Walsh Lecture 1
Fisher's Variance Decomposition and the Resemblance Between Relatives

Bruce Walsh. 25 Jan - 5 Feb 2016

Second Bangalore School on Population Genetics and Evolution

I: Fisher's Variance Decomposition

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 its mean (μ_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

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

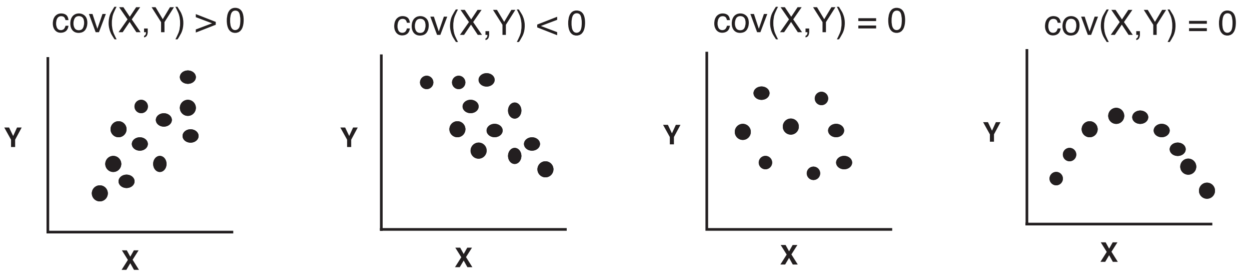
If the mean is zero, then $Var(x) = E[x^2]$.

The Covariance:

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

$$\begin{aligned} Cov(x, y) &= E[(\, x - \mu_x \,) * (\, y - \mu_y \,)] \\ &= E(\, xy \,) - \mu_x \mu_y \\ &= \text{mean of the product} - \text{product of the means} \qquad (1.1) \end{aligned}$$

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 *linear* association — 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 covariance is estimated for a sample of n paired observations (x_i, y_i) by

$$\begin{aligned} Cov(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}(\,x_i-\overline{x}\,)(\,y_i-\overline{y}\,) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i\,y_i - n\,\overline{x}\,\overline{y}\right) \end{aligned} \tag{1.2}$$

In the literature, $\sigma(x,y) = \sigma_{xy}$ is often used to denote the population covariance (Equation 1.1), while $Cov(x)$ denotes its estimated value (Equation 1.2). In these notes, we tend to use Cov interchangeably for both, although an occasional σ may appear. Likewise, the variance of x can be denoted as $Var(x)$, V_x, and σ_x^2.

The **correlation**, $r(x,y)$ is a scaled measure of the covariance, where

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

The notation $\rho(x,y)$ and ρ_{xy} is also used. 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:

Consider the best linear fit of some **response** (or **dependent**) y given a **predictor** (or **independent**) variable x,

$$y = a + b_{y\,|\,x}x + e = \widehat{y} + e \tag{1.4a}$$

where

$$\widehat{y} = a + b_{y\,|\,x}x \tag{1.4b}$$

is the best linear predictor of y given a specific value of x, and e is the residual error. A linear regression passes through the means ($\overline{x}$, $\overline{y}$) of the response and predictor variables, giving

$$y - \overline{y} = b_{y\,|\,x}(x - \overline{x}) + e \tag{1.4c}$$

The slope $b_{y\,|\,x}$ is given by

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

Correlations and regression slopes are related as follows:

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

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 $Cov(ax,y) = a \cdot Cov(x,y)$
- $Var(ax) = a^2Var(x)$. This follows since
  $Var(ax) = Cov(ax,ax) = a^2Cov(x,x) = a^2Var(x)$

- $Cov(x+y,z) = Cov(x,z) + Cov(y,z)$, 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 of quantitative genetics is that the **phenotypic value** P of a trait is the sum of a **genetic value** G plus an **environmental value** E,

$$P = G + E \tag{1.5a}$$

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. More generally, there can also be **genotype-by-environment interactions**, with the base model now being

$$P = G + E + GE \tag{1.5b}$$

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 diallelic 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 often set equal to zero. What matters is the difference $2a$ between the two homozygotes,

$$a = [\mathrm{G}(Q_2Q_2) - \mathrm{G}(Q_1Q_1)]/2 \tag{1.6a}$$

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

$$d = \mathrm{G}(Q_1Q_2) - \frac{\mathrm{G}(Q_1Q_1) + \mathrm{G}(Q_2Q_2)}{2} \tag{1.6b}$$

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 $d = -a$ ($k = -1$) 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 k measure the amount of dominance at this locus, and are related by

$$ak = d, \qquad \text{or} \quad k = \frac{d}{a} \tag{1.6c}$$

The reason for using both the d and k notation for the amount of dominance is that some expressions are simpler using one parameterization over another.

