

# Lecture 2

## Fisher's Variance Decomposition

Bruce Walsh. July 2005. Asian Institute on Statistical Genetics

### Covariances and Regressions

Quantitative genetics requires measures of variation and association. Thus we introduce some standard statistical measures of association (covariances, correlations, and regressions) and variation (variances).

#### The Variance:

The standard measure of variation is the **variance**,

$$Var(x) = E[(x - \mu_x)^2]$$

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 the 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

$$Var(x) = E[x^2] - \mu_x^2$$

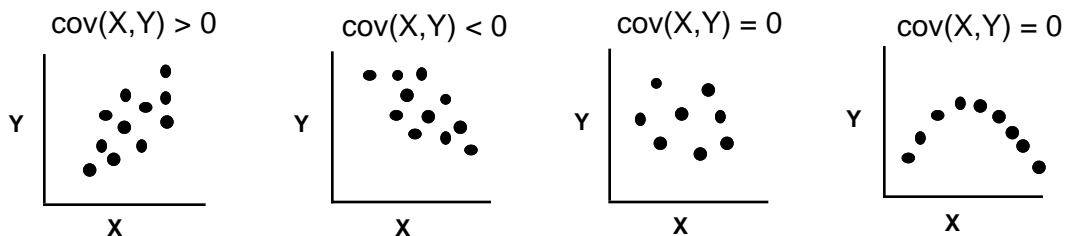
Note that if the mean is zero, then  $Var(x) = E[x^2]$ .

#### 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

$$\begin{aligned} Cov(x, y) &= E[(x - \mu_x) * (y - \mu_y)] \\ &= E(xy) - \mu_x \mu_y \\ &= \text{mean of the product} - \text{product of the means} \end{aligned} \tag{2.1}$$

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  $y$  in 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 covariance is estimated for a sample of  $n$  paired observations  $(x_i, y_i)$  by

$$\begin{aligned} Cov(x, y) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right) \end{aligned} \quad (2.2)$$

[In the literature,  $\sigma(x, y) = \sigma_{xy}$  is often used to denote the population covariance (Equation 2.1), while  $Cov(x)$  denotes its estimated value (Equation 2.2). In these notes, we tend to use  $Cov$  interchangeably for both, although an occasional  $\sigma$  may appear. ]

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

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x) Var(y)}} \quad (2.3)$$

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

$$b_{y|x} = \frac{Cov(x, y)}{Var(x)} \quad (2.4a)$$

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

$$\hat{y} = \bar{y} + b_{y|x}(x - \bar{x}) \quad (2.4b)$$

Correlations and regression slopes are related as follows:

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} = \frac{Cov(x, y)}{Var(x)} \sqrt{\frac{Var(x)}{Var(y)}} = b_{y|x} \sqrt{\frac{Var(x)}{Var(y)}} \quad (2.4c)$$

Thus, if the variances of  $x$  and  $y$  are the same, then  $r(x, y) = b_{y|x} = b_{x|y}$ .

### Useful Properties of Variances and Covariances:

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

$$Var(ax) = Cov(ax, ax) = a^2 Cov(x, x) = a^2 Var(x)$$

- $Cov(x + y, z) = Cov(x, z) + Cov(y, z)$ , i.e., the covariance of a sum is the sum of covariances. More generally,

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

- $Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$ . Hence, the variance of a sum,  $Var(x + y)$ , equals the sum of the variances,  $Var(x) + Var(y)$ , 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**  $E$ ,

$$P = G + E \quad (2.5a)$$

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. While it is often assumed that the genetic and environmental values are uncorrelated, this need not be the case. For example, a genetically higher-yielding dairy cow may also be fed more, creating a positive correlation between  $G$  and  $E$ . In such cases the basic model becomes

$$P = G + E + Cov(G, E) \quad (2.5b)$$

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:

	Genotypes		
	$Q_1Q_1$	$Q_1Q_2$	$Q_2Q_2$
Average Trait Value:	$C$	$C + a(1 + k)$	$C + 2a$
	$C$	$C + a + d$	$C + 2a$
	$C - a$	$C + d$	$C + a$

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

$$a = [G(Q_2Q_2) - G(Q_1Q_1)]/2 \quad (2.6a)$$

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

$$d = G(Q_1Q_2) - \frac{G(Q_1Q_1) + G(Q_2Q_2)}{2} \quad (2.6b)$$

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  $d = -a$  ( $k = -1$ ) then  $Q_1$  is dominant to  $Q_2$ . Finally if  $d > a$  ( $k > 1$ ) the locus shows **overdominance** with the heterozygote having a larger value than either homozygote. Thus  $d$  (and equivalently  $k$ ) measure the amount of dominance at this locus. Note that  $d$  and  $k$  are related by

$$ak = d, \quad \text{or} \quad k = \frac{d}{a} \quad (2.6c)$$

The reason for using both  $d$  and  $k$  is that some expressions are simpler using one parameterization over another.

