14

Short-term Changes in the Mean: 2. Truncation and Threshold Selection

Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. — Tukey (1962)

Version 29 May 2013.

This brief chapter first considers the theory of truncation selection on the mean, which is of general interest, and then examines a number of more specialized topics that may be skipped by the casual reader. Truncation selection (Figure 14.1) occurs when all individuals on one side of a threshold are chosen, and is by far the commonest form of artificial selection in breeding and laboratory experiments. One key result is that for a normally-distributed trait, the selection intensity $\overline{\imath}$ is fully determined by the fraction p saved (Equation 14.3a), provided that the choosen number of adults is large. This allows a breeder or experimentalist to predict the expected response given their choice of p.

The remaining topics are loosely organized around the theme of selection intensity and threshold selection. First, when a small number of adults are chosen to form the next generation, Equation 14.3a overestimates the expected $\overline{\imath}$, and we discuss how to correct for this small sample effect. This correction is important when only a few individuals form the next generation, but is otherwise relatively minor. The rest of the chapter considers the response in discrete traits. We start with a binary (present/absence) trait, and show how an underlying liability model can be used to predict response. We also examine binary trait response in a logistic regression framework (estimating the probability of showing the trait given some underlying liability scores) and the evolution of both the mean value on the liability scale and the threshold value. We conclude with a few brief comments on response when a trait is better modeled as Poisson, rather than normally, distributed.

TRUNCATION SELECTION

In addition to being the commonest form of artificial selection, truncation selection is also the most efficient, giving the largest selection intensity of any scheme culling the same fraction of individuals from a population (Kimura and Crow 1978, Crow and Kimura 1979). 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 discussion of this topic, we first assume a large number of individuals are saved, before turning to complications introduced by finite sample size.

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,

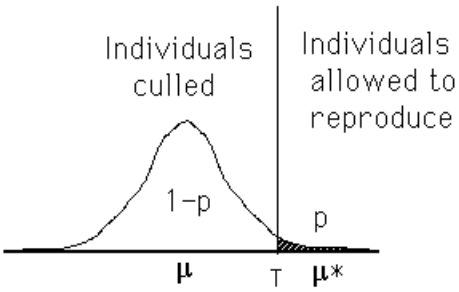


Figure 14.1. Under truncation selection, the uppermost (or lowermost) fraction p of a population is selected to reproduce. Alternatively, one could set a threshold level T in advance. To predict response given either p or T, we need to know the mean of the selected tail (μ^*), from which we can compute either $S = \mu^* - \mu$ or $\overline{\imath}$ and then apply the breeder's equation.

and we use this assumption throughout (unless stated otherwise). 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

$$S = \varphi\left(\frac{T-\mu}{\sigma}\right)\frac{\sigma}{p} \tag{14.1}$$

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

Usually the fraction saved p (rather than T) is preset by the investigator. Given p, to apply Equation 14.1, we must first find the threshold value T_p satisfying $\Pr(z \geq T_p) = p$. Notice that T in Equation 14.1 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), as satisfying

$$\Pr(\, U \leq x_{[p]}\,) = p \tag{14.2a}$$

Hence

$$\Pr(\, U > x_{[1-p]}\,) = p \tag{14.2b}$$

It immediately follows that $x_{[1-p]} = (T_p - \mu)/\sigma$, and Equation 14.1 gives the expected selection intensity as

$$\overline{\imath} = \frac{S}{\sigma} = \frac{\varphi(x_{[1-p]})}{p} \tag{14.3a}$$

Note that $\overline{\imath}$ is ***entirely a function of p***. This can be approximated by

$$\overline{\imath} \simeq 0.8 + 0.41\,\ln\!\left(\frac{1}{p} - 1\right), \tag{14.3b}$$

a result due to Smith (1969). 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, as did Saxton (1988). Montaldo (1997) gives an approximation for the standardized truncation value $z = (T - \mu)/\sigma$ in terms of $\bar{\imath}$.

Example 14.1. Consider selection on a normally distributed character in which the upper 5% of the population is saved ($p = 0.05$). Here $x_{[1-0.05]} = x_{[0.95]}$ is obtained in **R** by **`qnorm(0.95)`**, which returns 1.645, as $\Pr[U > 1.645] = 0.05$. Hence,

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

In **R**, this is obtained by **`dnorm(qnorm(0.95))/0.05`**. Smith's approximation gives the selection intensity as

$$\bar{\imath} \simeq 0.8 + 0.41 \ln\left(\frac{1}{0.05} - 1\right) \simeq 2.01$$

which is quite reasonable.

CORRECTING THE SELECTION INTENSITY FOR FINITE SAMPLE SIZE

If the number of individuals saved is small, Equation 14.1 *overestimates* the selection differential because of sampling effects (Nordskog and Wyatt 1952, Burrows 1972). To see this, suppose 100 observations are put randomly in ten groups of size ten and the largest value selected from each group. These will be, on average, not as extreme as selecting the best ten from the entire 100, as the best observation within a random group of ten can be the 11th largest (or smaller) from the entire group. To more formally treat this, assume M adults are sampled at random from the population and the largest N of these are used to form the next 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 $z_{k,M}$ denotes 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 (Harter 1961, Kendall and Stuart 1977, David 1981, Sarhan and Greenberg 1962, Harter 1970a,b), and these can also be obtained by simulation. Figure 14.2 plots 10,000 random draws of the largest order statistic in a sample of size ten ($p = 0.1$). Note that the distribution of realized differentials is asymmetric about its mean, implying that the variance alone is not sufficient for computing confidence intervals. Figure 14.3 plots exact values for the expected selection intensity for small values of N, showing that Equation 14.3a overestimates the intensity, although the difference is small unless N is small.