Example 1.1: 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 1.6c, $k = d/a = 0.17$.

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 later in this 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** α for each allele and a **dominance deviation** δ giving the deviation of the actual value for this genotype from the value predicted by the average contribution of each of the single alleles,

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij} \tag{1.7a}$$

The predicted genotypic value is

$$\widehat{G}_{ij} = \mu_G + \alpha_i + \alpha_j \tag{1.7b}$$

where μ_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 $G_{ij} - \widehat{G}_{ij} = \delta_{ij}$, so that δ is the residual error, the difference between the actual value and that predicted from the regression. Since α and δ represent deviations from the overall mean, they also have expected values of zero.

You might notice that Equation 1.7a looks like a regression, as $G_{ij} = \widehat{G}_{ij} + e$. Indeed it is, as we can express Equation 1.7a as

$$G_{ij} = a + bN + e \tag{1.8a}$$

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

$$a = \mu_G + 2\alpha_1 \qquad b = \alpha_2 - \alpha_1, \qquad e = \delta_{ij} \tag{1.8b}$$

Note that

$$2\alpha_1 + (\alpha_2 - \alpha_1)N = \begin{cases} 2\alpha_1 & \text{for } N = 0, \text{ e.g, } Q_1Q_1 \\ \alpha_1 + \alpha_2 & \text{for } N = 1, \text{ e.g, } Q_1Q_2 \\ 2\alpha_2 & \text{for } N = 2, \text{ e.g, } Q_2Q_2 \end{cases} \tag{1.9}$$

Thus we have a regression, where N (the number of copies of allele Q_2) is the independent variable, the genotypic value G the dependent variable, $(\alpha_2 - \alpha_1)$ is the regression slope, and the δ_{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 (δ_{ij}) is zero, so that $\sigma(\alpha_i, \delta_j) = \sigma(\alpha_k, \delta_j) = 0$.

To obtain the α, μ_G, and δ 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

where p_i =freq(Q_i), and (for two alleles) $p_1 = 1 - p_2$. A little algebra gives

$$\mu_G = 2\,p_1\,p_2\,a(1+k) + 2\,p_2^2\,a = 2\,p_2\,a(1+p_1k) \tag{1.10a}$$

Recall that the slope of a regression is simply the covariance divided by the variance of the predictor variable (Equation 1.4d), giving

$$\alpha_2 - \alpha_1 = \frac{\sigma(G, N_2)}{\sigma^2(N_2)} = a\,[\,1 + k\,(\,p_1 - p_2\,)\,] \tag{1.10b}$$

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

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

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

$$\alpha_2 = p_1 a\,[\,1 + k\,(\,p_1 - p_2\,)\,] \tag{1.10c}$$

$$\alpha_1 = -p_2 a\,[\,1 + k\,(\,p_1 - p_2\,)\,] \tag{1.10d}$$

The average effect α_i of allele Q_i can be thought of as how much a random individual that receives a copy of Q_i from one parent and a random allele from its other parent deviates from the mean.

Finally, the dominance deviations follow since

$$\delta_{ij} = G_{ij} - \mu_G - \alpha_i - \alpha_j \tag{1.10e}$$

Note the important point that both α and δ 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

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

$$BV(G_{ij}) = \alpha_i + \alpha_j \tag{1.11a}$$

Likewise, for n loci underlying the trait, the BV becomes

$$BV = \sum_{k=1}^{n} \left(\alpha_i^{(k)} + \alpha_j^{(k)} \right) \tag{1.11b}$$

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_xQ_w	1/4	$\mu_G + \alpha_x + \alpha_w + \delta_{xw}$
Q_xQ_z	1/4	$\mu_G + \alpha_x + \alpha_z + \delta_{xz}$
Q_yQ_w	1/4	$\mu_G + \alpha_y + \alpha_w + \delta_{yw}$
Q_yQ_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 terms (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

$$\mu_O - \mu_G = \left(\frac{\alpha_x + \alpha_y}{2}\right) = \frac{BV(\text{Sire})}{2} \tag{1.12a}$$