### Example 2.1: The Booroola ( $B$ ) gene

The Booroola ( $B$ ) gene influences fecundity (offspring number) in the Merino sheep of Australia. The mean litter sizes for the  $bb$ ,  $Bb$ , and  $BB$  genotypes based on 685 total records are 1.48, 2.17, and 2.66, respectively. Taking these to be estimates of the genotypic values ( $G_{bb}$ ,  $G_{Bb}$ , and  $G_{BB}$ ),

$$a = (2.66 - 1.48)/2 = 0.59, \quad d = 2.17 - (1.48 + 2.66)/2 = 0.10$$

This value of  $d$  suggests slight dominance of the Booroola gene. Using the alternative  $k$  notation, from Equation 2.6c,  $k = d/a = 0.17$ .

### 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. This insight is more fully developed below. 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 develop this point in the next lecture.

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,

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij} \quad (2.7)$$

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 2.7 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 2.7 as

$$G_{ij} = a + bN + e \quad (2.8a)$$

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

$$a = \mu_G + 2\alpha_1 \quad b = \alpha_2 - \alpha_1, \quad e = \delta_{ij} \quad (2.8b)$$

Note that

$$2\alpha_1 + (\alpha_2 - \alpha_1)N = \begin{cases} 2\alpha_1 & \text{for } N = 0, \text{ e.g. } Q_1Q_1 \\ \alpha_1 + \alpha_2 & \text{for } N = 1, \text{ e.g. } Q_1Q_2 \\ 2\alpha_2 & \text{for } N = 2, \text{ e.g. } Q_2Q_2 \end{cases} \quad (2.9)$$

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

Genotypes:	$Q_1Q_1$	$Q_1Q_2$	$Q_2Q_2$
Average Trait Value:	0	$a(1+k)$	$2a$
frequency (HW):	$p_1^2$	$2p_1p_2$	$p_2^2$

A little algebra gives

$$\mu_G = 2p_1p_2a(1+k) + 2p_2^2a = 2p_2a(1+p_1k) \quad (2.10a)$$

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

$$\alpha_2 - \alpha_1 = \frac{\sigma(G, N_2)}{\sigma^2(N_2)} = a[1+k(p_1-p_2)] \quad (2.10b)$$

See Lynch and Walsh, Chapter 4 for the algebraic details leading to Equation 2.10b. 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 2.10b this implies (again, see L & W Chapter 4)

$$\alpha_2 = p_1a[1+k(p_1-p_2)] \quad (2.10c)$$

$$\alpha_1 = -p_2a[1+k(p_1-p_2)] \quad (2.10d)$$

Finally, the dominance deviations follow since

$$\delta_{ij} = G_{ij} - \mu_G - \alpha_i - \alpha_j \quad (2.10e)$$

Note the important point that both  $\alpha$  and  $\delta$  are functions of allele frequency 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 and Breeding Values

The  $\alpha_i$  value is the **average effect** of allele  $Q_i$ . Breeders are concerned (indeed obsessed) with the **breeding values** (BV) of individuals, which are related to average effects. (The BV is also called the **additive genetic value**,  $A$ .) The BV associated with genotype  $G_{ij}$  is just

$$BV(G_{ij}) = \alpha_i + \alpha_j \quad (2.11a)$$

Likewise, for  $n$  loci underlying the trait, the BV is just

$$BV = \sum_{k=1}^n (\alpha_i^{(k)} + \alpha_k^{(k)}) \quad (2.11b)$$

namely, the sum of all of the average effects of the individual's alleles. Note that since the BVs are functions of the allelic effects, they change as the allele frequencies in the population change.

So, why all the fuss over breeding/additive-genetic values? Consider the offspring from the cross of a sire (genotype  $Q_xQ_y$ ) mated to a number of unrelated dams (let the genotype of one of these random dams be  $Q_wQ_z$  where  $w$  and  $z$  denote randomly-chosen alleles.) Since each parent passes along one of its two alleles, there are four equally-frequent offspring:

Genotype	Frequency	Value
$Q_x Q_w$	1/4	$\mu_G + \alpha_x + \alpha_w + \delta_{xw}$
$Q_x Q_z$	1/4	$\mu_G + \alpha_x + \alpha_z + \delta_{xz}$
$Q_y Q_w$	1/4	$\mu_G + \alpha_y + \alpha_w + \delta_{yw}$
$Q_y Q_z$	1/4	$\mu_G + \alpha_y + \alpha_z + \delta_{yz}$

