

Lecture 2

Basic Population and Quantitative Genetics

Bruce Walsh. Aug 2003. Nordic Summer Course

Allele and Genotype Frequencies

The frequency p_i for allele A_i is just the frequency of A_iA_i homozygotes plus half the frequency of all heterozygotes involving A_i ,

$$p_i = \text{freq}(A_i) = \text{freq}(A_iA_i) + \frac{1}{2} \sum_{i \neq j} \text{freq}(A_iA_j) \quad (2.1)$$

The $1/2$ appears since only half of the alleles in heterozygotes are A_i . Equation 2.1 allows us to compute *allele* frequencies from *genotypic* frequencies. Conversely, since for n alleles there are $n(n+1)/2$ genotypes, the same set of allele frequencies can give rise to very different genotypic frequencies. To compute genotypic frequencies solely from allele frequencies, we need to make the (often reasonable) assumption of random mating. In this case,

$$\text{freq}(A_iA_j) = \begin{cases} p_i^2 & \text{for } i = j \\ 2p_i p_j & \text{for } i \neq j \end{cases} \quad (2.2)$$

Equation 2.2 is the first part of the **Hardy-Weinberg theorem**, which allows us (assuming random mating) to predict genotypic frequencies from allele frequencies. The second part of the Hardy-Weinberg theorem is that allele frequencies will remain unchanged from one generation to the next, *provided*: (1) infinite population size (i.e., no genetic drift), (2) no mutation, (3) no selection, and (4) no migration. Further, for an autosomal locus, a single generation of random mating gives genotypic frequencies in **Hardy-Weinberg proportions** (i.e., Equation 2) and the genotype frequencies forever remain in these proportions.

Gamete Frequencies, Linkage, and Linkage Disequilibrium

Random mating is the same as gametes combining at random. For example, the probability of an $AABB$ offspring is the chance that an AB gamete from the father and an AB gamete from the mother combine. Under random mating,

$$\text{freq}(AABB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \quad (2.3a)$$

For heterozygotes, there may be more than one combination of gametes that gives rise to the same genotype,

$$\text{freq}(AaBB) = \text{freq}(AB|\text{father}) \cdot \text{freq}(aB|\text{mother}) + \text{freq}(aB|\text{father}) \cdot \text{freq}(AB|\text{mother}) \quad (2.3b)$$

If we are only working with a single locus, then the gamete frequency is just the allele frequency, and under Hardy-Weinberg conditions, these do not change over the generations. However, when the gametes we consider involve two (or more) loci, recombination can cause gamete frequencies to change over time, even under Hardy-Weinberg conditions. At **linkage equilibrium**, the frequency of a multi-locus gamete is just equal to the product of the allele frequencies. For example, for two and three loci,

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) \quad \text{for 2 loci}, \quad \text{freq}(ABC) = \text{freq}(A) \cdot \text{freq}(B) \cdot \text{freq}(C) \quad \text{for 3 loci}$$

In linkage equilibrium, the alleles at different loci are independent — knowledge that a gamete contains one allele (say A) provides no information on the allele from the second locus. More generally, loci show **linkage**

disequilibrium (LD), which is also called **gametic phase disequilibrium** as it can occur between unlinked loci. When LD is present,

$$\text{freq}(AB) \neq \text{freq}(A) \cdot \text{freq}(B)$$

Indeed, the disequilibrium D_{AB} for gamete AB is defined as

$$D_{AB} = \text{freq}(AB) - \text{freq}(A) \cdot \text{freq}(B) \quad (2.4a)$$

Rearranging Equation 2.4a shows that the gamete frequency is just

$$\text{freq}(AB) = \text{freq}(A) \cdot \text{freq}(B) + D_{AB} \quad (2.4b)$$

$D_{AB} > 0$ implies AB gametes are more frequent than expected by chance, while $D_{AB} < 0$ implies they are less frequent. If the recombination frequency between the two loci is c , then the disequilibrium after t generations of recombination is simply

$$D(t) = D(0)(1 - c)^t \quad (2.5)$$

Hence, with loose linkage (c near $1/2$) D decays very quickly and gametes quickly approach their linkage equilibrium values. With tight linkage, disequilibrium can persist for many generations.