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

$$BV(\text{Sire}) = 2(\mu_0 - \mu_G) \tag{1.12b}$$

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

$$\mu_0 - \mu_G = \frac{BV(\text{Sire})}{2} + \frac{BV(\text{Dam})}{2} \tag{1.12c}$$

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 δ corresponds to a residual. Recall from regression theory that the estimated value and its residual are uncorrelated, and hence α and δ are uncorrelated. Since μ_G is a constant (and hence contributes nothing to the variance), and α and δ are uncorrelated, we have

$$\sigma^2(G) = \sigma^2(\mu_g + (\alpha_i + \alpha_j) + \delta_{ij}) = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij}) \tag{1.13}$$

Equation 1.13 is the contribution from a single locus. Assuming linkage equilibrium (that $\sigma(\alpha_i^{(k)}, \alpha_j^{(\ell)}) = 0$ for all loci $k \neq \ell$), 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

$$\sigma^2_G = \sigma^2_A + \sigma^2_D \tag{1.14}$$

where σ^2_A is the **additive genetic variance** and represents the variance in breeding values in the population, while σ^2_D denotes the **dominance genetic variance** and is the variance in dominance deviations. These are also denoted $Var(A), V_A$ and $Var(D), V_D$.

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

$$\sigma_A^2 = E[\alpha_i^2 + \alpha_j^2\,] = 2E[\alpha^2\,] = 2\sum_{i=1}^{m} \alpha_i^2\, p_i, \qquad \text{and} \qquad \sigma_D^2 = E[\delta^2\,] = \sum_{i=1}^{m}\sum_{j=1}^{m} \delta_{ij}^2\, p_i\, p_j \tag{1.15}$$

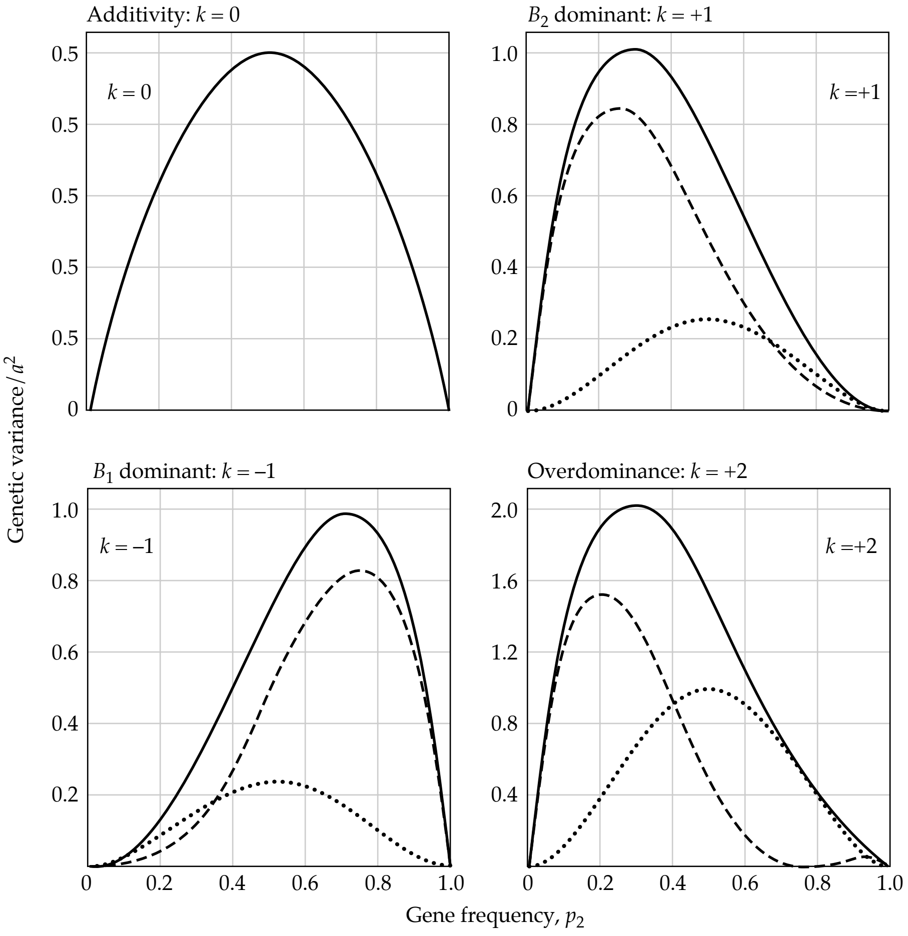
For one locus with two alleles, these become

