IBD. Applying the same approach as for half sibs, if we can compute: 1) the probability of each case; and 2) the contribution to the genetic covariance for each case.

Each full sib receives one paternal and one maternal allele. The probability that each sib receives the same paternal allele is 1/2, which is also the probability each sib receives the same maternal allele. Hence,

\[ \Pr(2 \text{ alleles IBD}) = \Pr(\text{paternal allele IBD}) \Pr(\text{maternal allele IBD}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \]

\[ \Pr(0 \text{ alleles IBD}) = \Pr(\text{paternal allele not IBD}) \Pr(\text{maternal allele not IBD}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \]

\[ \Pr(1 \text{ allele IBD}) = 1 - \Pr(2 \text{ alleles IBD}) - \Pr(0 \text{ alleles IBD}) = \frac{1}{2} \]

We saw above that when two relatives share one allele IBD, the contribution to the genetic covariance is \( Var(A) / 2 \). When two relatives share both alleles IBD, each has the same genotype at the locus being considered, and the contribution is

\[ \text{Cov}(\alpha_1 + \alpha_2 + \delta_{12}, \alpha_1 + \alpha_2 + \delta_{12}) = Var(\alpha_1 + \alpha_2 + \delta_{12}) = Var(A) + Var(D) \]

Putting these results together gives

<table>
<thead>
<tr>
<th>Case</th>
<th>Probability</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>( o_1 ) and ( o_2 ) have 0 alleles IBD</td>
<td>1/4</td>
<td>0</td>
</tr>
<tr>
<td>( o_1 ) and ( o_2 ) have 1 allele IBD</td>
<td>1/2</td>
<td>( Var(A)/2 )</td>
</tr>
<tr>
<td>( o_1 ) and ( o_2 ) have 2 alleles IBD</td>
<td>1/4</td>
<td>( Var(A) + Var(D) )</td>
</tr>
</tbody>
</table>

This results in a genetic covariance between full sibs of

\[ \text{Cov}(G_{o_1}, G_{o_2}) = \frac{1}{2} \cdot \frac{Var(A)}{2} + \frac{1}{4} (Var(A) + Var(D)) = \frac{Var(A)}{2} + \frac{Var(D)}{4} \]  \( (1.23) \)

**General relationships.**

Equations 1.20a and 1.20b suggest a general expression for the covariance between (noninbred) relatives, based on the probabilities that they share one and both alleles IBD.

If \( r_{xy} = (1/2) \) Prob(relatives \( x \) and \( y \) have one allele IBD) + Prob(relatives \( x \) and \( y \) have both alleles IBD), and \( u_{xy} = \) Prob(relatives \( x \) and \( y \) have both alleles IBD), then the genetic covariance between \( x \) and \( y \) is given by

\[ \text{Cov}(G_x, G_y) = r_{xy} Var_A + u_{xy} Var_D \]  \( (1.24a) \)

If epistatic genetic variance is present, this can be generalized to

\[ \text{Cov}(G_x, g_y) = r_{xy} Var_A + u_{xy} Var_D + r_{xy}^2 Var_{AA} + r_{xy} u_{xy} Var_{AD} + u_{xy}^2 Var_{DD} + \cdots \]  \( (1.24b) \)
Components of Environmental Variances

Just as we decomposed the total genotypic value into components, some shared, others not transmitted between relatives, we can do the same for environmental effects. For example, plants may be distributed within different plots within different fields, within different locations (e.g., different climatic regions) within different growing seasons (e.g., 2007 vs. 2008). This decomposition of the environmental variance is especially important when pure lines are used in cultivar trails, as the genotypes of all individuals within a line are the same, and these can be replicated over different times and places. Besides plot and field variances, breeders often focus on three specific components for a given pure line:

- $E_L$: Shared effects due to sharing the location
- $E_Y$: Shared effects from the year
- $E_{L \times Y}$: Location by year interaction

Lines showing small location and/or year variances tend to be more stable producers over time, but lines showing large location effects may be the highest producers in a particular environment, even more so than a stable line.

