10

Short-term Changes in the Mean:
The Breeders' Equation

Prediction is very difficult, especially if it's about the future. — Niels Bohr

Version 16 Feb 2010.

Our examination of selection of quantitative characters, and its consequences, comprises the remainder of this book. We start by discussing the simplest models — changes in the mean of a single character following a single generation of selection. As the reader will see over the next several chapters, even these simple models are not without complications. Over subsequent chapters we build on these models, making them more complex and biologically realistic.

Selection changes the distribution of both phenotypic and genotypic values from one generation to the next. While it is convenient to assume that only the means of these distributions change, this is generally not the case. Rigorous prediction of the expected change in the mean over several generations requires predictions of how higher moments (such as the variance and skew) of the genotypic distribution change as well (Chapters 13, 24). Unfortunately, this requires considerable knowledge about the underlying distributions of allelic effects and frequencies, both of which are almost always unknown. Despite this gloomy initial assessment, the change in mean over a few generations of selection can often be accurately projected using estimates of genetic variances from an unselected base population. This chapter considers the basis for much of this success: models that assume a linear parent-offspring regression and short-term stability of the genetic variance (i.e., over short time scales, selection-induced changes in variance are assumed to be negligible).

We will have much to say about selection-induced changes in the genetic variances and covariances in subsequent chapters (Chapters 13, 24). For now, the assumption of the infinitesimal model (a very large number of underlying loci, each with very small effect) ensures that negligible change in allele frequencies will occur over only a few generations of selection. However (Chapter 13), even under the infinitesimal model, *genotypic* frequencies can change, due to selection generating gametic-phase disequilibrium, even while allele frequencies remain essentially unchanged. Over longer time scales, changes in variances can be substantial and are extremely difficult to predict. Thus, while the results presented here hold for a single generation, they are expected to become increasingly less accurate as the number of generations of selection increases.

Besides the constancy of variances and linearity of parent-offspring regression, the third major assumption required here is that there is no selection acting on characters correlated to our character of interest. Breeders have the luxury of artificial selection and hence fairly precise control over the nature of selection, while evolutionary biologists face the additional problem of estimating the form of selection acting on the character(s) under consideration. However, even under highly controlled breeding designs, natural selection can also be acting on the character of interest (either directly, or through correlations with other characters under selection). Accounting for selection on multiple characters is discussed in Chapters 30–36.

There is a huge literature on different selection schemes that exploit special features of

specific organisms (such as artifical insemination in large farm animals or complex crosses in plants that are capable of selfing). Our focus in this chapter is **individual** (or **mass**) **selection**, wherein individuals are chosen solely on the basis of their phenotypic value (i.e., information from relatives, other characters, etc. is ignored). **Family selection**, where individuals are chosen based on either their family mean or their relative ranking within a family is discussed in Chapter 17. Using such information from relatives can improve our ability to accurately predict breeding values and hence increase response relative to individual selection. The general theory for selection using information from relatives relies on index selection and BLUP (LW Chapter 26), both of which are discussed in Chapters 32–34. Additional complications (such as overlapping generations, inbreeding, cross-breeding, and group selection) are examined in Chapters 17–23. Finally, the reader is also referred to Turner and Young (1969), Pirchner (1983), Ollivier (1988), Cameron (1997), Simm (1998), Kinghorn et al. (2000), and Weller (2001) for specialized applications in animal breeding and to Namkoong (1979), Hallauer and Miranda (1981), Wricke and Weber (1986), Mayo (1987), Stoskopf et al. (1993), Bernardo (2002), and Gallais (2003) for specialized applications in plant breeding.

SINGLE-GENERATION RESPONSE: THE BREEDERS' EQUATION

The simplest selection model is the classic **breeders' equation**

$$R = h^2 S \tag{10.1}$$

which relates response to heritability times the selection differential. This is perhaps the most well known expression in quantitative genetics. The actual origin of the breeders' equation is somewhat unclear, but it was clearly suggested (in multivariate form) in the early writings of Pearson and popularized by Lush (1937). The simplicity of this equation is compelling, relating the change in mean *across* a generation (the response R) to the product of the *within*-generation change in the mean (the **directional selection differential** S) with a measure of how character value is passed across generations (the slope h^2 of the midparent-offspring regression). The breeders' equation immediately follows by assuming the parent-offspring regression is linear. In particular, the expected phenotypic value z_o of an offspring given the mean phenotypic value of its parents z_{mp} is given by the midparent-offspring regression,

$$z_o = \mu + h^2(z_{mp} - \mu)$$

Taking the average over all the selected parents, $E_s[z_{mp} - \mu] = S$, and likewise the difference between the expected value of the offspring from such parents and the overall mean is the response R, giving

$$E_s[z_o - \mu] = R = h^2 E_s[z_{mp} - \mu] = h^2 S$$

Even when the parent-offspring regression is not assumed to be linear, the breeders' equation still approximately holds (Equations 17.31b and 17.35b).

Response is the Change in Mean Breeding Value

A key concept is that response (being the change in offspring mean) equals the mean breeding value of the selected parents. Recall that (non-inbred, sexually reproducing) parents pass along only a fraction of their total genotypic value, namely their breeding value A, to their offspring. The expected offspring value is just the average breeding values of its parents (LW Chapter 4). In an unselected base population with a nonzero σ_A^2, potential parents show a range of breeding values distributed about an expected breeding value of zero.

Selection response occurs by choosing parents with breeding values favorable to character improvement, and the expected change in offspring mean from this set of parents is simply the average breeding value of the selected parents.

Hence, one goal of selection is to choose those parents in a sample with the most favorable breeding values. The problem is that we cannot completely predict the breeding value of an individual from its phenotype alone (unless $h^2 = 1$). Phenotype is an imperfect predictor of breeding value, and therefore the offspring of phenotypically exceptional parents are generally not themselves as exceptional. To see the direct connection between changes in breeding value and the breeders' equation, recall that the predicted value $\widehat{A}$ of an individual's breeding value given their phenotypic value z is just

$$\widehat{A} - \mu_A = \frac{\sigma(A, z)}{\sigma_z^2}(z - \mu_z), \quad \text{or} \quad \widehat{A} = h^2(z - \mu_z)$$

This follows from standard regression theory (LW Chapter 3) as (i) the mean of the regression passes through the mean of A and z, which are 0 and μ_z (respectively), and (ii) the slope of the regression for predicting A given z is just $\sigma(z, A)/\sigma_z^2 = \sigma_A^2/\sigma_z^2 = h^2$, which follows since

$$\sigma(z, A) = \sigma(G + E, A) = \sigma(A + D + E, A) = \sigma(A, A) = \sigma_A^2$$

Hence, the expected breeding value for a set of selected parents is just

$$E_s[h^2(z - \mu_z)] = h^2 E_s[z - \mu_z] = h^2 S$$

where E_s denotes the average value over the selected parents. The change in the mean value of their offspring (relative to the base population) is just the mean parental breeding value. Thus response = $h^2 S$ and we recover the breeders' equation via another route.

Thinking in terms of breeding values will prove a most useful way of treating many features of selection. As we will see, a number of selection schemes (such as using information from relatives) have been proposed to improve response by improving our ability to estimate the breeding value of potential parents. Further, thinking about the breeding values of individuals will prove a helpful way of considering selection of multiple characters.

The Importance of Linearity

A variety of factors can result in a nonlinear parent-offspring regression (Chapter 24, LW Chapter 17). In such cases, the mean of the selected parents (and hence the selection differential S) is not sufficient to predict the offspring mean. As Figure 10.1 shows, two selected parental populations with the same mean, but different variances, can have different expected responses when the parent-offspring regression is nonlinear. Even if phenotypes are normally distributed and the character is completely determined by additive loci (no dominance or epistasis), if the underlying distribution of genotypic values shows skew, selection on the variance also results in a change in the mean (see Equation XX9.35b). A sufficient condition for linearity is that the joint distribution of breeding and phenotypic values is bivariate normal (LW Chapter 8).
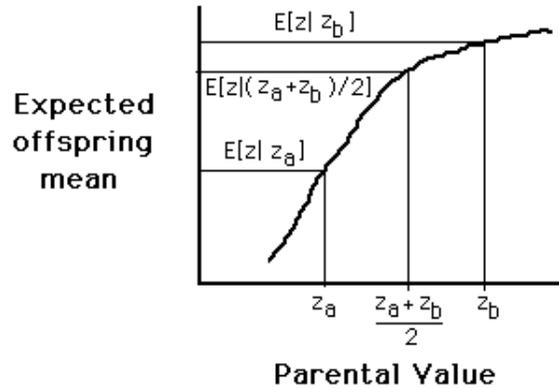
Figure 10.1.   The importance of linearity in the parent-offspring regression. If this regression is nonlinear, different subsets of the population with the same mean can have different offspring means. Suppose equal numbers of parents with values z_a and z_b are chosen. Denoting the expected value of offspring from parents with value z_x by $E[\,z \,|\, z_x\,]$, the offspring mean is given by $(\,E[\,z\,|\,z_a\,] + E[\,z\,|\,z_b\,]\,)\,/\,2$. Conversely, choosing parents all with value $(z_a + z_b)/2$, gives the same parental mean and hence the same S, but the expected offspring mean is now $E[\,z\,|\,(z_a + z_b)/2\,]$, which, as shown, can deviate considerably from $(\,E[\,z\,|\,z_a\,] + E[\,z\,|\,z_b\,]\,)\,/\,2$.

Selection generally causes the distribution of genotypic values to depart from normality, creating at least slight departures from linear parent-offspring regressions. Response under such non-normal distributions can be very complicated, depending on underlying genetic factors that do not easily translate into standard (and measurable) variance components (see Chapter 24).

Response Under More General Parent-Offspring Regressions

The regression coefficients for parent-offspring regressions can vary with the sex of both parents and offspring. In such cases, the breeders' equation can be extended by incorporating these sex-dependent regression slopes. To see how this is done, denote the phenotypic values of the father and mother by z_{fa} and z_{mo} and an offspring by z_o (if its sex is unimportant) or by z_{so} and z_{da} for sons and daughters (respectively) when sex is important. Let $E(z_o \,|\, z_{fa}, z_{mo})$ be the expected phenotypic value of an offspring whose parents have phenotypic values z_{mo} and z_{fa}. The importance of this conditional expectation (the **biparental regression**) is that the expected character value in the next generation (assuming no fertility differences) is the average of $E_s(z_o \,|\, z_{fa}, z_{mo})$ over all selected parents. Taking expectations is particularly straightforward when the biparental regression is linear, i.e.,

$$E(z_o \,|\, z_{fa}, z_{mo}) = \mu_o + b_{o,fa}\,(z_{fa} - \mu_{fa}) + b_{o,mo}\,(z_{mo} - \mu_{mo}) \tag{10.2}$$

where μ_{fa} and μ_{mo} are the mean character values of males and females before selection and μ_o the mean for the offspring sex being considered. Denoting the expectation taken over all selected parents by E_s, the expected offspring mean after selection is

$$E_s\big[\,E(z_o \,|\, z_{fa}, z_{mo})\,\big] = \mu_o + b_{o,fa}\,S_{fa} + b_{o,mo}\,S_{mo} \tag{10.3}$$

where S_{fa} and S_{mo} are the directional selection differentials on fathers and mothers (respectively).

Equations 10.2 and 10.3 allow for the possibility of differences between sexes in regression coefficients, in which case separate equations for sons and daughters are required. For example, the expected change in the mean character value of daughters, R_{da}, equals the expected mean of daughters of selected parents minus the mean of females before selection. Applying Equation 10.3,

$$E_s \left[E(z_{da} \,|\, z_{fa}, z_{mo}) \right] = \mu_{mo} + b_{da,fa} \, S_{fa} + b_{da,mo} \, S_{mo}$$

implying

$$R_{da} = b_{da,fa} \, S_{fa} + b_{da,mo} \, S_{mo} \tag{10.4a}$$

where $b_{da,fa}$ is the regression coefficient of daughters on their fathers and $b_{da,mo}$ the mother-daughter regression coefficent. Likewise, for sons

$$R_{so} = b_{so,fa} \, S_{fa} + b_{so,mo} \, S_{mo} \tag{10.4b}$$

Example 10.1.   Coyne and Beecham (1987) estimated the following parent-offspring regression coefficients for abdominal bristle number in laboratory populations of *Drosophila melanogaster*:

Mother-son	$b_{so,mo} = 0.39 \pm 0.08$
Mother-daughter	$b_{da,mo} = 0.32 \pm 0.08$
Father-son	$b_{so,fa} = 0.13 \pm 0.10$
Father-daughter	$b_{da,fa} = 0.40 \pm 0.08$

Note that the father-son regression has a significantly smaller slope than the three other parent-offspring sex combinations. (Gimelfard and Willis 1994 give other *Drosophila* examples where the regressions differ significantly between sons and daughters). Assume that these estimated values are indeed the true values and that different amounts of selection are applied to fathers and mothers. Suppose the mean increase in bristle number in selected fathers is 2, while selected mothers show a mean decrease of 1 bristle. What is the expected change in mean bristle number in the male and female offspring using these estimated regression coefficients, assuming all parent-offspring regressions are linear? Here $S_{mo} = -1$ and $S_{fa} = 2$, and from Equation 10.4a, the expected change in bristle number in females is

$$R_{da} = b_{da,mo} \, S_{mo} + b_{da,fa} \, S_{fa} = 0.32 \,(-1) + 0.40 \,(2) = 0.48$$

Likewise, from Equation 10.4b, the expected change in males is

$$R_{so} = b_{so,fa} \, S_{fa} + b_{so,mo} \, S_{mo} = 0.13 \,(2) + 0.39 \,(-1) = -0.13$$

Even though we selected for increased bristle number in males and decreased number in females, the expected response is the exact opposite: a decrease in males and an increase in females.