$$\sigma_A^2 = 2p_1\, p_2\, a^2[\,1 + k\,(\,p_1 - p_2\,)\,]^2 = 2p_1\, p_2\,[\,a + d\,(\,p_1 - p_2\,)\,]^2 \tag{1.16a}$$

and

$$\sigma_D^2 = (2p_1\, p_2\, ak)^2 = (2p_1\, p_2\, d)^2 \tag{1.16b}$$

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 figure and from Equation 1.16) that there is plenty of additive variance (dashed line) even in the face of complete dominance ($k = \pm 1$). This is not surprising as the allelic effects α arise from the best-fitting line, which will accommodate some of the nonlinearity (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 σ_D^2 is symmetric in allele frequency, as $p_1 p_2 = p_1(1 - p_1)$ is symmetric about $1/2$.

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 (expressed as a deviation from the population mean) 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 genotypic value when epistasis is present is a straight-forward extension of the no-epistasis version. For two loci, the genotypic value is decomposed as

$$\begin{aligned} G_{ijkl} &= \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kj}) \\ &\quad + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\ &\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl} + \alpha\delta_{kij} + \alpha\delta_{lij}) \\ &\quad + (\delta\delta_{ijkl}) \\ &= \mu_G + A + D + AA + AD + DD \end{aligned} \tag{1.17}$$

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. Terms in Equation 1.17 are uncorrelated, so that we can write the genetic variance as

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DD} \tag{1.18}$$

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

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

II: Resemblance Between Relatives

We now apply the above results to express the phenotype resemblance (covariance) between relatives as a function of these genetic variances. This, in turn, allows us to estimate these variances simply from these phenotypic covariances. For example, the parent-offspring covariance is $\sigma^2_A/2$, so that twice this phenotypic covariance provides an estimate of the additive variance.

Heriability

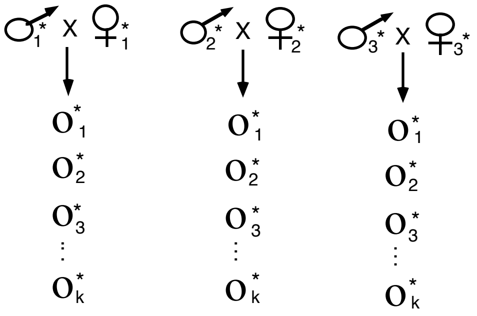
The **heritability** of a trait, a central concept in quantitative genetics, is the proportion of variation among individuals in a population that is due to variation in the additive genetic (i.e., breeding) values of individuals:

$$h^2 = \frac{\sigma^2_A}{\sigma^2_P} = \frac{\text{Variance of breeding values}}{\text{Phenotypic Variance}}$$

Since an individual's phenotype can be directly scored, the phenotypic variance σ^2_P can be estimated from measurements made directly on the population.

In contrast, an individual's breeding value cannot be observed directly, but rather must be inferred from the mean value of its offspring (or more generally using the phenotypic values of other known relatives). Thus, estimates of σ^2_A require known collections of relatives. The most common situations (which we focus on here) are comparisons between parents and their offspring or comparisons among sibs. We can classify relatives as either **ancestral** or **collateral**, and we focus here on designs with just one type of relative. In a more general pedigree, information from both kinds of relatives is present.

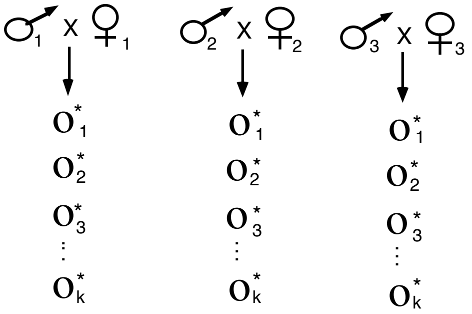
Ancestral relatives: e.g., parent and offspring