Contribution of a Locus to the Phenotypic Value of a Trait

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 ,

$$P = G + E \quad (2.6a)$$

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 not be the case. For example, a genetically higher-yield dairy cow may also be feed more, creating a positive correlation between G and E , and in this case the basic model becomes

$$P = G + E + Cov(G, E) \quad (2.6b)$$

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

Genotypes:	Q_1Q_1	Q_1Q_2	Q_2Q_2
	C	$C + a(1 + k)$	$C + 2a$
Average Trait Value:	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, $a = [G(Q_2Q_2) - G(Q_1Q_1)]/2$, and the relative position of the heterozygotes compared to the average of the homozygotes. If it 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

$$ak = d, \quad \text{or} \quad k = \frac{d}{a} \quad (2.7)$$

The reason for using both d and k is that some expressions are simpler using one parameterization than the other.

Example: the Booroola (*B*) gene

The Booroola (*B*) gene influences fecundity 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}), the homozygous effect of the *B* allele is estimated by $a = (2.66 - 1.48)/2 = 0.59$. The dominance coefficient is estimated by taking the difference between *bb* and *Bb* genotypes, $a(1 + k) = 0.69$, substituting $a = 0.59$, and rearranging to obtain $k = 0.17$. This suggests slight dominance of the Booroola gene. Using the alternative d notation, from Equation 4, $d = ak = 0.59 \cdot 0.17 = 0.10$

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 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 an individual carrying a 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,

$$G_{ij} = \mu_G + \alpha_i + \alpha_j + \delta_{ij} \quad (2.8)$$

The predicted genotypic value is $\hat{G}_{ij} = \mu_G + \alpha_i + \alpha_j$ and $G_{ij} - \hat{G}_{ij} = \delta_{ij}$. Here 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. Since α and δ represent deviations from the overall mean, they have expected values of zero.

You might notice that Equation 2.8 looks like a regression. Indeed it is. Suppose we have only two alleles, Q_1 and Q_2 . Notice that we can re-express Equation 2.8 as

$$G_{ij} = \mu_G + 2\alpha_1 + (\alpha_2 - \alpha_1)N + \delta_{ij} \quad (2.9)$$

where N is the number of copies of allele Q_2 , so that

$$2\alpha_1 + (\alpha_2 - \alpha_1)N = \begin{cases} 2\alpha_1 & \text{for } N = 0, \text{ e.g., } Q_1Q_1 \\ \alpha_1 + \alpha_2 & \text{for } N = 1, \text{ e.g., } Q_1Q_2 \\ 2\alpha_2 & \text{for } N = 2, \text{ e.g., } Q_2Q_2 \end{cases} \quad (2.10)$$

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 δ_{ij} are the residuals of the actual values from the predicted values.

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

A little algebra gives

$$\mu_G = 2p_1 p_2 a(1 + k) + 2p_2^2 a = 2p_2 a(1 + p_1 k) \quad (2.11a)$$

From the rules of regressions, the slope is just the

$$\alpha_2 - \alpha_1 = \frac{\sigma(G, N_2)}{\sigma^2(N_2)} = a [1 + k (p_1 - p_2)] \quad (2.11b)$$

See Lynch and Walsh, Chapter 4 for the details leading to Equation 2.11b. Since we have chosen the α to have mean value zero, it follows that

$$p_1 \alpha_1 - p_2 \alpha_2 = 0$$

When coupled with Equation 2.11b this implies (again, see L & W Chapter 4)

$$\alpha_2 = p_1 a [1 + k (p_1 - p_2)] \quad (2.11c)$$

$$\alpha_1 = -p_2 a [1 + k (p_1 - p_2)] \quad (2.11d)$$

Finally, the dominance deviations follow since

$$\delta_{ij} = G_{ij} - \mu_G - \alpha_i - \alpha_j \quad (2.11e)$$

Average Effects and Breeding Values

The α_i value is the **average effect** of allele Q_i . Note that the α and δ are functions of allele frequencies and that these change as the allele frequencies change. Breeders are concerned (indeed obsessed) with the **breeding values** (BV) of individuals, which are related to average effects. The BV associated with genotype G_{ij} is just

$$BV(G_{ij}) = \alpha_i + \alpha_j \quad (2.12a)$$

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

$$BV = \sum_{k=1}^n (\alpha_i^{(k)} + \alpha_k^{(k)}) \quad (2.12b)$$

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 values? Consider the offspring from the cross of a sire (genotype $Q_x Q_y$) mated to a number of unrelated dams (let the genotype of one of these random dams be $Q_w Q_z$ 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 w and z 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

$$\mu_O - \mu_G = \left(\frac{\alpha_x + \alpha_y}{2} \right) = \frac{BV(\text{Sire})}{2} \quad (2.13a)$$

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

$$BV(\text{Sire}) = 2(\mu_O - \mu_G) \quad (2.13b)$$

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

$$\mu_O - \mu_G = \frac{BV(\text{Sire})}{2} + \frac{BV(\text{Dam})}{2} \quad (2.13c)$$

The focus on breeding values thus arises because they predict offspring means.