Equation 10.4 is very general, requiring only that the biparental regression is linear, in which case $b_{o,fa}$ and $b_{o,mo}$ are partial regression coefficients and can be obtained from covariances between relatives. Again, linearity is ensured if the joint distribution of both parents and their offspring is multivariate normal. If there is no correlation between the phenotypes

of parents (which is guaranteed under random mating), the partial regression coefficients are standard univariate regression coefficients (LW Chapter 8), so that LW Equation 3.14b gives

$$b_{o,fa} = \frac{\sigma(z_o, z_{fa})}{\sigma^2(z_{fa})} \qquad \text{and} \qquad b_{o,mo} = \frac{\sigma(z_o, z_{mo})}{\sigma^2(z_{mo})}$$

If mating is random, and genotype $\times$ environmental interactions, shared environmental effects, epistasis, and sex-specific effects can all be neglected, the regression slope (for each parent-offspring combination) is $b_{o,p} = h^2/2$ (LW Chapters 7, 17). Defining the total selection differential as the average of both parental values, $S = (S_{fa} + S_{mo})/2$, again recovers the breeders' equation

$$R = \frac{h^2}{2} S_{fa} + \frac{h^2}{2} S_{mo} = h^2 S \tag{10.5}$$

Equation 10.5 shows how differential selection on parents is incorporated into the breeders' equation. For example, consider selection on dioecious plants. If plants to form the next generation are chosen *after* pollination, fathers (pollen donors) are chosen at random with respect to the character under selection and $S_{fa} = 0$, giving $R = (h^2/2)S_{mo}$. Conversely, if parents are selected before pollination with equal amounts of selection (S) on both sexes, $R = h^2 S$.

The Selection Intensity, $\bar{\imath}$

While the selection differential S is a convenient and simple measure of the selection, it does not really tell us much about the *strength* of selection. Consider selection acting on the same character in two different populations. In one, the largest five percent of measured individuals are allowed to reproduce while in the second the largest 25 percent reproduce. Clearly selection is more intense in the first population. However, if the characters are normally distributed, Equation 10.22a (below) gives the selection differentials for these two populations as $S_1 = 2.06 \, \sigma_1$ and $S_2 = 1.27 \, \sigma_2$, where σ_k^2 is the character variance in population k. Provided the second population is sufficiently more variable than the first, it can have the larger selection differential even though it clearly experiences less intense selection.

A better measure is to use the **standardized directional selection differential** (or **selection intensity**), which is the selection differential expressed in phenotypic standard deviations,

$$\bar{\imath} = S/\sigma_z \tag{10.6a}$$

The selection intensity accounts for differences in the phenotypic variances, much akin to the correlation being a better measure of the strength of association than is the covariance (LW Chapter 3). Substituting $\bar{\imath} \, \sigma_z$ for S gives the selection intensity version of the breeders' equation,

$$R = h^2 \, \bar{\imath} \, \sigma_z = \bar{\imath} \, h \, \sigma_A \tag{10.6b}$$

which follows since

$$h^2 \sigma_z = \frac{\sigma_A^2}{\sigma_z^2} \sigma_z = \frac{\sigma_A}{\sigma_z} \sigma_A = h \, \sigma_A$$

Equation 10.6b will prove to be a useful starting point for generalizations (developed below) of the Breeders' Equation to accommodate more general types of selection.

The Robertson-Price Identity, $S = \sigma(w, z)$

As introduced in LW Chapter 3, the selection differential can also be written as the covariance between relative fitness and trait value,

$$S = \sigma(w, z) \tag{10.7}$$

This relationship was first noted by Robertson (1966) and late elaborated on by Price (1970, 1972), and hence we refer to Equation 10.7 as the **Robertson-Price identity**. To derive Equation 10.7, let z_i, p_i, and w_i be the trait value, frequency before selection, and relative fitness (respectively) of class i. The selection differential is simply the mean after selection minus the mean before selection,

$$S = \mu_s - \mu = \sum z_i \, w_i \, p_i - \sum z_i \, p_i = E[\, z \, w \,] - E[\, z \,]$$

Since $E[\, w \,] = 1$, we can write this as

$$S = E[\, z \, w \,] - E[\, z \,] \, E[\, w \,] = \sigma(w, z)$$

recovering the Robertson-Price identity.

Correcting for Reproductive Differences: Effective Selection Differentials

In artificial selection experiments, S is usually estimated as the difference between the mean of the selected adults and the sample mean of the population before selection. Selection need not stop at this stage. For example, strong artificial selection to increase a character might be countered by natural selection due to a decrease in the fertility of individuals with extreme character values. Biases introduced by such differential fertility can be removed by randomly choosing the same number of offspring from each selected parent, ensuring equal fertility.

Alternatively, biases introduced by differential fertility can be accounted for by using **effective selection differentials**, S_e,

$$S_e = \frac{1}{n_p} \sum_{i=1}^{n_p} \left(\frac{n_i}{\overline{n}} \right) (z_i - \mu_z) \tag{10.8}$$

where z_i and n_i are the phenotypic value and total number of offspring of the ith parent, n_p the number of parents selected to reproduce, $\overline{n}$ the average number of offspring for selected parents, and μ_z is the mean before selection. If all selected parents have the same number of offspring ($n_i = \overline{n}$ for all i), then S_e reduces to S. However, if there is variation in the number of offspring n_i among selected parents, S_e can be considerably different from S. This corrected differential is also referred to as the **realized selection differential**.

Example 10.2.   Consider a trait with a heritability of 0.3 and a (before selection) mean of 30. Suppose 5 parents are selected, with the following trait values and offspring number:

Parent	Phenotypic value	Number of offspring
1	45	1
2	40	2
3	35	3
4	33	5
5	32	5

The resulting (unweighted) mean is 37, giving $S = 7$ and and expected response of $R = 0.3 \cdot 7 = 2.1$.

However, computing the effective selective differential by weighting the selected parents proportional to the number of offspring they leave gives

i	z_i	n_i	$n_i/\overline{n}$
1	45	1	0.3125
2	40	2	0.6250
3	35	3	0.9375
4	33	5	1.563
5	32	5	1.563

$$\frac{1}{n_p} \sum_{i=1}^{n_p} \left(\frac{n_i}{\overline{n}} \right) z_i = 34.69$$

Hence, $S_e = 4.69$, for an expected response of $R = 0.3 \cdot 4.69 = 1.4$. In this case, not using the effective differential results in an overestimation of the expected response.

The derivation of Equation 10.8 follows directly from applying the Robertson-Price identity (LW Equation 3.21), $S = \sigma(z, w)$, which expresses the selection differential as a covariance between trait value z and relative fitness w. (Relative fitness is simply fitness scaled to have a mean value of one, i.e, the relative fitness of individual i is its absolute fitness divided by the mean population fitness, $w_i = W_i/\overline{W}$.) If a total of n potential parents are examined, n_p of which are selected as parents, then

$$S = \sigma(z, w) = E[\, w\, z\,] - E[\, z\,]\, E[\, w\,] = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i}{\overline{W}} \right) z_i - \mu_z \cdot 1$$

where the fitness of individual i is $W_i = n_i$ (with $n_i = 0$ for individuals not chosen as parents). The mean fitness becomes

$$\overline{W} = \frac{1}{n} \sum_{i=1}^{n} n_i = \frac{\overline{n}\, n_p}{n}, \qquad \text{where} \qquad \overline{n} = \sum_{i=1}^{n} n_i/n_p$$

so that $\overline{n}$ is the mean number of offspring left by adults selected to reproduce. Hence

$$\frac{W_i}{\overline{W}} = \frac{n_i\, n}{\overline{n}\, n_p}, \quad \text{giving} \quad \sigma(z, w) = \sum_{i=1}^{n} z_i \frac{n_i}{\overline{n}\, n_p} - \mu_z$$

Rearranging recovers Equation 10.8. Chapter 28 examines individual fitness and fitness calculations such as this in detail.

EXPANDING THE BASIC BREEDERS' EQUATION

Accuracy

Under appropriate linearity assumptions, the breeders' equation generalizes to much more general settings, predicting the mean of character y measured in one group when selection

occurs on character x measured in another group. Assuming the regression of y on x is linear, standard regression theory (LW Equation 3.14) gives

$$E[\,y - \mu_y\,|\,x\,] = \frac{\sigma(x,y)}{\sigma_x^2}\,(x - \mu_x)$$

giving the expected change in y from selection on x as

$$R_y = \mu_y^* - \mu_y = \frac{\sigma(x,y)}{\sigma_x^2}\,(\,\mu_x^* - \mu_x\,) = \frac{\sigma(x,y)}{\sigma_x^2}\,S_x = \frac{\sigma(x,y)}{\sigma_x}\,\overline{\imath}_x \qquad (10.9)$$

where μ^* denotes the mean following selection.

Example 10.3.   As an example of applying Equation 10.9, consider selection on clones or other pure lines. With clones, parents pass on their entire genome to their offspring. The phenotypic value z of an offspring from a parent with genotypic value G can be written as $z = G + E$, so that parent-offspring covariance (in the absence of any genotype $\times$ environment covariance and/or interactions) equals the total genetic variance σ_G^2. The resulting parent-offspring regression has slope $b_{op} = \sigma_G^2/\sigma_z^2 = H^2$, the **broad-sense heritability** (LW Chapter 20), giving

$$R = H^2\,S$$

Since $H^2 \geq h^2$ (as $\sigma_G^2 \geq \sigma_A^2$), the single-generation response to selection is larger for clones than for a sexual population with the same variance components. When selection continues for several generations, using clones is expected to be less efficient than using a sexual population. Selection among clones very rapidly removes genetic variation from the population without any mechanism (other than mutation) to generate new variation. With selection among sexual individuals, recombination generates an almost endless supply of new variation if a large number of segregating loci underlie the trait. For this reason, selection schemes involving clones often randomly mate lines every few generations to allow for recombination, generating variation required for continued response. Selection and development of pure lines is covered in detail in Chapter 20.

The selection intensity version of Equation 10.9 is

$$R_y = \frac{\sigma_{x,y}}{\sigma_x}\,\overline{\imath}_x = \overline{\imath}_x\,\sigma_y\,\rho(x,y) \qquad (10.10a)$$

where $\rho(x,y)$ is the correlation between x and y. This correlation is referred to as the **accuracy** in predicting the response in y from knowledge of x, and one immediately sees that by improving the accuracy of our selection scheme, we improve the response. Expressing Equation 10.10a in terms of the **relative response**, the change in y in phenotypic standard deviations of y, gives

$$\frac{R_y}{\sigma_y} = \overline{\imath}_x\,\rho(x,y) \qquad (10.10b)$$

Equation 10.10b allows one to compare the relative efficiencies of different selection schemes. Fixing the selection intensity $\overline{\imath}$, the maximal expected response in y occurs when we select on the measure x that has the largest correlation with y. The relative response of two different schemes (x_1 versus x_2) is given by $\rho(x_1,y)/\rho(x_2,y)$.

Most commonly, y is the breeding value A of an individual (for our focal trait) so that R_y is the change in breeding value from selection on x, giving

$$R_A = \bar{\imath}_x \, \rho(x, A) \, \sigma_A \tag{10.11}$$

Hence, the breeders equation can be consider as the following product:

Response = (Intensity)*(Accuracy in Predicting Breeding Value)*($\sqrt{\text{Usable Variance}}$)

If x is the phenotypic value z of the character whose breeding value is of interest, the correlation between x and y here is the correlation between an individual's phenotype z and its breeding value A,

$$\rho(z, A) = \sigma(z, A)/(\sigma_z \, \sigma_A) = \sigma_A/\sigma_z = h$$

recovering Equation 10.6b.

In evaluating other breeding schemes h is replaced by the appropriate correlation between the breeding value and the measure x being selected on. For example, we could take x to be the family mean of the character for the individual being considered, in which case the correlation between an individual's breeding value and its family mean determines the response to selection under this scheme (Chapter 17). Assuming $\bar{\imath}$ and σ_A are constant, the largest response occurs by taking the measure x that gives the largest correlation with breeding value. This idea forms the foundation of index selection (Chapters 32 and 33) wherein x is a linear combination of several characters $x = \sum a_i z_i$.