* Measure phenotypes of one or both parents, and k offspring

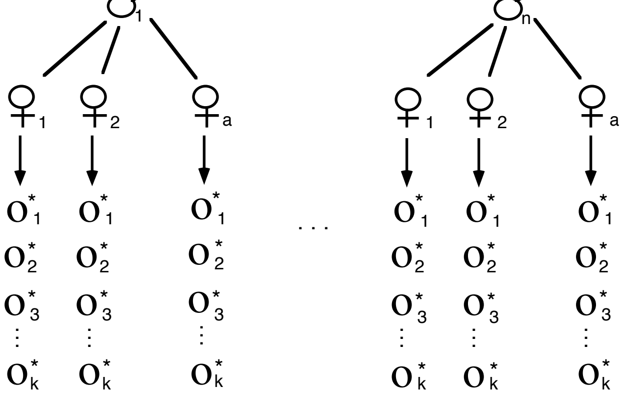
Collateral relatives:

Full Sibs have both parents in common



*Measure k offspring in each family, but not the parents.

Half Sibs have one parent in common



* Measure phenotype of k progeny of each family, but not the parents. Note that if $k > 1$, this design involves both full- (within any column) and half-sibs (between columns from the same sire) and is referred to as a **nested half-sib/full-sib design**.

Key observation: ***The amount of phenotypic resemblance among relatives for a trait provides an indication of the amount of genetic variation for that trait. If trait variation has a significant genetic basis, the closer the relatives, the more similar their appearance.***

Causes of Phenotypic Covariance Among Relatives

Relatives resemble each other for quantitative traits more than they do unrelated members from 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, $Cov(G_x, G_y)$, is the covariance of the genotypic values (G_x, G_y) of the related individuals x and y. We now show how the genetic covariances between parent and offspring, full sibs, and half sibs depend on the genetic variances σ_A^2 and σ_D^2.

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.

Consider the offspring of two (unrelated) parents and label the four allelic copies in the parents by 1 - 4, independent of whether or not any are identical in state.

$$\text{Parents: } A_1A_2 \times A_3A_4$$

$$\text{Offspring: } o_1 = A_1A_3 \quad o_2 = A_1A_4 \quad o_3 = A_1A_3 \quad o_4 = A_2A_4$$

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.

1. 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 δ_{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

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

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

$$\begin{aligned} Cov(G_o, G_p) =& Cov(\alpha_1, \alpha_1) + Cov(\alpha_1, \alpha_3) + Cov(\alpha_1, \delta_{13}) \\ &+ Cov(\alpha_2, \alpha_1) + Cov(\alpha_2, \alpha_3) + Cov(\alpha_2, \delta_{13}) \\ &+ Cov(\delta_{12}, \alpha_1) + Cov(\delta_{12}, \alpha_3) + Cov(\delta_{12}, \delta_{13}) \end{aligned}$$

By the way have (intentionally) constructed α and δ, 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,

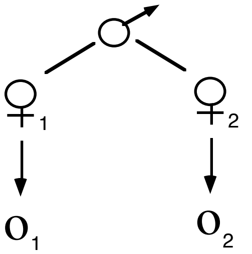
$$Cov(\delta_{xy}, \delta_{wz}) = \begin{cases} 0 & \text{if } xy \neq wz, \text{ i.e., both alleles are not IBD} \\ Var(D) & \text{if } xy = wz, \text{ both alleles are IBD} \end{cases} \tag{1.20b}$$

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

$$Cov(G_o, G_p) = Cov(\alpha_1, \alpha_1) = Var(A)/2 \tag{1.21}$$

2. Half-sibs.

Here, one parent is shared, the other is drawn at random from the population. For paternal half-sibs,



The genetic covariance between half-sibs is the covariance of the genetic values between o_1 and o_2.

To compute this, consider a single locus. First note that o_1 and o_2 share either one allele IBD (from the father) or no alleles IBD (since the mothers are assumed unrelated, these sibs cannot share both alleles IBD as they share no maternal alleles IBD). The probability that o_1 and o_2 both receive the same allele from the male is one-half (because whichever allele the male passes to o_1, the probability that he 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:

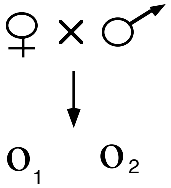
Case	Probability	Contribution
o_1 and o_2 have 0 alleles IBD	$1/2$	0
o_1 and o_2 have 1 allele IBD	$1/2$	$Var(A)/2$

giving the genetic covariance between half sibs as

$$Cov(G_{o_1}, G_{o_2}) = Var(A)/4 \tag{1.22}$$

3. Full-Sibs.

Both parents are in common,



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,

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

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

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

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

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

Putting these results together gives

Case	Probability	Contribution
o_1 and o_2 have 0 alleles IBD	1/4	0
o_1 and o_2 have 1 allele IBD	1/2	$Var(A)/2$
o_1 and o_2 have 2 allele IBD	1/4	$Var(A) + Var(D)$