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 (Lecture 1), 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,

$$\sigma^2(G) = \sigma^2(\mu_g + (\alpha_i + \alpha_j) + \delta_{ij}) = \sigma^2(\alpha_i + \alpha_j) + \sigma^2(\delta_{ij}) \quad (2.14)$$

Equation 2.14 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

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 \quad (2.15)$$

where σ_A^2 is the **additive genetic variance** and represents the variance in breeding values in the population, while σ_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 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

$$\sigma_A^2 = 2E[\alpha^2] = 2 \sum_{i=1}^m \alpha_i^2 p_i, \quad \text{and} \quad \sigma_D^2 = 2E[\delta^2] = \sum_{i=1}^m \sum_{j=1}^m \delta_{ij}^2 p_i p_j \quad (2.16)$$

For one locus with two alleles, these become

$$\sigma_A^2 = 2p_1 p_2 a^2 [1 + k(p_1 - p_2)]^2 \quad (2.17a)$$

and

$$\sigma_D^2 = (2p_1 p_2 ak)^2 \quad (2.17b)$$

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 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

$$\begin{aligned} G_{ijkl} &= \mu_G + (\alpha_i + \alpha_j + \alpha_k + \alpha_l) + (\delta_{ij} + \delta_{kl}) \\ &\quad + (\alpha\alpha_{ik} + \alpha\alpha_{il} + \alpha\alpha_{jk} + \alpha\alpha_{jl}) \\ &\quad + (\alpha\delta_{ikl} + \alpha\delta_{jkl} + \alpha\delta_{kij} + \alpha\delta_{lij}) \\ &\quad + (\delta\delta_{ijkl}) \\ &= \mu_G + A + D + AA + AD + DD \end{aligned} \tag{2.18}$$

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 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 (eg. 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 2.18 are uncorrelated, so that we can write the genetic variance as

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_{AA}^2 + \sigma_{AD}^2 + \sigma_{DD}^2 \tag{2.19}$$

Problems

1. Suppose loci A and B are linked, with $c = 0.25$. Further, suppose $\text{freq}(AB) = 0.1$, $\text{freq}(A) = 0.5$ and $\text{freq}(B) = 0.5$. Assume a random mating population.
 - a. Under Hardy-Weinberg, what is the frequency of an AA homozygote? A BB homozygote?
 - b. Assuming gametes combine at random, what is the expected frequency of an $AABB$ individual assuming the above gamete frequencies.
 - c. What is the initial disequilibrium for the AB gamete, D_{AB} ?
 - d. After four generations of recombination, what is the disequilibrium, $D_{AB}(4)$? What is $\text{freq}(AB)$? What is $\text{freq}(AABB)$?
2. Consider the Booroola gene mentioned early in the notes.
 - a. For $\text{freq}(B) = 0.3$, compute α_B , α_b , and the breeding values of all three genotypes.
 - b. For $\text{freq}(B) = 0.8$, compute α_B , α_b , and the breeding values of all three genotypes.
3. For the above two frequencies for Booroola, compute σ_G^2 , σ_A^2 , and σ_D^2
4. 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.
5. What is the best linear predictor of an individual's breeding value A given that we observe their phenotypic value P

Solutions to Chapter 2 Problems

1. a. $0.5^2 = 0.25$ for both homozygotes. Hence, one might expect $\text{freq}(AABB) = 0.25^2 = 0.0625$
- b. $\text{freq}(AABB) = \text{freq}(AB)^2 = 0.1^2 = 0.01$
- c. $D_{AB}(0) = \text{freq}(AB) - \text{freq}(A) \cdot \text{freq}(B) = 0.1 - 0.5 \cdot 0.5 = -0.15$
- d. $D_{AB}(4) = (1 - c)^4 D_{AB}(0) = -0.15(1 - .25)^4 = -0.047$,
 $\text{freq}(AB)(4) = \text{freq}(A)\text{freq}(B) + D_{AB}(4) = 0.20$.
 $\text{freq}(AABB) = \text{freq}(AB)(4) \cdot \text{freq}(AB)(4) = 0.04$

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

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

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

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

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

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

5. 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., $\mu_A = 0$