Example 10.4.   Progeny testing, using the mean of a parent's offspring to predict the parent's breeding value, is an alternative predictor of an individual's breeding value. In this case, the correlation between the mean x of n offspring and the breeding value A of the parent is

$$\rho(x, A) = \sqrt{\frac{n}{n + a}}, \quad \text{where} \quad a = \frac{4 - h^2}{h^2}$$

From Equation 10.11, the response to selection under progeny testing is

$$R = i\sigma_A \sqrt{\frac{n}{n + a}} = i\sigma_A \sqrt{\frac{h^2 n}{4 + h^2(n - 1)}}$$

Note that for very large n that the accuracy approaches one. Progeny testing gives a larger response than simple selection on the phenotypes of the parents (**mass selection**) when

$$\sqrt{\frac{n}{4 + h^2(n - 1)}} > 1, \quad \text{or} \quad n > \frac{4 - h^2}{1 - h^2}$$

In particular, $n > 4$, 5, and 7, for $h^2 = 0.1$, 0.25, and 0.5. Also note that the ratio of response for progeny testing (R_{pt}) to mass selection (R_{ms}) is just

$$\frac{R_{pt}}{R_{ms}} = \frac{1}{h} \sqrt{\frac{h^2 n}{4 + h^2(n - 1)}} = \sqrt{\frac{n}{4 + h^2(n - 1)}}$$

which approaches $1/h$ for large n.

Example 10.5.   **Sib selection**. Suppose our trait of interest is extremely hard to measure in live individuals (such as meat quality). How can we select on individuals if we (essentially) have to kill them to measure the trait? Likewise, consider traits expressed in only one sex, such as milk production. How can we select on males if they do not express the trait themselves? The answer in both cases is sib selection (Chapter 17), wherein we use sibs of exceptional individuals as the parents which form the next generation. For example, we chose brothers based on the milk production values of their sisters. Here the selection unit x is the phenotype of the sib. What is the correlation between the phenotypic value P_{s1} of one sib and the breeding value A_{s2} of another?

$$Cov(P_{s1}, A_{s2}) = \begin{cases} \sigma_A^2/2 & \text{for full sibs} \\ \sigma_A^2/4 & \text{for half-sibs} \end{cases} , \quad \text{hence,} \quad \rho(x, A) = \begin{cases} h^2/2 & \text{for full sibs} \\ h^2/4 & \text{for half-sibs} \end{cases}$$

Thus the response to selecting on a sib is

$$R = \begin{cases} \bar{\imath}_x \, (h/2) \, \sigma_A & \text{for full sibs} \\ \bar{\imath}_x \, (h/4) \, \sigma_A & \text{for half-sibs} \end{cases}$$

Hence, the response (using full sibs) is half that of mass selection, while using half-sibs gives only a quarter of the expected response under mass selection. In the case of milk production, which is a mixture of mass and sib selection, the expected response is

$$R = (1/2)\bar{\imath} \, h \, \sigma_A + (1/2)\bar{\imath}_x \, (h/2) \, \sigma_A = (3/4)\bar{\imath} \, h \, \sigma_A$$

where the first term is the response from using superior daughters (mass selection) and the second term the response using the brothers of superior daughters (sib selection). Another way to think about this is that $\bar{\imath} \, h \, \sigma_A$ is the change in breeding values in the mothers and $\bar{\imath}_x \, (h/2) \, \sigma_A$ the change in breeding values of the fathers. The change in the offspring mean (the response) is just the average of these two breeding value changes.

Reducing Environmental Noise: Stratified Mass Selection

Accuracy (and hence response) can also be increased by using design schemes that reduce environmental noise. One popular approach is Gardner's (1961) method of **stratified mass selection**, where a population is stratified into a number of blocks (potentially representing different microenvironments) and selection occurs within each block. The motivation for Gardner's method was to improve individual selection for yield in maize. At the time of Gardner's paper, selection based solely on individual phenotype for yield resulted in a very poor response, largely because environmental effects overwhelm genetic differences. Simply by selecting for plants within blocks of presumably similar environments, Gardner was able to use mass selection to obtain fairly significant gains (about 4% per year).

Stratified mass selection is performed as follows. Suppose n individuals are measured within each block, and selection occurs on the deviation from the block mean, e.g., on $z_{ij} - \overline{z}_i$ where z_{ij} is the jth individual from block i and $\overline{z}_i$ is the block mean. An individual's phenotypic value can be expressed as its genotypic value G_{ij} plus an environmental value consisting of a block effect bl_j and the residual environmental value e_{ij},

$$z_{ij} = \mu + G_{ij} + bl_j + e_{ij} \tag{10.12a}$$

The total environmental variance equals the variance between blocks σ_{bl}^2 plus the within-block variance σ_{Wbl}^2 (the variance of the residuals e_{ij}), giving the total variance as

$$\sigma_z^2 = \sigma_G^2 + \sigma_E^2 = \sigma_G^2 + \sigma_{bl}^2 + \sigma_{Wbl}^2 \tag{10.12b}$$

To predict the response to selection under stratified mass selection, we use Equation 10.9, where the selection critera x here is $z_{ij} - \overline{z}_i$ and and y is the breeding value A_{ij} of individual z_{ij}. The resulting covariance required for Equation 10.9 is

$$\sigma(x, y) = \sigma(z_{ij} - \overline{z}_i, A_{ij}) = \sigma(z_{ij}, A_{ij}) - \frac{1}{n}\sum_{k=1}^{n}\sigma(z_{ik}, A_{ij})$$

$$= \sigma_A^2\left(1 - \frac{1}{n}\right) \simeq \sigma_A^2 \tag{10.13}$$

as the assumption is that individuals within blocks are unrelated and that a large number of individuals are scored within each block. The phenotypic variance within the block is $\sigma_z^2 = \sigma_G^2 + \sigma_{Wbl}^2$, giving the response under stratified mass selection as

$$R = \frac{\sigma(x, y)}{\sigma_x}\,\overline{\imath} = \frac{\overline{\imath}\,\sigma_A^2(1 - 1/n)}{\sqrt{\sigma_G^2 + \sigma_{Wbl}^2}} \simeq \frac{\overline{\imath}\,\sigma_A^2}{\sqrt{\sigma_G^2 + \sigma_{Wbl}^2}} \tag{10.14a}$$

In contrast, if the effects of blocks are ignored and individuals are simply selected from the entire population, the response becomes

$$R = \frac{\overline{\imath}\,\sigma_A^2}{\sqrt{\sigma_G^2 + \sigma_{bl}^2 + \sigma_{Wbl}^2}} \tag{10.14b}$$

where σ_{bl}^2 is the variance between blocks in environmental effects. The relative efficiency of stratification (assuming the block size is modest to large so that $1 - 1/n \simeq 1$) is

$$\sqrt{\frac{\sigma_G^2 + \sigma_{bl}^2 + \sigma_{Wbl}^2}{\sigma_G^2 + \sigma_{Wbl}^2}} = \sqrt{1 + \frac{\sigma_{bl}^2}{\sigma_G^2 + \sigma_{Wbl}^2}} \tag{10.15}$$

Thus if the between-block variance is considerable, the response can be significantly improved by taking blocks into consideration.

Reducing Environmental Noise: Repeated-measures Selection

The **repeated-measures** design is a second example of increasing accuracy (and response) by providing some control over environmental noise. Here the character of interest is measured at n different times in an individual, and selection occurs on $\overline{z}_i$, the mean value for individual i. For example, one could use the mean value of three litters (as opposed to simply using the first litter). This is an especially common design in behavioral experiments, wherein a single measure (such as wheel-running speed) may vary greatly within an individual over time.

The model here depends on the *repeatability* (LW Chapter 6) of the character. Decompose the character value for the jth measure of individual i as

$$z_{ij} = G_i + E_i + e_{ij} \tag{10.16a}$$

where G and E are the genotypic and environmental values common to all measures of i and e_{ij} is the special environmental value restricted to the jth measure of i. The *repeatability* of the character is defined as

$$r = \frac{\sigma_G^2 + \sigma_E^2}{\sigma_z^2} = 1 - \frac{\sigma_e^2}{\sigma_z^2} \tag{10.16b}$$

Note that

$$r\sigma_z^2 = \sigma_G^2 + \sigma_E^2 \quad \text{and} \quad (1-r)\sigma_z^2 = \sigma_e^2 \tag{10.16c}$$

Now consider the prediction of an individual's breeding value A_i from the mean of the j measures, $\overline{z}_i$. The regression of A_i on $\overline{z}_i$ is given by

$$A_i = \mu + b\overline{z}_i + e = \beta(\overline{z}_i - \overline{z}) + \epsilon \tag{10.17a}$$

where

$$\beta = \frac{\sigma(A_i, \overline{z}_i)}{\sigma^2(\overline{z}_i)} \tag{10.17b}$$

The covariance between breeding value and the mean of repeated measures is just

$$\sigma(A_i, \overline{z}_i) = \frac{1}{n} \sum_{j=1}^{n} \sigma(A_i, z_{ij}) = \frac{1}{n} n\, \sigma(A_i, A_i) = \sigma_A^2 \tag{10.17c}$$

Writing

$$\overline{z}_i = G_i + E_i + \frac{1}{n} \sum_{j=1}^{n} e_{ij} \tag{10.17d}$$

it immediately follows from Equation 10.16c that

$$\sigma^2(\overline{z}_i) = \sigma_G^2 + \sigma_E^2 + \sigma_e^2/n$$

$$= \sigma_z^2 r + \sigma_z^2 \frac{1-r}{n} = \sigma_z^2 \left(\frac{1+(n-1)r}{n} \right) \tag{10.17e}$$

Substituting these results into Equation 10.17a, and recalling that the expected value of the regression is zero ($E[\epsilon] = 0$), gives the predicted breeding value for individual i as

$$\widehat{A}_i = h^2 \left(\frac{n}{1+(n-1)r} \right) (\overline{z}_i - \overline{z}) \tag{10.18a}$$

giving the response to selection as

$$R = h^2 \left(\frac{n}{1+(n-1)r} \right) S \tag{10.18b}$$

Likewise, the resulting accuracy in predicting an individual's breeding value based on n repeated measures of a trait is

$$\rho(\overline{z}_i, A) = \frac{\sigma(\overline{z}_i, A)}{\sigma(\overline{z}_i)\, \sigma_A} = h \sqrt{\frac{n}{1+(n-1)r}} \tag{10.19a}$$

The ratio of accuracies under repeated-measures vs. mass selection becomes

$$\frac{\rho(\overline{z}_i, A)}{\rho(z, A)} = \sqrt{\frac{n}{1+(n-1)r}} \tag{10.19b}$$

which approaches $1/\sqrt{r}$ for large n. Hence, when repeatability is low ($\sigma_e^2 > \sigma_G^2 + \sigma_E^2$), considerable improvement in response can be accomplished by using repeated-measured selection.

Example 10.6.   As an example of the consequences of single versus multiple measures, consider the following data set, simulated by assuming a character with a heritability of 0.1, a total variance of $\sigma_z^2 = 100$, a mean of 50, and a repeatability of $r = 0.2$. The values for twenty individuals from either an initial measurement ($z(1)$) or the average of five measurements ($z(5)$) were then computed using the simulation parameters and drawing values from the appropriate normal distributions.

j	$z(1)$	$z(5)$	j	$z(1)$	$z(5)$
1	54.97	56.80	11	49.81	48.76
2	64.01	54.51	12	51.92	46.76
3	42.64	52.61	13	43.56	51.79
4	42.70	38.69	14	41.60	47.23
5	61.62	56.42	15	51.80	48.90
6	39.86	47.70	16	52.88	47.21
7	56.54	48.63	17	63.86	54.03
8	35.88	47.26	18	39.76	49.62
9	54.32	53.93	19	36.45	47.78
10	57.85	45.10	20	59.07	51.16

Suppose the uppermost 25 percent (top 5 of the 20) are chosen for selection. Based on the initial measure, individuals 2, 5, 10, 17, and 20 would be selected, while based on five measures, individuals 1, 2, 5, 9, and 17 would be selected. The mean overall initial measure scores is 50.05, while the mean of selected individuals is 61.28, giving an S of 11.23. Using repeated-measures selection, the overall mean is 49.74, while the mean of selected individuals is 55.14, for an S of 5.39.

Hence, for the same selection intensity (best 5 out of 20), the selection differential is less for individuals with multiple measures. The expected response based on single measures is

$$R = h^2 S = 0.1 \cdot 11.23 = 1.12$$

Applying Equation 10.18b, the expected response under repeated measures is

$$R = h^2 \left(\frac{n}{1 + (n-1)r} \right) S = 0.1 \cdot 2.78 \cdot 5.39 = 1.50$$

Thus, the reduction in S under repeated measures is more than made up for by the increase in the regression slope, yielding a larger expected response relative to mass-selection. As a final check, from Equation 10.19b, the ratio of the accuracies of five-measure to single-measure is 1.67, slightly above the predicted ratio of response (for this realization of a simulation) of 1.33. For large n, this ratio approaches $1/\sqrt{r} \simeq 2.24$.