This results in a genetic covariance between full sibs of

$$Cov(G_{o_1}, G_{o_2}) = \frac{1}{2}\,\frac{Var(A)}{2} + \frac{1}{4}\,(Var(A) + Var(D)) = \frac{Var(A)}{2} + \frac{Var(D)}{4} \qquad (1.23)$$

4. 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

$$Cov(G_x, G_y) = r_{xy}\sigma^2_A + u_{xy}\sigma^2_D \qquad (1.24a)$$

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

$$Cov(G_x, G_y) = r_{xy}\sigma^2_A + u_{xy}\sigma^2_D + r^2_{xy}\sigma^2_{AA} + r_{xy}u_{xy}\sigma^2_{AD} + u^2_{xy}\sigma^2_{DD} + \cdots \qquad (1.24b)$$

So that the coefficient on $\sigma^2_{A^m D^n}$ is $r^m_{xy}\, u^n_{xy}$

Environmental Causes of Relationship Between Relatives

Shared environmental effects (such as a common maternal environment) also contribute to the covariance between relatives, and care must be taken to distinguish sharded environmental effects from shared genetic effects.

If members of a family are reared together, they share a common environmental value, E_c. If the common environmental circumstances are different for each family, the variance due to common environmental effects, σ^2_{Ec}, causes greater similarity among members of a family, and greater differences among families, than would be expected just from the proportion of genes they share. Thus, σ^2_{Ec} inflates the phenotypic covariance of sibs over what is expected from their genotypic covariance.

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. In particular, we can write the total environmental effect E as the sum of a common environmental effect shared by the relatives E_c, a general environmental effect E_g, and a specific environmental effect E_s (unique to each indivdiual). Hence, we can write $E = E_c + E_g + E_s$, partitioning the environmental variance as

$$\sigma^2_E = \sigma^2_{Ec} + \sigma^2_{Eg} + \sigma^2_{Es} \tag{1.25}$$

We can further consider different possible sources of the common environmental effect E_c:

- E_{cS} or E_{cL}: Shared effects due to sharing the space/location (different farms, cages)
- E_{cT}: Temporal (changes in climactic or nutritional conditions over time)
- E_{cM}: Maternal (pre- and post-natal nutrition)

Thus, we can partition the environmental variance as

$$\begin{aligned}\sigma^2_E &= \sigma^2_{Ec} + \sigma^2_{Eg} + \sigma^2_{Es} \\ &= \sigma^2_{EcS} + \sigma^2_{EcT} + \sigma^2_{EcM} + \sigma^2_{Ec} + \sigma^2_{Eg} + \sigma^2_{Es}\end{aligned}$$

Common environment effects mainly contribute to resemblance of sibs, but maternal environment effects can contribute to resemblance between mother and offspring as well.

σ^2_{EcS} and σ^2_{EcT} can be eliminated, or estimated, by using the correct experimental design, but it is very difficult (except by cross-fostering) to eliminate or estimate σ^2_{EcM} from the covariance of full sibs. Further, cross-fostering only removes post-natal (past birth) maternal effects, it does not remove shared pre-natal maternal effects.

Phenotypic Covariance Among Relatives and h^2

Summarizing the above results, the resulting covariances between common sets of relatives, the associated regression slopes (b) or intra-class correlations (t), and how these relate to estmates of h^2 are as follows:

Relative Pair	Cov	t or b	h^2
Parent-offspring (P-O)	$\sigma^2_A/2$	$b_{O\|P} = \frac{1}{2}\,\sigma^2_A/\sigma^2_P$	$2b_{O\|P}$
Midparent-offspring (MP-O)	$\sigma^2_A/2$	$b_{O\|MP} = \sigma^2_A/\sigma^2_P$	$b_{O\|MP}$
Half-sibs (HS)	$\sigma^2_A/4$	$t_{HS} = (1/4)\sigma^2_A/\sigma^2_P$	$4t_{HS}$
Full-sibs (FS)	$\sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{Ec}$	$t_{FS} = \dfrac{\sigma^2_A/2 + \sigma^2_D/4 + \sigma^2_{Ec}}{\sigma^2_P}$	$2t_{FS} > h^2$