The average value of the offspring thus becomes

$$\mu_O = \mu_G + \left( \frac{\alpha_x + \alpha_y}{2} \right) + \left( \frac{\alpha_w + \alpha_z}{2} \right) + \frac{\delta_{xw} + \delta_{xz} + \delta_{yw} + \delta_{yz}}{4}$$

Taking the average of this expression over the random collection of dams (the sire alleles  $x$  and  $y$  remain constant, but dam alleles  $y$  and  $w$  are random), the last two expressions (the average effects of the dams and the dominance deviations) have expected values of zero. Hence, the expected value for the offspring of the sire becomes

$$\mu_O - \mu_G = \left( \frac{\alpha_x + \alpha_y}{2} \right) = \frac{BV(\text{Sire})}{2} \quad (2.12a)$$

Thus one (simple) estimate of the sire's BV is just twice the deviation from its offspring and overall population mean,

$$BV(\text{Sire}) = 2(\mu_O - \mu_G) \quad (2.12b)$$

Similarly, the expected value of the offspring given the breeding values of both parents is just their average,

$$\mu_O - \mu_G = \frac{BV(\text{Sire})}{2} + \frac{BV(\text{Dam})}{2} \quad (2.12c)$$

The focus on breeding values thus arises because they predict offspring means — the largest expected offspring mean is generated by crossing the parents with the largest breeding values.

## 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,

$$\sigma^2(G) = \sigma^2(\mu_g + (\alpha_i + \alpha_j) + \delta_{ij}) = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij}) \quad (2.13)$$

Equation 2.13 is the contribution from a single locus. Assuming linkage equilibrium, we can sum over loci,

$$\sigma^2(G) = \sum_{k=1}^n \sigma^2(\alpha_i^{(k)} + \alpha_j^{(k)}) + \sum_{k=1}^n \sigma^2(\delta_{ij}^{(k)})$$

This is usually written more compactly as

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 \quad (2.14)$$

where  $\sigma_A^2$  is the **additive genetic variance** and represents the variance in breeding values in the population, while  $\sigma_D^2$  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

$$\sigma_A^2 = E[\alpha_i^2 + \alpha_j^2] = 2E[\alpha^2] = 2 \sum_{i=1}^m \alpha_i^2 p_i, \quad \text{and} \quad \sigma_D^2 = E[\delta^2] = \sum_{i=1}^m \sum_{j=1}^m \delta_{ij}^2 p_i p_j \quad (2.15)$$

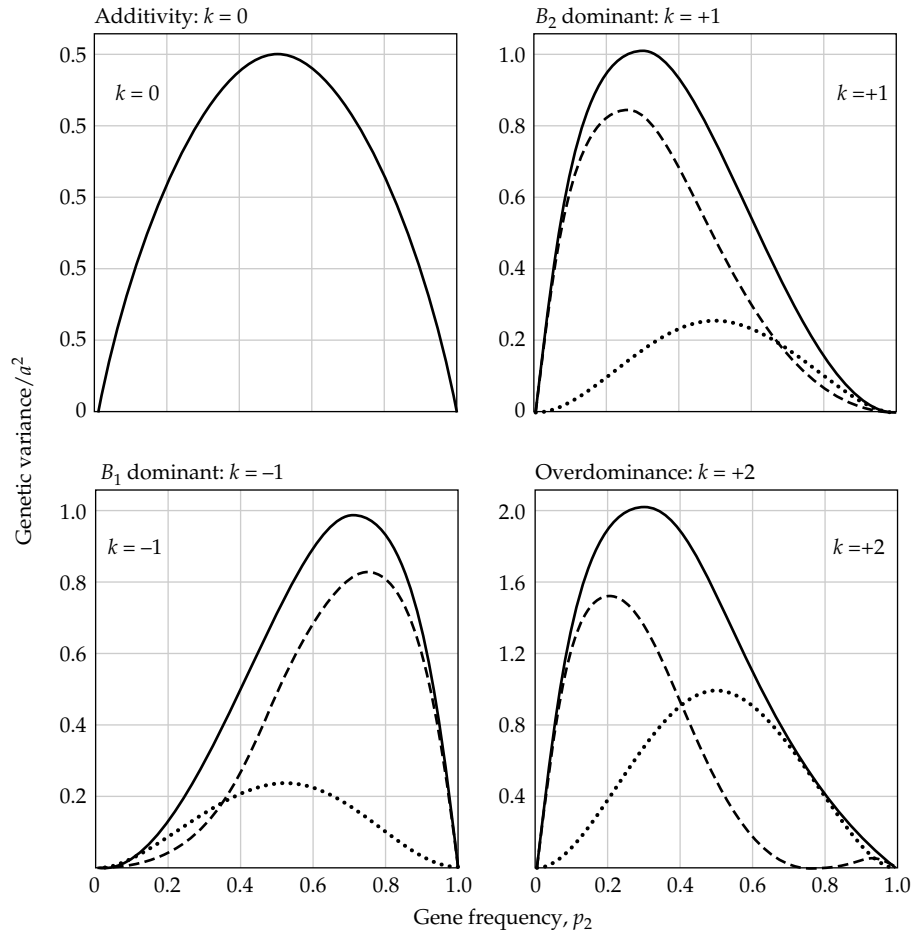
For one locus with two alleles, these become