Narrow-vs. Broad-sense Heritability

One of the key parameters in quantitative genetics is the heritability $h^2$. The reason for this focus, indeed obsession, on the heritability is that it determines the degree of resemblance between parents and offspring, which in turn determines the response to selection. In particular, the slope of a midparent-offspring regression is just $h^2 = V_A/V_P$. The fact that the regression involves midparents implies sexual reproduction. In many plant breeding settings, the parent-offspring regression involves offspring that are asexual clones of the single parent. In this case, the parent-offspring regression has slope given by the broad-sense heritability, $H^2 = V_G/V_P$.

When we refer to heritability (without making use of either $h^2$ or $H^2$), we are by default referring to the narrow-sense heritability $h^2$. Use of the broad-sense heritability $H^2$ is generally restricted to discussions of clones (individuals from a pure = fully-inbred line). While $H^2$ also gives the total fraction of variation in a trait due to differences in genotypic values, for sexually reproducing species only variation in breeding values is (easily) converted into selection response. Hence, $h^2$ rather than $H^2$ is a better measure for sexual species of the fraction of (easily) usable genetic variance.

Defining $H^2$ for Plant Populations

Our last point about heritability deals with how plant breeders define the broad-sense heritability $H^2$. Our key point is that identical populations may have different $H^2$ values, depending on the unit of analysis chosen by the investigator.

In plant breeding, pure lines are often used, and instead of measuring individuals directly for trait values (such as yield), one often measures a block or plot of individuals as the sampling unit. Suppose our design is to measure each line in $r$ plots, each consisting of $n$ individuals, over $e$ environments. The resulting linear model for the $\ell$th individual of $i$ genotype in plot $k$ in environment $j$

$$z_{ijkt} = G_i + E_j + E_{i}^{G} + p_{ijk} + e_{ijk}$$

(4.20)

where $p_{ijk}$ is the plot effect for the $k$th replicate of the plot for genotype $i$ in environment $j$, and $e_{ijk}$ the residual value for the $\ell$th individual. If we simply take $z_i = z_{i \cdots}$ (the average value of genotype $i$ over all plots, environments, and individuals) as our measurement then $\sigma^2_G$ is unchanged, but the phenotypic variances of the $z_i$ becomes

$$\sigma^2(z_i) = \sigma^2_G + \sigma^2_E + \frac{\sigma^2_{G \times E}}{e} + \frac{\sigma^2_p}{e r} + \frac{\sigma^2_e}{e r n}$$
here $\sigma_p^2$ is the between-plot environmental variance and $\sigma_e^2$ the within-plot individual variance. The problem with defining $H^2$ in a consistent fashion is that different investigators may chose different values of $r$, $e$ and $n$ to measure the trait, and hence (even with identical variance components) get different values for the phenotypic variance, and hence for $H^2$.

**General and Specific Combining Abilities: GCA, SCA**

When dealing with pure (i.e., inbred) lines, one often sees heterosis when two lines are crossed (Lecture 5). Here, the $F_1$ mean from the cross is larger than the mean value of the two parental lines. When dealing with a collection of pure lines, and in deciding which to cross, we often consider the **general and specific combining abilities** of these lines. The **GCA** is akin to the average effect of an allele — it is the mean value of all crosses over our collection using a particular line. Thus, if line 1 has a GCA of 15, then the average mean from all crosses involving line 1 if 15 units over the mean value of all lines. The **SCA** is a function of two specific crosses and is akin to the dominance deviation. Suppose line 2 has a GCA of -5. We then expect the mean from a line 1 by line 2 cross is 15-5 = 10 units over the mean. However, the specific value for that cross is the value actually seen, and suppose this is 20. The SGA would then be the true value minus the values predict from the GCAs, SCA(1,2) = 20 - GCA(1)-GCA(2) = 20-15 + 5 = 10.