The midparent-offspring slope is computed as follows: using the properties of covariances,

$$\begin{aligned}Cov(O, MP) = Cov(0, [P_f + P_m]/2) &= \frac{Cov(0, P_f)}{2} + \frac{Cov(0, P_m)}{2} \\ &= \frac{\sigma^2_A/2}{2} + \frac{\sigma^2_A/2}{2} = \sigma^2_A/2\end{aligned}$$

The variance of the midparent values also follows from the properties of covariances, with

$$Var(MP) = Var\left(\frac{P_f + P_m}{2}\right) = \frac{Var(P_f)}{4} + \frac{Var(P_m)}{4} = \sigma^2_P/2$$

The last equality assumes equal trait variances in both parents and that parental values are uncorrelated (i.e., no assortative mating).

The regression slope equals the covariance between midparent and offspring divided by the midparent variance,

$$b_{O|MP} = \frac{Cov(O, MP)}{Var(MP)} = \frac{\sigma^2_A/2}{\sigma^2_P/2} = \frac{\sigma^2_A}{\sigma^2_P} = h^2$$

Thus, the slope of a midparent-offspring regression is the heritability h^2. Likewise, the slpe of a single-parent offsrping regression is $h^2/2$. Hence, ***parent-offspring regressions provide an easy way to estimate the heritability of a trait***.

Lecture 1 Problems

1. Consider the Booroola gene from Example 1.1.

   a. For freq$(B) = 0.3$, compute α_B, α_b, and the breeding values of all three genotypes.

   b. For freq$(B) = 0.8$, compute α_B, α_b, and the breeding values of all three genotypes.

2. For the above two frequencies for Booroola, compute σ^2_G, σ^2_A, and σ^2_D

3. What is the covariance between an individual's breeding value A and its phenotypic value P? 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?

5. The sm (small) allele is a bristle mutation that segregates in an Australian *Drosophila* population, where the genotypic values for the wildtype (+) and small (sm) alleles are as follows: $++ : +sm : sm\,sm$ have values of $44 : 40 : 22$. Suppose the environmental variance of bristle number is 6, and there are no common environmental effects due to maternal environment or rearing families together is vials. Assuming the sm locus is the only source of genetic variance, compute the regressions or intraclass correlations of bristle number between the following relatives:

   a: Offspring and midparent

   b: Half sibs

   c: Full sibs

   Perform these calculations for (i) populations where freq(sm) = 0.1 and (ii) populations where freq(sm) = 0.9.

Solutions to Lecture 1 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 $\text{freq}(B) = 0.8$,

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

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

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

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

$$\sigma^2_D = (2p_1 p_2 ak)^2 = 0.002, \quad \sigma^2_G = \sigma^2_A + \sigma^2_D = 0.169$$

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

$$\sigma^2_A = 0.090, \quad \sigma^2_D = 0.001, \quad \sigma^2_G = 0.091$$

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

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

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

5. Rescaling the genotypic values to $a : d : -a$ gives $11 : 7 : -11$, or $a = 11, d = 7$. Hence, the genetic variances are given by

$$\sigma^2_A = 2pq(a + d(q - p))^2 = 2pq(11 + 7(q - p))^2$$

and

$$\sigma^2_D = (2pqd)^2 = q^2 p^2\, 196$$

where $q = \text{freq}(sm)$. Hence:

q	σ^2_A	σ^2_D	σ^2_G
0.1	5.25	1.59	6.84
0.9	49.60	1.59	51.19

Here $\sigma^2_E = 6$, giving $\sigma^2_P = \sigma^2_G + 6$.

a) Parent-offspring regression: $b = \sigma^2_A/\sigma^2_p$, or 0.41 for $q = 0.1$, 0.87 for $q = 0.9$.

b) Half-sib correlation: $t_{HS} = (1/4)\sigma^2_A/\sigma^2_P$, or 0.10 for $q = 0.1$, 0.22 for $q = 0.9$.

c) Full-sib correlation: $t_{FS} = (\sigma^2_A/2 + \sigma^2_D/4)/\sigma^2_P$, or 0.24 for $q = 0.1$, 0.44 for $q = 0.9$.