Adjustments for Non-overlapping Generations

So far, we have been assuming non-overlapping generations — all parents only reproduce in one generation interval. Domesticated animals, perennial plants, and many species in nature live multiple years and can have progeny over more than one year. In such cases, the response should be expressed in terms of response per year. To express the breeders' equation in terms of *rate of response* (typically the response per year), we first need to compute the **generation intervals** L_x (the average age of parents when progeny are born) for both sexes.

Example 10.7.   Compute the sire L_s and dam L_d generation intervals for the following age structure:

Age at Birth of Progeny

Sires	year 2	year 3	year 4	year 5	Total
Number	60	30	0	0	90

Dams	year 2	year 3	year 4	year 5	Total
Number	400	600	100	40	1140

The resulting sire generation interval is

$$L_s = \frac{2 \cdot 60 + 3 \cdot 30}{60 + 30} = 2.33,$$

while the dam generation interval is

$$L_d = \frac{2 \cdot 400 + 3 \cdot 600 + 4 \cdot 100 + 5 \cdot 40}{400 + 600 + 100 + 40} = 2.81$$

Incorporating the generation intervals, the yearly rate of response can be expressed as

$$R_y = \left(\frac{\bar{\imath}_s + \bar{\imath}_d}{L_s + L_d} \right) h^2 \sigma_p = \left(\frac{\bar{\imath}_s + \bar{\imath}_d}{L_s + L_d} \right) h \sigma_A \qquad (10.20)$$

This result (in a slightly different form) is due to Rendel and Robertson (1950). Thus, one way to increase response is to *reduce the generation intervals*, for example by using younger parents. The problem is that there is a tradeoff between generation interval and selection intensity. In species that are reproductively-limited (few offspring per dam), using younger dams means that a higher fraction of the dams must be chosen to replace the population (i.e., to keep the same number of animals in a herd). As a consequence, the selection intensity on these parents (which increases as fewer parents are chosen) is reduced. Chapter 23 examines the effects of age structure on selection response in much greater detail.

Maximizing Response Under the Breeders' Equation

We can combine both selection accuracy and generation interval to give a more general version of the breeders' equation:

$$R = \left(\frac{\bar{\imath}_s + \bar{\imath}_d}{L_s + L_d} \right) \rho(A, x) \, \sigma_A \qquad (10.21)$$

Here, x is the selection scheme measure used to choose the parents to form the next generation. Expressed this way, there are three components of response that the breeder has some control over:

- (i) selection intensity, $\bar{\imath}$
- (ii) generation interval, L
- (iii) selection accuracy, ρ

Note that not much can be done with increasing σ_A^2. Response is increased by decreasing L and/or increasing ρ and $\bar{\imath}$. We have already discussed tradeoffs between L and $\bar{\imath}$, and there are similar tradeoffs between L and ρ. Clearly, the longer we wait to allow a parent to reproduce, the more accurate we can predict their breeding value, as information from other relatives and from progeny-testing accumulates over time. However, these increases in ρ also result in increases in L. The optimal selection program must balance all of these competing interests.

Example 10.8.   As an example of the tradeoff between accuracy and generation intervals, consider a trait with $h^2 = 0.25$ and selection only on sires. One scheme is to simply select on the sire's phenotype, which results in a sire generation interval of 1.5 years. Alternatively, one might perform progeny testing to improve the accuracy of the selected sires. This results in an increase of the sire generation interval to (say) 2.5 years. Suppose in both cases, the dam interval is steady at 1.5 years.

Since the intensity of selection and additive genetic variation are the same in both schemes, the ratio of response under mass selection to response under progeny testing is just

$$\frac{R(\text{Sire phenotype})}{R(\text{progeny mean})} = \frac{\rho(A, \text{Sire phenotype})/(L_s + L_d)}{\rho(A, \text{progeny mean})/(L_s + L_d)}$$

Here, $\rho(A, \text{Sire phenotype}) = h = \sqrt{0.25} = 0.5$, with generation intervals $L_s + L_d = 1.5 + 1.5 = 3$. With progeny testing, (Example 10.4)

$$\rho(A, \text{progeny mean}) = \sqrt{\frac{n}{n + a}} = \sqrt{\frac{n}{n + 15}}$$

as $a = (4 - h^2)/(h^2) = 15$, with a total generation interal of $L_s + L_d = 2.5 + 1.5 = 4$. Hence,

$$\frac{R(\text{Sire phenotype})}{R(\text{progeny mean})} = \frac{0.5/3.0}{\sqrt{\frac{n}{n+15}}/4} = \frac{2}{3} \cdot \sqrt{\frac{n+15}{n}}$$

If (say) $n = 2$ progeny are tested per sire, this ratio is 1.95, giving a much larger rate of response under sire-only selection. For $n = 12$, the ratio is exactly one, while for a very large number of offspring tested per sire, the ratio approaches 2/3, or a 1.5-fold increase in the rate of response under progeny testing, despite the increase in sire generation interval.

Equation 10.21 highlights the importance to animal breeding of advances in reproductive technologies such as artificial insemination (AI) and multiple ovulation embryo transplant (MOET) schemes (Chapter 23). The more offspring a parent can produce, the stronger a selection intensity we can apply and still keep a required fixed number of animals in our herd. Hence, AI has resulted in the potential for far greater selection intensities (and unfortunately far more inbreeding) than would be possible under natural insemination. Likewise, MOET schemes to increase the number of offspring from females potentially allow for increases in the selection intensity on dams as well as decreases in the generation interval.

Maximizing the Economic Rate of Response

Example 10.8 also makes another important point – economics. Notice that by scoring more than 12 offspring we can obtain a larger expected rate of response using progeny testing. So, why not simply score (say) 30 progeny, giving a 122% rate of response relative to simple mass selection? The reason is economics – it costs money to raise all of those progeny. Hence, much of animal breeding is really more concerned with the *economic* rate of response – trying to maximize the rate of response per unit capital, although this point is often underappreciated, even by some breeders. Besides selecting for increases in quality and quantity, more and more selection is for increased efficiency, and hence greater economic gain. Weller (2001) presents a nice development of the economic aspects of animal breeding.

Example 10.9.   To see how our results from Example 10.8 would change when economics are taken into account, suppose the economic gain for unit of rate of response is C, while the cost of scoring a progeny is α. Hence the economic rate of response to simple mass selection (sire phenotype only) is

$$EcR(\text{Sire phenotype}) = C \cdot 0.5/3 = 0.167 \cdot C$$

while the economic rate using progeny testing is the gain minus the cost of scoring n progeny,

$$EcR(\text{progeny mean}) = C \cdot \frac{1}{4} \sqrt{\frac{n}{n+15}} - \alpha \cdot n$$

Thus, the ratio of economic response can we written as

$$\frac{EcR(\text{progeny mean})}{EcR(\text{Sire phenotype})} = 1.5 \cdot \sqrt{\frac{n}{n+15}} - n \cdot \frac{\alpha}{C}$$

With a little work, one can show that if $\alpha/C > 0.0076$ (the cost per progeny is greater than 0.76% of the economic gain per unit of response), then simple selection on the sire always result in a greater rate of economic gain than progeny testing. The figure below (for $\alpha/C = 0.01$ and 0.005) shows there is an optimal number of progeny for maximizing response when progeny testing is chosen.
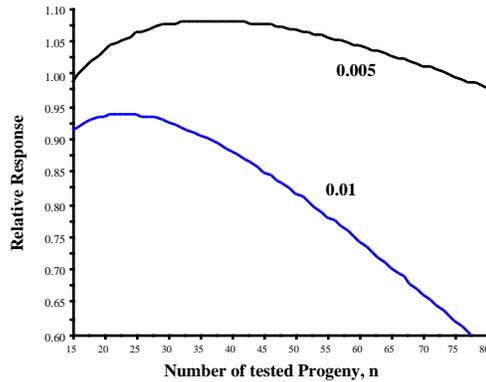
Prelude to the Multivariate Breeders' Equation

Several alternate parameterizations of the breeders' equation are popular in the evolutionary literature, and we introduce them here to motivate their multivariate versions (discussed in Chapters 30 and 31). Expressing the heritability in terms of additive genetic and phenotypic variance,

$$R = \sigma_A^2 \, \sigma_z^{-2} \, S \tag{10.22a}$$

While this decomposition seems rather trivial, it suggests that the multivariate version of the breeders' equation (under appropriate linearity assumptions) is given by

$$\mathbf{R} = \mathbf{G}\mathbf{P}^{-1}\mathbf{S} \tag{10.22b}$$

where $\mathbf{R}$ is the vector of responses (cross-generational changes in the mean of each character), $\mathbf{S}$ the vector of selection differentials for each character, and $\mathbf{G}$ and $\mathbf{P}$ are the additive

genetic and phenotypic covariance matrices (LW Chapter 21). We will examine this equation, and its consequences, in Chapters 30–36.

A second parameterization follows from the Robertson-Price identity, $S = \sigma(z, w)$. Since the slope of the least-squares linear regression of relative fitness w on phenotypic value z is $\beta = \sigma(z, w)/\sigma_z^2$, it follows that $S = \sigma_z^2 \beta$. The regression slope β is called the **directional selection gradient**, for reasons that will become apparent shortly. Substituting $S = \sigma_z^2 \beta$ into Equation 10.1 gives

$$R = \sigma_A^2 \beta \tag{10.23a}$$

Noting that $\sigma_A^2 = h^2 \sigma_z^2$, we can rearrange this to give

$$\frac{R}{\sigma_z^2} = h^2 \beta \tag{10.23b}$$

so that $h^2 \beta$ is the expected response in units of phenotypic variance. The multiple linear regression of relative fitness w on a vector $\mathbf{z}$ of characters has slope $\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{S}$ (LW Equation 8.10c), where $w = 1 + \boldsymbol{\beta}\mathbf{z}^T + e$. Substituting into Equation 10.22b gives

$$\mathbf{R} = \mathbf{G}\,\boldsymbol{\beta} \tag{10.23c}$$

The importance of the selection gradient is that under appropriate conditions it relates how a within-generation change in the vector of trait means maps into a change in the mean fitness of a population. In particular, if $w(z)$ denotes the expected fitness of an individual with character value z, then when phenotypes are normally distributed, and fitness is frequency-independent (individual fitnesses are not a function of the mean of the character), Lande (1976) showed that the directional selection gradiate statisfies $\beta = \partial \ln \overline{w}/\partial \mu$. Hence we can express the breeders equation as

$$R = \sigma_A^2 \left(\frac{\partial \ln \overline{w}}{\partial \mu} \right) \tag{10.23d}$$

The multivariate version is the **gradient of mean fitness** (Chapter 30) with respect to the vector of character means,

$$\boldsymbol{\beta} = \frac{\partial \ln \overline{w}}{\partial \boldsymbol{\mu}}$$

giving

$$\mathbf{R} = \mathbf{G}\,\frac{\partial \ln \overline{w}}{\partial \boldsymbol{\mu}} \tag{10.23e}$$

The gradient terms represents the changes in character means that produce the maximal change in population fitness, while the actual response involves the product of this vector with the genetic covariance matrix $\mathbf{G}$. The resulting response vector is generally not parallel to $\boldsymbol{\beta}$ and hence the genetic covariance structure causes the character means to change in a direction that does not necessarily result in the optimal change in population fitness. We examine these issues in detail in Chapter 30.

These alternative expressions point out that we can distinguish between **phenotypic selection**, the change in a phenotypic distribution *within* a generation (measured by S or β), and the **evolutionary response to selection**, the transmission of these within-generation changes to the next generation (R). Lande and Arnold (1983) and Arnold and Wade (1984a,b), following Fisher (1930, 1958) and Haldane (1954), have stressed the utility of this approach. Attempts to measure selection by comparing phenotypic distributions across generations are confounded by inheritance, as R depends on β through σ_A^2 (or $\mathbf{G}$ in the multivariate

case). For example, if a character has no additive genetic variation ($\sigma^2_A = 0$), there is no permanent change in the mean across generations regardless of the selection intensity, and a cross-generation comparison would conclude that selection was not operating. Chapters 28 and 29 examine in detail methods for estimating the nature of phenotypic selection in natural populations.