$$\sigma_A^2 = 2p_1 p_2 a^2 [1 + k(p_1 - p_2)]^2 = 2p_1 p_2 [a + d(p_1 - p_2)]^2 \quad (2.16a)$$

and

$$\sigma_D^2 = (2p_1 p_2 ak)^2 = (2p_1 p_2 d)^2 \quad (2.16b)$$

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 2.16) 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 (or higher) locus genotypic values. 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

$$\begin{aligned}
 G_{ijkl} &= \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kj}) \\
 &\quad + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\
 &\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl} + \alpha\delta_{kij} + \alpha\delta_{lij}) \\
 &\quad + (\delta\delta_{ijkl}) \\
 &= \mu_G + A + D + AA + AD + DD
 \end{aligned} \tag{2.17}$$

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  $B_j B_k$ ), 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 2.17 are uncorrelated, so that we can write the genetic variance as

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 \tag{2.18}$$

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

$$\sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 + \sigma_{AAA}^2 + \sigma_{AAD}^2 + \sigma_{ADD}^2 + \sigma_{DDD}^2$$



## Lecture 2 Problems

1. Consider the Booroola gene from Example 2.1.
  - a. For  $\text{freq}(B) = 0.3$ , compute  $\alpha_B$ ,  $\alpha_b$ , and the breeding values of all three genotypes.
  - b. For  $\text{freq}(B) = 0.8$ , compute  $\alpha_B$ ,  $\alpha_b$ , and the breeding values of all three genotypes.
2. For the above two frequencies for Booroola, compute  $\sigma_G^2$ ,  $\sigma_A^2$ , and  $\sigma_D^2$
3. What is the covariance between an individual's breeding value  $A$  and its phenotypic value  $P$ ? (Assume  $\text{Cov}(G, E) = 0$ .) Hint, use the properties of the covariance and decompose  $P$  into its various genetic and environmental components.
4. What is the best linear predictor of an individual's breeding value  $A$  given that we observe their phenotypic value  $P$ ?

## Solutions to Lecture 2 Problems

1. For Booroola,  $a = 0.59$ ,  $k = 0.17$ . In our notation,  $p_2 = \text{freq}(B)$

a. For  $p_2 = \text{freq}(B) = 0.3$ ,  $p_1 = \text{freq}(b) = 0.7$

$$\alpha_2 = \alpha_B = p_1 a [1 + k(p_1 - p_2)] = 0.7 \cdot 0.59 [1 + 0.17(0.7 - 0.3)] = 0.441$$

$$\alpha_1 = \alpha_b = -p_2 a [1 + k(p_1 - p_2)] = -0.189$$

$$BV(BB) = 2\alpha_B = 0.882, \quad BV(Bb) = \alpha_B + \alpha_b = 0.252, \quad BV(BBb) = 2\alpha_b = -0.378,$$

b. For  $\text{freq}(B) = 0.8$ ,

$$\alpha_B = 0.106, \quad \alpha_b = -0.423$$

$$BV(BB) = 2\alpha_B = 0.211, \quad BV(Bb) = \alpha_B + \alpha_b = -0.318, \quad BV(BBb) = 2\alpha_b = -0.848,$$

2. a For  $p_2 = \text{freq}(B) = 0.3$

$$\sigma_A^2 = 2p_1 p_2 a^2 [1 + k(p_1 - p_2)]^2 = 0.167$$

$$\sigma_D^2 = (2p_1 p_2 a k)^2 = 0.002, \quad \sigma_G^2 = \sigma_A^2 + \sigma_D^2 = 0.169$$

b For  $p_2 = \text{freq}(B) = 0.8$

$$\sigma_A^2 = 0.090, \quad \sigma_D^2 = 0.001, \quad \sigma_G^2 = 0.091$$

3.  $Cov(P, A) = Cov(G + E, A) = Cov(A + D + E, A) = Cov(A, A) = Var(A)$

4. The regression is  $A = \mu_A + b_{A|P}(P - \mu_P)$ . The slope is

$$b_{A|P} = \frac{Cov(P, A)}{V_P} = \frac{Cov(A, A)}{V_P} = \frac{Var(A)}{V_P} = h^2$$

Hence,  $A = h^2(P - \mu_P)$  as the mean breeding value (by construction) is zero, i.e.,  $\mu_A = 0$