TRUNCATION SELECTION

Truncation selection is by far the commonest form of artificial selection in plant and animal breeding and in laboratory experiments. Under truncation selection, only the individuals with the largest (or smallest) trait values are allowed to reproduce (Figure 10.2). Truncation selection is usually described by either the percent p of the population saved or the threshold phenotypic value T below (above) which individuals are **culled**. The investigator usually sets these in advance of the actual selection. Hence, while S is trivially computed *after* the parents are chosen, we would like to *predict* the expected selection differential given either T or p. Specifically, given either T or p, what is the expected mean of the selected parents? In our discussions of this issue, we will initially assume a large number of individuals are saved, before turning to complications introduced by finite population size.
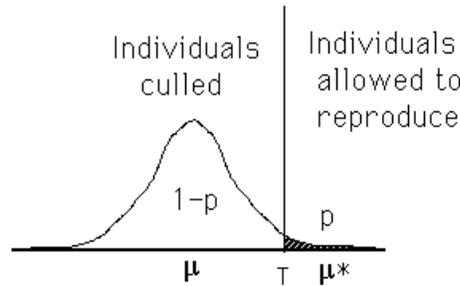

Figure 10.2.   Under truncation selection, the uppermost (lowermost) fraction p of a population is selected to reproduce. Alternatively, one could set a threshold level T in advance, above (below) which individuals are allowed to reproduce. To predict response given either p or T, we need to know the mean of the selected tail (μ^*), from which we can compute $S = \mu^* - \mu$ and then apply the breeders' equation.

Selection Intensities and Differentials Under Truncation Selection

Given a threshold cutoff T, the expected mean of the selected adults is given by the conditional mean, $E(z \mid z \geq T)$. Generally it is assumed that phenotypes are normally distributed, and we use this assumption throughout. With initial mean μ and variance σ^2, this conditional mean is given by LW Equation 2.14, which gives the expected selection differential as

$$S = \varphi\left(\frac{T - \mu}{\sigma}\right) \frac{\sigma}{p} \tag{10.24}$$

where $p = \Pr(z \geq T)$ is the fraction saved and $\varphi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ is the unit normal density function evaluated at x.

Generally, it is the fraction to be saved p (rather than T) that is preset by the investigator (for example, if we save the uppermost 5 percent, what is the expected S?). Given p, to apply

Equation 10.24, we must first find the threshold value T_p satisfying $\Pr(z \geq T_p) = p$. Notice that T in Equation 10.24 enters only as $(T - \mu)/\sigma$, which transforms T_p to a scale with mean zero and unit variance. Hence,

$$\Pr(z \geq T_p) = \Pr\left(\frac{z - \mu}{\sigma} > \frac{T_p - \mu}{\sigma}\right) = \Pr\left(U > \frac{T_p - \mu}{\sigma}\right) = p$$

where $U \sim \mathrm{N}(0, 1)$ denotes a unit normal random variable. Define $x_{[p]}$, the **probit transformation** of p (LW Chapter 11), by

$$\Pr(U < x_{[p]}) = p \tag{10.25a}$$

Hence

$$\Pr(U \geq x_{[1-p]}) = p \tag{10.25b}$$

It immediately follows that $x_{[1-p]} = (T_p - \mu)/\sigma$, and Equation 10.24 gives the expected selection intensity as

$$\bar{\imath} = \frac{S}{\sigma} = \frac{\varphi(x_{[1-p]})}{p} \tag{10.26a}$$

One can obtain $x_{[1-p]}$, and hence $\bar{\imath}$, from standard statistical packages. For example, in **R**, the command **qnorm(1-p)** returns $x_{[1-p]}$, while **dnorm(qnorm(1-p))/p** returns $\bar{\imath}$ for a pre-assigned p value (e.g., **p <- 0.05**). Alternatively, a number of approximations have been suggested for Equation 10.26a. Assuming normality, Smith (1969) has suggested

$$\bar{\imath} \simeq 0.8 + 0.41 \ln\left(\frac{1}{p} - 1\right) \tag{10.26b}$$

Simmonds (1977) found that this approximation is generally quite good for $0.004 \leq p \leq 0.75$, and offered alternative approximations for p values outside this range. The most precise approximation is Saxton's (1988), with

$$\bar{\imath} \simeq \frac{2.97425 - 3.35197\,p^{0.2} - 1.9319\,p^{0.4} + 2.3097\,p^{0.6}}{0.51953 + 0.88768\,p^{0.2} - 2.38388\,p^{0.4} + p^{0.6}} \tag{10.26c}$$

Likewise, Montaldo (1997) gave an approximation for the standard truncation value $z = (T - \mu)/\sigma$ in terms of $\bar{\imath}$,

$$z \simeq -1.411053 + 2.077585\,\bar{\imath} - 0.454318\,\bar{\imath}^2 + 0.097020\,\bar{\imath}^3 - 0.008137\,\bar{\imath}^4 \tag{10.26d}$$

Example 10.10.   Consider selection on a normally distributed character in which the upper 5% of the population is saved ($p = 0.05$). Here $x_{[1-0.05]} = 1.645$ as $\Pr[U \geq 1.645] = 0.05$. Hence,

$$\bar{\imath} = \frac{\varphi(1.645)}{0.05} = \frac{0.103}{0.05} \simeq 2.06$$

Applying Equation 10.6b gives the expected response to this amount of selection as $R = h^2\,\sigma\,2.06$. Smith's approximation gives the selection intensity as

$$\bar{\imath} \simeq 0.8 + 0.41 \ln\left(\frac{1}{0.5} - 1\right) \simeq 2.01$$

which is quite reasonable. Saxton's approximation gives $\bar{\imath} \simeq 2.06$.

Correcting the Selection Intensity for Finite Samples

If the number of individuals saved is small, Equation 10.24 overestimates the selection differential because of sampling effects (Nordskog and Wyatt 1952, Burrows 1972). To see this, assume M adults are sampled at random from the population and the largest N of these are used to form the new generation, giving $p = N/M$. The expected selection coefficient is computed from the distribution of **order statistics**. Rank the M observed phenotypes as $z_{1,M} \geq z_{2,M} \ldots \geq z_{M,M}$ where $z_{k,M}$ is referred to as the kth order statistic when M observations are sampled. The expected selection intensity is given by the expected mean of the N selected parents, which is the average of the first N order statistics,

$$E(\bar{\imath}) = \frac{1}{\sigma} \left(\frac{1}{N} \sum_{k=1}^{N} E(z_{k,M}) - \mu \right) = \frac{1}{N} \sum_{k=1}^{N} E(z'_{k,M})$$

where $z'_{k,M} = (z_{k,M} - \mu)/\sigma$ are the **standardized order statistics**. Properties of order statistics have been worked out for many distributions (Kendall and Stuart 1977, David 1981, Sarhan and Greenberg 1962, Harter 1970a,b). For the unit normal, see Harter (1961) for expected values and Sarhan and Greenberg (1962) for variances and covariances. Figure 10.3 plots exact values for the expected selection intensity for small values of N when phenotypes are normally distributed. Note that finite population size results in Equation 10.26a overestimating the actual selection intensity, although the difference is small unless N is very small.
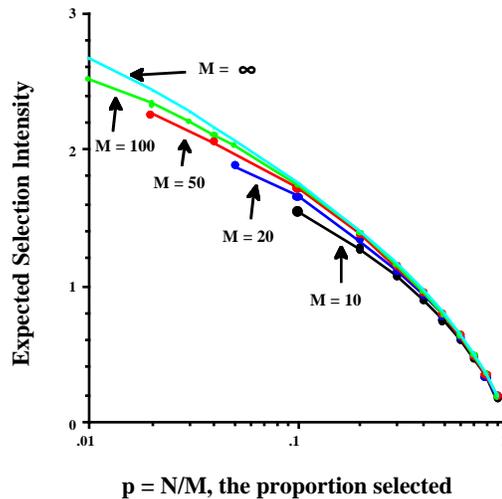

Figure 10.3. The expected selection intensity $E(\bar{\imath})$ under truncation selection with normally-distributed phenotypes, as a function of the total number of individuals measured M and the fraction of these saved $p = N/M$, N being the number of these adults allowed to reproduce. The curve $M = \infty$ is given by using Equation 10.26a, which is exact if N and M are infinite. The values on the curves for $M = 10, 20, 50,$ and 100

were obtained from the average of the expected values of the N largest unit normal order statistics (Harter 1961). Note that Equation 10.26a is generally a good approximation, unless N is very small.

Burrows (1972) developed a finite-sample approximation for the expected selection intensity for any reasonably well-behaved continuous distribution. Using the standardized variable $y = (z-\mu)/\sigma$ simplifies matters considerably. Letting $\phi(y)$ be the probability density function of the phenotypic distribution, and y_p the truncation point (i.e., $\Pr(y \geq y_p) = p$), Burrows' approximation is

$$E(\bar{\imath}_{(M,N)}) \simeq \mu_{y_p} - \frac{(M-N)\,p}{2N(M+1)\,\phi(y_p)} \qquad (10.27a)$$

where

$$\mu_{y_p} = E\,(y\,|\,y \geq y_p\,) = \frac{1}{p}\int_{y_p}^{\infty} x\,\varphi(x)\,dx$$

is the truncated mean, which can be obtained by numerical integration. Since the second term of Equation 10.27a is positive, if M is finite the expected truncated mean overestimates the expected standardized selection differential.

For a unit normal distribution, $\mu_{y_p} = \varphi(y_p)/p$, giving

$$\begin{aligned} E(\bar{\imath}_{(M,N)}) &\simeq \bar{\imath} - \left[\frac{M-N}{2N(M+1)}\right]\frac{1}{\bar{\imath}} \\ &= \bar{\imath} - \left[\frac{1-p}{2p(M+1)}\right]\frac{1}{\bar{\imath}} \end{aligned} \qquad (10.27b)$$

where $\bar{\imath}$ is given by Equation 10.26a. Lindgren and Nilsson (1985) examined the accuracy of Burrows' approximation. They found the error is rather constant independent of M for $M > 10$, and decreases with N roughly as $1/N^2$. They suggest that Burrows' approximation can be used for all N if an error of 0.025 (less than 5% of the selection intensity) is acceptable, for $N > 2$ if an error less than 0.01 is acceptable, and for $N > 6$ if an error of less than 0.001 is required.

Bulmer (1980) suggests an alternative finite-sample approximation for $E(\bar{\imath}_{(M,N)})$ when phenotypes are normally distributed, using Equation 10.26a with p replaced by

$$\widetilde{p} = \frac{N + 1/2}{M + N/(2M)} \qquad (10.27c)$$

Example 10.11.   Consider the expected selection intensity on males when the upper 5% of the sampled males are used to form the next generation and phenotypes are normally distributed. If the number of males sampled is very large, then from Example 10.10, the expected selection intensity is $\bar{\imath} \simeq 2.06$. Suppose, however, that only 20 males are sampled so that only the largest male is allowed to reproduce in order to give $p = 0.05$. The expected value for this male is the expected value of the largest order statistic for a sample of size 20. For the unit normal, the expected value of the largest order statistic in a sample of 20 is $\simeq 1.87$ (Harter 1961) and hence $E(\bar{\imath}_{(20,1)}) \simeq 1.87$. There is considerable spread about this expected value, as the standard deviation of this order statistic is 0.525

(Sarhan and Greenberg 1962). How well do the approximations of $E(\bar{\imath}_{(20,1)})$ perform? Burrows' approximation gives

$$E(\bar{\imath}_{(20,1)}) \simeq 2.06 - \frac{(20-1)}{2\,(20+1)\,2.06} = 2.06 - 0.22 = 1.84$$

while Bulmer's approximation uses

$$\widetilde{p} = \frac{1+1/2}{20+1/40} \simeq 0.075$$

which gives $x_{[1-0.075]} \simeq 1.44$. Since $\varphi(1.44) = 0.1415$,

$$E(\bar{\imath}_{(20,1)}) \simeq 0.1415/0.075 \simeq 1.89$$

A final correction for finite population size was noted by Rawlings (1976) and (especially) Hill (1976, 1977c). If families are sampled, such that n individuals are chosen per family, then the selection intensity is further reduced because the correlations between family members. In particular, if a total of M individuals are sampled, with n individuals per family, then Burrows' correction (Equation 10.27b) is modified to become

$$\bar{\imath} - \left[\frac{1-p}{2p(M+1)(1-\tau+\tau/n)} \right] \frac{1}{\bar{\imath}} \tag{10.28}$$

where τ is the intra-class correlation of family members. Burrows (1975) also gives expressions of the variance of $\bar{\imath}$ in finite populations. However, since the distribution of realized differentials is asymmetry (Figure 10.4), the variance alone is not sufficient for computing confidence intervals.
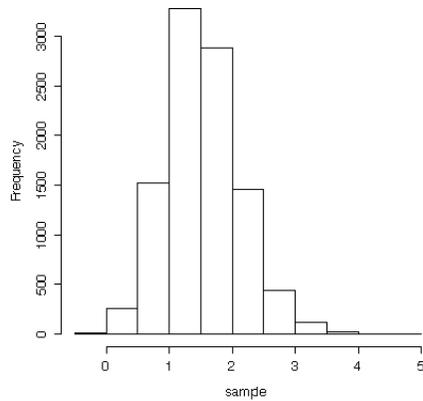

Figure 10.4.   The distribution of 10,000 random draws of $\bar{\imath}_{(10,1)}$, the largest order statistic in a sample of 10. The mean value is 1.54, as opposed to the expected value of $\bar{\imath} = 1.75$ for $p = 0.1$ in an infinite population. Burrows' approximation gives an expected value of 1.52. Notice that there is a consideriable spread of values and that the distribution is not symmetric about the mean, but rather is skewed towards higher values.

RESPONSE WITH DISCRETE TRAITS: BINARY CHARACTERS

A potentially interesting complication for the analysis of quantitative traits are **binary characters**, those characterized simply by presence/absence (such as normal/diseased). Our basic model to this point has assumed a continuous character that is influenced by both genetic and environmental value, which initially seem a bit at odds with the notion of a binary trait. The connection (LW Chapters 11 and 25) is that we assumed some underling continuous trait that maps into the observed discrete trait. Roff (1996) reviews a number of examples of such threshold-determined morphological traits in animals.

The Threshold/Liability Model

As discussed in LW Chapters 11 and 25, discrete characters can often be modeled by mapping an underlying continuous character, the **liability**, to the observed discrete character states (Figure 10.5). Consider the simplest case where the character is either present ($z \geq T$) or absent ($z < T$). Let μ_t be the mean liability and q_t the frequency of individuals displaying the character in generation t. If liability is well enough behaved to satisfy the assumptions of the breeders' equation (e.g., a linear biparental-offspring regression, no epistasis, genotype-environmental interactions or correlations), then $\mu_{t+1} = \mu_t + h^2 S_t$. We index the selection differential S_t by generation, as the amount of selection on threshold characters changes each generation. The problem is to estimate the mean liability μ_t from the observed frequency of the trait q_t. Our analysis will be restricted to a single threshold, but extension to multiple thresholds is straightfoward (see Lande 1978, Korsgaard et al. 2002).
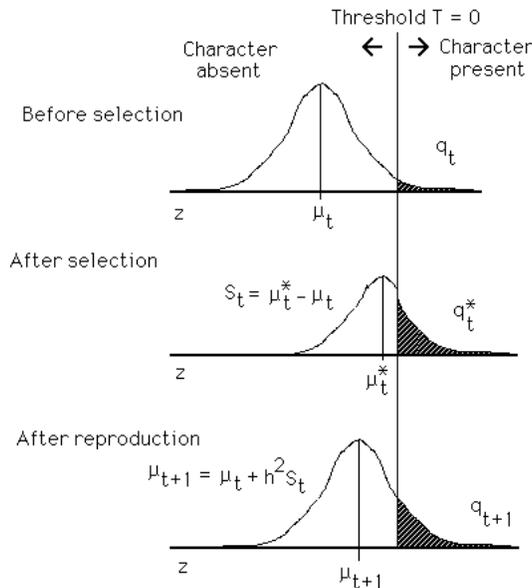

Figure 10.5.   Selection on a character with a single threshold T. z is the value on the underlying scale of liability. Assuming z is a well-behaved quantitative character, $\mu_{t+1} = \mu_t + S_t h^2$, where $S_t = \mu_t^* - \mu_t$. Using the probit transform (Equation 10.25b) we can translate q, the frequency of individuals displaying the character, into μ, the mean of z.

If the values on the underlying scale are normally distributed, we can choose a scale that sets the threshold value at $T = 0$ and assigns z a variance of one. Since $z - \mu_t$ is a unit normal, $\Pr(z \geq 0) = \Pr(z - \mu_t \geq -\mu_t) = \Pr(U \geq -\mu_t) = q_t$ and from Equation 10.25b

$$\mu_t = -x_{[1-q_t]} \tag{10.29}$$

where $x_{[p]}$, the probit transformation of p is given by Equation 10.25a. For example, if 5% of a large population displays the trait, $\mu = -x_{[0.95]}$. From normal probability tables, $\Pr(U < 1.645) = 0.95$, hence $x_{[0.95]} = 1.645$ and $\mu = -1.645$. For small samples, estimation of μ requires the use of order statistics.

The response to selection, as measured by the change in the frequency of the character,

is

$$q_{t+1} = \Pr(U \geq -\mu_{t+1})$$
$$= \Pr(U \geq -\mu_t - h^2 S_t)$$
$$= \Pr(U \geq x_{[1-q_t]} - h^2 S_t) \qquad (10.30)$$

It remains to obtain $S_t = \mu_t^* - \mu_t$, where μ^* is the mean liability value in the selected parents in generation t. The selected population may consist entirely of adults displaying the character. However, more individuals than this may be required to keep the population at constant size, especially if q_t is small. In this case, the selected adults consist of two populations: those displaying the trait (hence $z \geq 0$) and those not ($z < 0$). Letting p_t be the fraction of selected adults displaying the character,

$$\mu_t^* = (1 - p_t)\, E(z|z < 0; \mu_t) + p_t\, E(z|z \geq 0; \mu_t) \qquad (10.31)$$

Applying LW Equation 2.14 and noting that the unit normal density function satisfies $\varphi(x) = \varphi(-x)$, gives

$$E(z|z \geq 0; \mu_t) = \mu_t + \frac{\varphi(\mu_t)}{q_t}$$
$$E(z|z < 0; \mu_t) = \mu_t - \frac{\varphi(\mu_t)}{1 - q_t}$$

Substituting into Equation 10.31 gives

$$S_t = \mu_t^* - \mu_t = \frac{\varphi(\mu_t)}{q_t} \frac{p_t - q_t}{1 - q_t} \qquad (10.32)$$

As expected, if $p_t > q_t$, $S_t > 0$. S_t depends critically on q_t and is very unlikely to remain constant over several generations of selection. Maximal selection occurs if only individuals displaying the trait are saved ($p_t = 1$), in which case Equation 10.32 reduces to $S_t = \varphi(\mu_t)/q_t$.

The careful reader might ask why did we simply not estimate μ_t^* using $x_{[1-q_t^*]}$ in Equation 10.31 (i.e., using the frequency q^* of the trait in the selected parents)? The reason is that the distribution of z values in selected parents is a weighted average of two truncated normal density functions (Equation 10.31), and this distribution is not normal. However, we assume that normality is restored in the liability distribution at the start of the next generation due to segregation plus the addition of the environmental value. We examine the validity of this assumption in Chapter 24.

Example 10.12.   Consider a threshold character whose liability has heritability $h^2 = 0.25$. What is the expected response to selection if the initial frequency of individuals displaying the character is 5% and selection is practiced by selecting only adults displaying the character? As was calculated earlier, $q_0 = 0.05$ implies $\mu_0 = -1.645$ (the mean liability is 1.65 standard deviations below the threshold). In each generation, only individuals displaying the trait are saved. Equation 10.32 gives

$$S_0 = \varphi(-1.645)/0.05 \simeq 0.106./0.05 \simeq 2.062$$

giving
$$\mu_1 = \mu_0 + 0.25 \cdot S_0 = -1.645 + 0.25 \times 2.062 = -1.129$$

and from Equation 10.30,

$$q_1 = \Pr(U \geq -\mu_1) = \Pr(U \geq 1.129) = 0.129$$

Thus, after one generation of selection, the character frequency is expected to increase to 12.9%. Changes in q and S after further iterations (again where selection occurs by only allowing adults displaying the trait to reproduce) are plotted below, where solid circles denote q_t, open squares denote S_t. Note that only six generations are required to increase the frequency of the trait to 50% ($\mu = 0$).
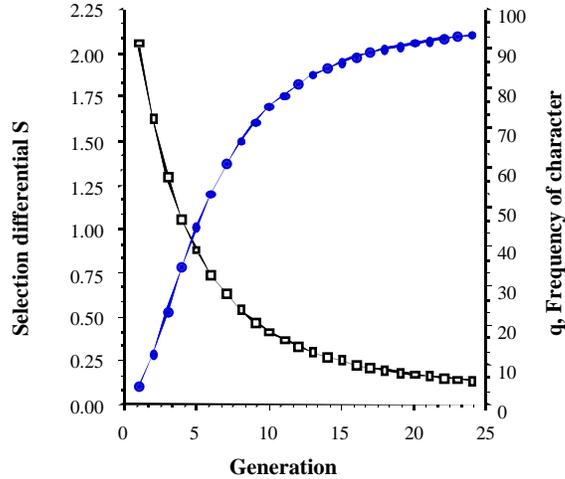
Example 10.13. Matsumura (1996) examined the effectiveness of selection on wing morphs in females of the whitebacked planthopper *Sogatella furcifera*. While this hemipterian is a serious rice pest in Japan, it is unable to overwinter. Rather, each year it undergoes migration from southern China to recolonize Japan. Females exhibit two wing morphs, while males have a single winged morph. *Macropterous* females are fully winged while *brachypterous* females have reduced wings and cannot fly. Further, increasing nymphal population density increases the frequency of macropterous females (leading to increased dispersal). Using three replicate experiments at each of three densities, Matsumura selected for increased macroptery in one replicate, decreased in another, and a control (no selection) in the third. For the replicates with one nymph per generation, roughly 40-90 adults were scored, and 20 chosen to form the next generation. The resulting data for the first five generations in the up-selected line was as follows (Matsumura pers. comm.):

Generation	q	μ	p	S	R
1	0.224	-0.76	1.00	1.34	0.35
2	0.340	-0.41	0.80	0.75	0.54
3	0.551	0.13	1.00	0.72	0.33
4	0.675	0.45	1.00	0.53	-0.07
5	0.651	0.39	1.00	0.57	0.16
6	0.708	0.55			

Here q is the frequency of macroptery before selection and p the frequency of macroptery in the selected parents. Translation from q into the mean liability μ follows from Equation 10.29. Hence, the response to selection in generation one is

$$R(1) = \mu_2 - \mu_1 = -x_{[1-0.340]} - (-x_{[1-0.224]}) = -0.41 - (-0.76) = 0.35$$

Likewise, the total response was

$$\mu_6 - \mu_1 = 0.55 - (-0.76) = 1.31$$

Selection differentials were calculated using Equation 10.32. For example, for generation two,

$$S_2 = \frac{\varphi(\mu_2)}{q_2} \frac{p_2 - q_2}{1 - q_2} = \frac{\varphi(-0.41)}{0.34} \frac{(0.80 - 0.34)}{1 - 0.34} = 0.75$$

The total selection differential is

$$\sum_{i=1}^{5} S_i = 3.91$$

One key summary statistic for any selection experiment is the *realized heritability*, the ratio of response to selection differential. As detailed in Chapter 14, there are several ways to compute this with a multi-generational selection experiment. One simply estimate is just the total response over the total differential,

$$\widehat{h^2} = \frac{\sum R_i}{\sum_i S_i} = \frac{1.31}{3.91} = 0.33$$

Hence, the estimated heritability of the underlying liability for macroptery is around 30%.

One important feature about selection on threshold traits is that *response is not necessarily symmetric* – a selected 5% increase in the trait may not yield the same response as a selected 5% decrease. The reason for this is that the mapping from the phenotype to the underlying liability is non-linear. Even though the parent-offspring regression on the liability scale is linear, the parent-offspring regression on the phenotypic level is not.

Finally, in animal breeding it is often natural to select on the estimated breeding values of individuals, using information from relatives in addition to their phenotypic value (Chapters 17, 34). Foulley (1992) examines the response to binary traits when selection occurs directly on estimated breeding values (Foulley et al. 1983), as opposed to observed phenotypes. Korsgaard et al. (2002) examine the response using estimated breeding value when there are $k \geq 2$ thresholds in the mapping of liability into different character states.

Logistic Regressions and the Logistic Distribution

The threshold approach is one way to model the mapping from some underlying continuous variable into a discrete character space. This is a deterministic approach, in that all individuals with liability values above the threshold value T show the trait, while all those below the threshold do not display the trait. A potentially more realistic model might (at first blush) appear to be more stochastic, with the underlying liability mapping into a probability of displaying the trait. Under the threshold model, this probability is one at or above T, and zero below T.

From a biological standpoint, a more appealing model might be that the probability of displaying the trait is a monotonically increasing function of the underlying liability value, which approaches zero for very low values and one for very high values. Under this model, a high value of the liability does not imply that the trait is always displayed, nor does a low liability value imply that the trait is never display. One common model for translating a liability value z into a probability of showing the trait is the **logistic regression**,

$$\Pr(X = 1 \,|\, z) = \frac{1}{1 + e^{-\alpha z}} \tag{10.33a}$$

where $\alpha > 0$ is a scaling factor. Note that for $z << -1$, $\Pr(X = 1) \simeq 0$, while for $z >> 1$, $\Pr(X = 1) \simeq 1$. Likewise, note that at $z = 0$, $\Pr(X = 1) = 1/2$.
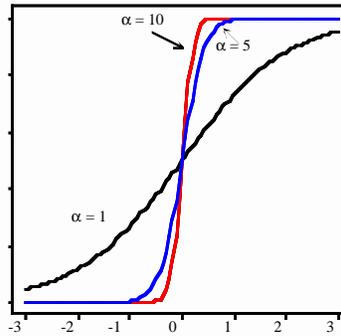

Figure 10.6.   The logistic function for differential values of α. Note for α values in excess of five that the logistic function essentially recovers the discrete threshold model.

As shown in Figure 10.6, the larger the α value, the more abrupt the transition from a probability near zero to a probability near one. Thus, for large α, the logistic mapping essentially equals the threshold model.

A slightly more general version of the logistic regression, which shifts the 50% probability value to $z = m$, is

$$\Pr(X = 1 \,|\, z) = \frac{1}{1 + e^{-\alpha\,(z-m)}} \tag{10.33b}$$

Notice that the logistic regression can be considered as a *cumulative distribution function* (LW Chapter 2) for the threshold value T, e.g., $\Pr(x = 1 \,|\, z) = \Pr(T \leq z)$. Hence, we can define the **logistic distribution** $\phi(x, \alpha, m)$ as the function which satisfies

$$\int_{\infty}^{z} \phi(x, \alpha, m)dx = \frac{1}{1 + e^{-\alpha\,(z-m)}} \tag{10.34a}$$

as the integral gives the cumulative distribution function, which is the logistic function. We can obtain this function by simply taking derivatives of both sides to obtain

$$\phi(x, \alpha, m) = \frac{\alpha\,e^{-\alpha\,(z-m)}}{(1 + e^{-\alpha\,(z-m)})^2} \tag{10.34b}$$

The mean, variance, and skew for this distribution are (Johnson and Kotz 1970b)

$$\mu = m, \quad \sigma^2 = \frac{1}{3}\left(\frac{\pi}{\alpha}\right)^2, \quad \mu_3 = 0 \tag{10.34c}$$

As shown in Figure 10.7, the normal and logistic distributions have very similar cumulative distribution functions. Indeed, for a unit normal random variable U,

$$\Pr(U \leq x) \simeq \frac{1}{1 + e^{-\alpha\,x}}, \quad \text{where} \quad \alpha = \frac{\pi}{\sqrt{3}} \tag{10.35}$$

which is the logistic distribution with variance one (see Equation 10.34c).

Thus, we have two approaches for mapping liability values into binary traits: the strict threshold approach (a deterministic mapping of liability into the discrete trait) or the logistic regression approach (a stochastic mapping translating a liability value into a probability of observing the trait). For most purposes, the threshold model is a reasonable approach, even if the underlying mapping is stochastic. However, as Example 10.14 (below) illustrates, the logistic distribution can prove useful for the analysis of threshold values when we know the liability itself.
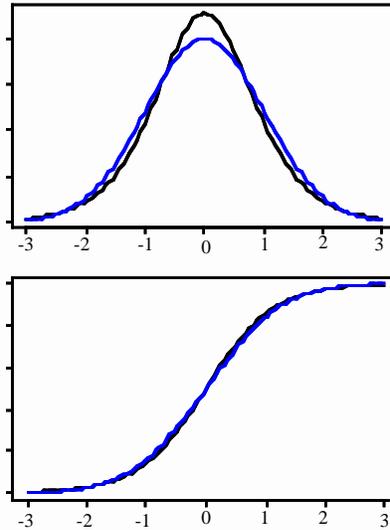

Figure 10.7.   A compairson of the unit normal and unit logistic ($\mu = 0$, $\sigma^2 = 1$) distributions. **Top**: Probability density functions: the logistic is more peaked, with positive scaled kutosis. **Bottom**: Cumulative distribution functions. As the figure shows, the normal and logistic distributions have extremely similar cumulative distributions.

Direct Selection on the Threshold T

We have been presenting the threshold model as having a fixed value for T. Of course, it is biologically quite reasonable to imagine that there is variation in the threshold value itself (Hazel et al. 1990). For example, suppose our trait of interest appears when the size of an organism exceeds some critical threshold T (Example 10.14). Hence, in the threshold model the liability value is size, with $z = g_s + e_s$, with a genotypic and an environmental value for size. Likewise, suppose the threshold value T itself is not fixed, but rather is also a quantitative trait, with $T = g_T + e_T$. The trait appears when $z \geq T$, or

$$g_s + e_s - (g_T + e_T) = g_s - g_T + e_s - e_T \geq 0$$

Thus, even though both the liability and threshold values are variable, we can simply consider a single new "risk liability", the difference between the liability value and the threshold value, and the analysis proceeds as above. Hence, if our interest is simply on presence/absence of the binary trait, we don't have to worry as to whether the liability or threshold, or both, show variation. However, as Example 10.14 shows, there are situations where we can directly measure the threshold value itself, in which case we can directly measure the heritability of the threshold level by a selection experiment.

Example 10.14. An interesting analysis of selection on a threshold trait using logistic regressions is given by Wesselingh and de Jong (1995), who studied the connection between plant size and flowering in hound's-tongue (*Cynoglossum officinale*). This plant is a *facultative biennial*, which means that like an annual plant, it flowers only once, but unlike an annual, it may live several years before flowering. This represents a trade-off between the risk of survival over several years versus a larger seed set with larger size at flowering. For *Cynoglossum* it has been shown that vernalization (cold treatment) followed by an appropriate photoperiod is required for flowering. However, unless plants are at (or above) a certain threshold size, they are unresponsive to vernalization and hence grow without flowering through the next growing cycle.

The authors were interested in the threshold size itself that triggers the binary trait (vernalization sensitivity), and in particular whether this size is both variable and heritable. To examine this, the authors grew plants for different number of days (ranging from 31 to 86) to generate plants of different sizes before vernalization treatment. This generated two selection groups: the smallest plants that flowered following vernalization where chosen as the low-line parents, while those plants that did not respond to the first vernalization treatment were allowed to grow a second cycle and these were chosen as the parents for the high lines, and the F_1 offspring for intercrosses within each set of selected parents were examined.

The data available to the authors were 0/1 (insensitive/sensitive to vernalization) data as a function of size. To directly estimate the distribution of threshold sizes, they performed a logistic regression on these data, using maximum likelihood to fit the variance (α) and mean (m) terms of Equation 10.33b. Data for the high and low lines are presented below along with the ML solution. Each individual has a zero/one data point (individual ticks), while the circles represent the average value for weight classes with more than ten individuals.
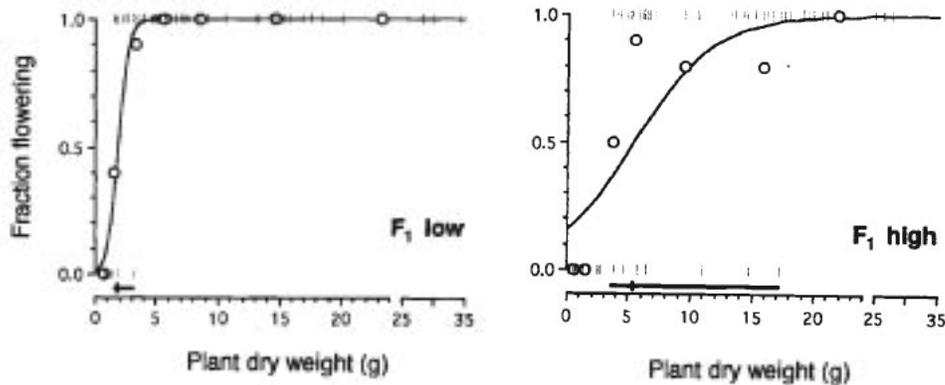
Logistic regressions were estimated for the P generation from which the high and low lines were derived, and three regressions corresponding to a replicate of P_r contemporaneously-grown with the progeny of the low parents (low line) and the high parents (high line). The ML estimates of the mean m (which corresponds to the weighting yielding 50% flowering) and α for these regressions were

Line	m	α
P	5.12	0.45
P_r	3.30	0.97
Low	1.85	2.58
High	5.41	0.31

Using these estimates for α and m, Equation 10.33b yields the expected percent of vernalization sensitivity (flowering) for a given weight. For example, for a 3 gram plant with is 0.43 in the control, but 0.96 in the low line. Basing estimates of response using the contemporaneously grown P_r line as a standard, the response in the high line was 5.41 - 3.30 = 2.11, while the response in the low line was 1.85 - 3.30 = -1.45. Likewise, the selection truncation point for the low line is the largest low parent (2.74 grams, or 25.5% of the left tail of the founding P line), while the smallest flowering high parent was 9.95 grams (corresponding to the upper 12.2% of the P lines). From Equation 10.26a, these translate into selection intensities of 1.26 and 1.66, respectively. To obtain the selection differentials S for each line, recall that $S = \bar{\imath}\sigma_p$. To estimate σ_p, the authors note that the 0.25 quartile for a normal distribution is 0.674σ from the mean. Although the assumption is that the threshold values follow a logistic distribution, the cumulative probability functions are rather similar for both the normal and logistic (Figure 10.7). Hence, taking the observed 0.25 quartile (in the P lines) of 2.68, and its mean of 5.12, suggests

$$\sigma_p = \frac{5.12 - 2.68}{0.674} = 3.63$$

(One could also use Equation 10.34c to translate the estimated α value into an estimate of the variance of the logistic distribution, which, for $\alpha = 0.45$ gives $\sigma_p = 4.94$; however this is not as robust an estimate as the 0.25 quartile which is much less sensitive to outliers.)

Thus, the response, selection intensity, and estimated heritability $\widehat{h^2} = R/S$ for the high and low lines are

Line	$\bar{\imath}$	S	R	$\widehat{h^2} = R/S$
Low	1.26	4.58	1.45	0.32
High	1.66	6.02	2.11	0.35

Thus, there is heritable variation in the threshold size, as there was response to selection for both larger and smaller threshold sizes. Further, the estimated heritability (based on the single-generation response to selection) was around 0.3.

RESPONSE WITH DISCRETE TRAITS: POISSON-DISTRIBUTED CHARACTERS

Many discrete characters, such as number of leaves on a tree, can be treated as a continuous trait with little error. However, what about a discrete trait with a rather compact distribution? A common example would be number of offspring, such as the **clutch size** for a bird, which may range from (say) 0 to 10 eggs in our observed sample with a mean of (say) four. This discreteness is of special concern when the trait has a significant probability mass at a particular value (especially zero), as often happens with offspring number.

A natural way to model such traits is to use the Poisson distribution, where the probability of observing a character state of k is given by

$$\Pr(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{10.36}$$

where $\lambda = E(Y)$ is the expected value of the trait. Motivated by our above treatment of liability, one might imagine that on some appropriate scale the mean value λ is akin to the liability of an individual (we will define this a bit more precisely below). In particular, from the theory of generalized linear models (Lindsey1997), the typical assumption is that

the log of λ is a linear function of underlying well-behaved (normally-distributed) random variables, with

$$\ln(\lambda) = \mu + g + e \tag{10.37}$$

Here, g is the genotypic value and e the environmental value, both of which are assumed to be normal with mean zero and variances σ_g^2 and σ_e^2. Hence, the average trait value is

$$
\begin{aligned}
E(Y) &= E(\exp(\mu + g + e)) \\
&= E[\exp(\mu) \cdot \exp(g) \cdot \exp(e)] \\
&= \exp(\mu) \cdot E[\exp(g)] \cdot E[\exp(e)]
\end{aligned} \tag{10.38}
$$

where the last step follows since (by construction) g and e are uncorrelated, while μ is a constant. To compute these expectations, recall that the expression $E(e^{tx})$ is the **moment-generating function** of the random variable x. For a normal (Johnson and Kotz 1970a),

$$E(e^{tx}) = \exp\left(\mu t + \frac{\sigma^2}{2}t\right) \tag{10.39a}$$

Hence, for a normal random variable x with mean zero and variance σ^2, setting $t = 1$ we have

$$E\left[\exp(x)\right] = \exp\left(\frac{\sigma^2}{2}\right) \tag{10.39b}$$

Substituting into Equation 10.38 shows that the expected mean trait value is a function of both the mean μ and variance σ_z^2 of the underlying liability value,

$$E(Y) = \exp(\mu) \cdot \exp\left(\frac{\sigma_g^2 + \sigma_e^2}{2}\right) = \exp(\mu) \cdot \exp(\sigma_z^2/2) \tag{10.40a}$$

We can now define liability a little more carefully. One might initially expect that if g is the genotypic value for liability, then its mean phenotype would simply be $\exp(\mu+g)$. However, Equation 10.40a shows that

$$E(Y \mid g) = \exp(\mu + g) \cdot \exp(\sigma_e^2/2) \tag{10.40b}$$

which reflects how variation about the expected value maps into phenotypic variation.

Following a single generate of selection, the distribution of liability values has the same (approximate) variance, but now the mean is shifted to $\mu + h^2 S$ (where S is the selection differential *on the liability scale*). Hence, response on the phenotypic scale becomes

$$
\begin{aligned}
R &= E(Y_{t+1}) - E(Y_t) \\
&= \left(\exp(\mu + h^2 S) - \exp(\mu)\right) \cdot \exp(\sigma_z^2/2) \\
&= \left(\exp(h^2 S) - 1\right) \cdot \exp(\mu) \cdot \exp(\sigma_z^2/2) \\
&= \left(\exp(h^2 S) - 1\right) \cdot E(Y_t)
\end{aligned} \tag{10.41}
$$

Notice, as was the case for selection on a binary trait, that the response is not symmetric. An S of $+\delta$ does not necessarily give the same increment of response as an S of $-\delta$.

Computing the selection differential on the log-liability scale based on the distribution of phenotypes chosen is complex. Foulley (1993) and Korsgaard et al. (2002) obtain expressions when selection is on the estimated breeding values of individuals, and even here the solution is a complex conditional expectation.

SUMMARY: LIMITATIONS OF THE BREEDERS' EQUATION

As many of the previous sections have illustrated, there are a number of situations that can compromise the breeders' equation, even if we are interested only in the single-generation response to selection and can assume a linear parent-offspring regression. One particularly important (and usually unstated) assumption is that we start from an *unselected* base population. If the base population itself has been under selection, decay of transient response components from previous selection compromises the predicted single-generation response (Chapter 11). Another troublesome assumption is that our model has accounted for all the selection on the character of interest. This is especially tricky as selection on any character correlated with the one of interest can introduce significant bias. The problem of selection of multiple characters is discussed in Chapters 30–36, but often there is no easy solution, or even any indication of a problem before the experiment begins. Thus, even in the best of situations (linearity and no selection-induced changes in allele and gamete frequencies), there are pitfalls in predicting even a single generation of response from the slope of the parent-offspring regression. Things get worse if the parent-offspring regression is nonlinear, where the single-generation change in the mean can depend on higher order moments of the genotypic distribution, and hence is not predictable from simple variance components (e.g., Equation 17.35b).

Table 10.1 reviews some of the various factors that can compromise the breeders' equation, and the chapters in which these complications are examined in detail. Provided one can assume linearity of the regressions of relatives, many of these complications can be accounted for. The importance of linearity is that if the regression of an individual on all its direct relatives selected in previous generations (back to the original unselected base population) remains linear, response is entirely determined by the covariances between a current individual and these previous relatives (Chapter 11).

Even if we have corrected for all of the potential complications listed in Table 10.1, the breeders' equation is expected to be an increasingly poor predictor as selection proceeds. Even a single generation of selection can significantly change the underlying variance components, which in turn changes the regression coefficients. Further, selection can introduce nonlinearities into an initially linear regression by transforming the starting distribution away from normality (Chapter 24). In the absence of major genes, allele frequency changes over the first few generations of selection are expected to be rather small, but *genotype* frequencies can change dramatically due to selection generating gametic-phase disequilibrium (Chapters 13, 24). Directional selection generates negative disequilibrium, decreasing heritability and hence reducing response. This reduction can be significant if heritabilty is high. Likewise, selection on the variance itself (through disruptive or stabilizing selection) also creates disequilibrium which changes the genetic variance. Chapter 13 examines such short-term changes in disequilibrium. As selection continues over several generations, allele frequencies themselves start to change, even if all loci have very small effects (Chapter 25). Drift and mutation also become increasingly important and these complications are examined in Chapters 24–27.

Table 10.1.   Summary of various factors that complicate prediction of short-term selection response in the phenotypic mean, even assuming all regressions are linear and that we are considering a single generation of selection from an unselected base population.

G × E interactions (Chapters 13, 24)	Possibility of nonlinear parent-offspring regressions. Possibility of reversed response.
Age-structure	Several generations are required to propagate

(Chapter 23)	genetic change uniformly through the population.
Epistasis (Chapter 11)	Component of response due to epistasis is transient. Parent-offspring covariance overestimates permanent response.
Correlated environmental effects (Chapter 11)	Contribution from parent-offspring correlation decays away after selection relaxed.
Maternal effects (Chapter 11 3)	Complicated lags in response — mean changes unpredictably after selection is relaxed. Possibility of reversed response.
Inbreeding (Chapter 19)	Response depends on additional variance components that are difficult to estimate (σ^2_{DI}, σ_{ADI}, etc). Response has permanent and transient components.
Drift (Chapters 14, 15)	Generates variance in the short-term response.
Gametic-phase disequilibrium (Chapter 13)	Changes additive genetic variance. Directional selection generates negative gametic-phase disequilibrium, reducing h^2 and slowing response.
Assortative Mating (Chapter 13)	Generates gametic-phase disequilibrium which either enhances (positive correlation between mates) or retards (negative correlation between mates) response.
Selection on Correlated Characters (Chapters 30, 31)	Response completely unpredicatable unless selection on correlated characters accounted for.
Environmental Correlations (Chapter 16)	Enviromental factors influence both the trait and fitness confounding the true amount of genetic change.
Environmental Change (Chapters 15, 16)	A significant change in the environment can obscure the true amount of genetic change.

Literature Cited

Arnold, S. J., and M. J. Wade. 1984a. On the measurment of natural and sexual selection: theory. *Evolution* 38: 709–719. [10]

Arnold, S. J., and M. J. Wade. 1984b. On the measurment of natural and sexual selection: applications. *Evolution* 38: 720–734. [10]

Bernardo, R. 2002. *Breeding for Quantitative Traits in Plants*, Stemma Press. [10]

Bulmer, M. G. 1980. *The mathematical theory of quantitative genetics.* Oxford Univ. Press, NY. [10]

Burrows, P. M. 1972. Expected selection differentials for directional selection. *Biomet.* 28: 1091–1100. [10]

Burrows, P. M. 1975. Variance of selection differentials in normal samples. *Biometrics* 31: 125–133. [10]

Cameron, N. D. 1997. *Selection Indices and Prediction of Genetic. Merit in Animal Breeding*. CAB international, Wallingford. [10]

Coyne, J. A., and E. Beecham. 1987. Heritability of two morphological characters within and among natural populations of *Drosophila melanogaster. Genetics* 117: 727–737. [10]

David, F. N. 1981. *Order statistics*, 2nd Ed. Wiley, New York. [10]

Fisher, R. A. 1930. *The genetical theory of natural selection*. Clarendon, Oxford. [10]

Fisher, R. A. 1958. *The genetical theory of natural selection.* Dover Publ., NY. [10]

Foulley, J. L. 1992. Prediction of selection response from threshold dichotomous traits. *Genetics* 132: 1187–1194. [10]

Foulley, J. L. 1993. Prediction of selection response for Poisson distributed traits. *Genet. Sel. Evol* 25: 297–303. [10]

Foulley, J. L., D. Gianola, and R. Thompson. 1983. Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty, birth weight, and pelvic opening. *Genet. Sel. Evol* 15: 401–424. [10]

Gallais, A. 2003. *Quantitative Genetics and Breeding Methods in Autopolyploid Plants.* INRA, Paris. [10]

Gardner, C. O. 1961. An evaluation of effects of mass selection and seed irradition with thermal neutrons on yields of corn. *Crop Sci.* 1: 241–245. [10]

Gmelfarb, A., and J. H. WIllis. 1994. Linearity versus nonlinearity of offspring-parent regression: An experimental study in *Drosophila melanogaster*. *Genetics* 138: 343–352. [10]

Haldane, J. B. S. 1954. The measurement of natural selection. *Proc. IX Internal. Cong. Genet.* 1: 480–487. [10]

Hallauer, A. R. and J. B. Miranda. 1981. *Quantitative genetics in maize breeding*. Iowa State University Press, Ames. [10]

Harter, H. L. 1961. Expected values of normal order statistics. *Biometrika* 48: 151–166. [10]

Harter, H. L. 1970a. *Order statistics and their use in testing and estimation. Volume 1: Tests based on range and studentized range of samples from a normal population.* U. S. Government Printing Office, Washington, D. C. [10]

Harter, H. L. 1970b. *Order statistics and their use in testing and estimation. Volume 2: Estimates based on order statistics of samples from various populations.* U. S. Government Printing Office, Washington, D. C. [10]

Hazel, W. N. R. Smock, and M. D. Johnson. 1990. A polygenic model for the evolution and maintenance of conditional strategies. *Proceed. Royal Society London B* 242: 181–187. [10]

Hill, W. G. 1976. Order statistics of correlated variables and implications in genetic selection programmes. *Biometrics* 32: 889–902. [10]

Hill, W. G. 1977c. Order statistics of correlated variables and implications in genetic selection programmes. II. Response to selection *Biometrics* 33: 703–712. [10]

Johnson, N. L., and S. Kotz. 1970a. *Continuous univariate distributions – 1.* John Wiley & Sons, NY. [10]

Johnson, N. L., and S. Kotz. 1970b. *Continuous univariate distributions – 2.* John Wiley & Sons, NY. [10]

Kendall, M., and A. Stuart. 1977. *The advanced theory of statistics. Vol. 1. Distribution theory.* 4th Ed. Macmillan, NY. [10]

Kinghorn, B. P., Van der Werf, J. H .J. and Ryan, M. (Eds.) 2000. "Animal Breeding: Use of New Technologies." The Post Graduate Foundation in Veterinarian Science of the University of Sydney. [10]

Korsgaard, I. R., A. H. Andersen, and J. Jensen. 2002. Prediction error variance and expected response to selection, when selection is based on the best predictor – for Gaussian and threshold characters, traits following a Poisson mixed model and survival traits. *Genet. Sel. Evol* 34: 307–333. [10]

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. *Evolution* 30: 314–334. [10]

Lande, R. 1978. Evolutionary mechanisms of limb loss in tetrapods. *Evolution* 32: 73–XXX. [10]

Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters. *Evolution* 37: 1210–1226. [10]

Lindgren, D. and J.-E. Nilsson 1985. Calculations concerning selection intensity. Report 5, Swedish University of Agricultural Sciences, Ulmea. [10]

Lindsey, J. K. 1997. *Applying generalized linear models.* Springer, New York. [10]

Lush, J. L. 1937. *Animal breeding plans.* Iowa State University Press, Ames. [10]

Matsumura, M. 1996. Genetic analysis of a threshold trait: density-dependent wing dimorphism in *Sogatella fucifera* (Horvath) (Hemiptera: Delphacidae), the whitebacked planthopper. *Heredity* 76: 229 – 237. [10]

Mayo, O. 1987. *The theory of plant breeding, 2nd Edition.* Clarendon, Oxford. [10]

Montaldo, H. H. 1997. Optimization of selection response using artifical insemination and new reproductive technologies in diary cattle. Thesis, Department of Animal Science, University of Nebraska. [10]

Namkoong, G. 1979. *Introduction to quantitative genetics in forestry.* U. S. Forest Serivce Technical Bulletin 1588, Washington, D. C. [10]

Nordskog, A. W., and A. J. Wyatt. 1952. Genetic improvement as related to size of breeding operations. *Poulty Sci* 31: 1062–1066. [10]

Ollivier, L. 1988. Current principles and future prospects in selection of farm animals. *In* B. S. Weir, E. J. Eisen, M. M. Goodman, and G. Namkoong (eds.), *Proceedings of the second international conference on quantitative genetics,* pp. 438–450. Sinauer Assocs., Sunderland, MA. [10]

Pirchner, F. 1983. *Population genetics in animal breeding*, 2nd Edition. Plenum, New York. [10]

Price, G. R. 1970. Selection and covariance. *Nature* 227: 520–521. [10]

Price, G. R. 1972. Extension of covariance selection mathematics. *Ann. Hum. Genet.* 35: 485–490. [10]

Rawlings, J. O. 1976. Order statistics for a special class of unequally correlated multinormal variates. *Biometrics* 32: 875–887. [10]

Rendel, J. M. and A. Robertson. 1950. Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. *J. Genetics* 50: 1–8. [10]

Robertson, A. 1966. A mathematical model of the culling process in dairy cattle. *Anim. Prod.* 8: 95–108. [10]

Roff, D. A. 1996. The evolution of threshold traits in animals. *Quarterly Review of Biology* 71: 3–35. [10]

Sarhan, A. E. and B. G. Greenberg. 1962. *Contributions to order statistics*. Wiley, New York. [10]

Saxton, A. M. 1988. Further approximations for selection intensity. *Theor. Appl. Genet* 76: 465–466. [10]

Simm, G. 1998. *Genetic Improvement of Cattle and Sheep*. Farming press, Suffolk. [10]

Simmonds, N. W. 1977. Approximations for i, intensity of selection. *Heredity* 38: 413–414. [10]

Smith, C. 1969. Optimum selection procedures in animal breeding. *Anim. Prod.* 11: 433–442. [10]

Stoskopf, N. C., D. T. Tomes, and B. R. Christie. 1993. *Plant Breeding: Theory and Practice*, Westview Press. [10]

Turner, H. N., and S. S. Y. Young. 1969. *Quantitative Genetics in Sheep Breeding*. Cornell, Ithaca. [10]

Weller, J. I. 2001. *Economic Aspects of Animal Breeding*, Springer. [10]

Wesselingh, R. A., and T. J. De Jong. 1995. Bidirectional selection on threshold size for flowering in *Cynoglossum officinale* (hound's-tough). *Heredity* 74: 415–424. [10]

Wricke, G., and W. E. Weber. 1986. *Quantitative genetics and selection in plant breeding*. Walter de Gruyter and Co., New York. [10]