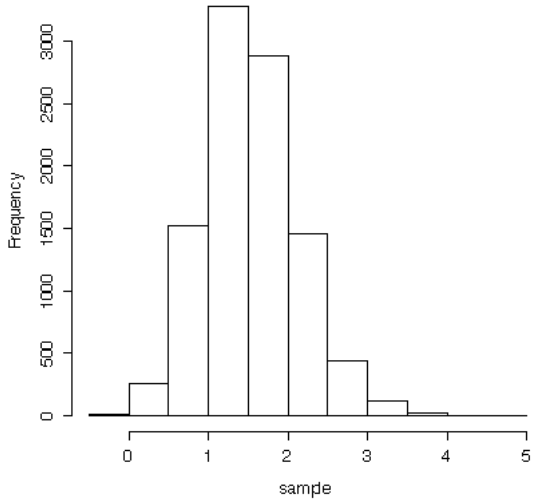


Figure 14.2. The distribution of 10,000 random draws of $\overline{z}_{(10,1)}$, the largest order statistic in a sample of ten. The mean value is 1.54, as opposed to the expected value of $\overline{z} = 1.75$ for $p = 0.1$ in an infinite population (Equation 14.3a). Notice that there is a consideriable spread about the mean, and that the distribution is not symmetric but rather is skewed towards higher values.

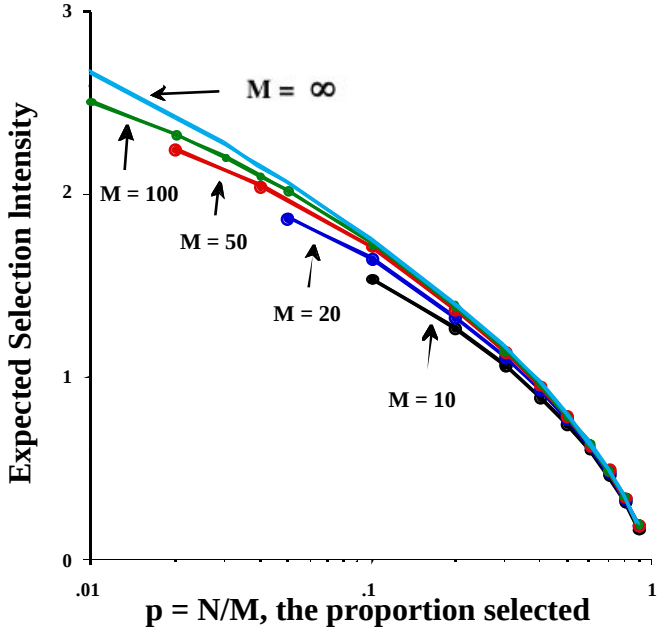


Figure 14.3. The expected selection intensity $E(\overline{z})$ under truncation selection with normally-distributed phenotypes, as a function of the total number of individuals measured M and the fraction of these saved $p = N/M$. The curve $M = \infty$ is given by Equation 14.3a, while the curves for $M = 10, 20, 50$, and 100 were obtained from the average of the expected values of the $N = pM$ largest unit normal order statistics (Harter 1961). Note that Equation 14.3a 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

$$E(\, \overline{\imath}_{(M,N)} \,) \simeq \mu_{y_p} - \frac{(M-N)\,p}{2N(M+1)\,\phi(\, y_p \,)} \tag{14.4a}$$

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 14.4a is positive, if M is finite the expected truncated mean overestimates the expected standardized selection differential. Expressions of the variance of $\overline{\imath}$ in finite populations are given by Burrows (1975).

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

$$E(\, \overline{\imath}_{(M,N)} \,) \simeq \overline{\imath} - \left[\frac{M-N}{2N(M+1)}\right] \frac{1}{\overline{\imath}} = \overline{\imath} - \left[\frac{1-p}{2p(M+1)}\right] \frac{1}{\overline{\imath}} \tag{14.4b}$$

where $\overline{\imath}$ is given by Equation 14.3a. Lindgren and Nilsson (1985) found this approximation to be quite accurate for $N \geq 5$. Bulmer (1980) suggests an alternative approximation under normally, using Equation 14.3a with p replaced by

$$\widetilde{p} = \frac{N + 1/2}{M + N/(2M)} \tag{14.4c}$$

Example 14.2. Consider the expected selection intensity on males when the upper 5% are used to form the next generation and phenotypes are normally distributed. If the number sampled is large, $\overline{\imath} \simeq 2.06$ (Example 14.1). Suppose, however, that only 20 males are sampled, with only the largest allowed to reproduce in order to give $p = 0.05$. The expected value for this individual is the expected value of the largest order statistic for a sample of size 20. For the unit normal, this is $\simeq 1.87$ (Harter 1961) and hence $E(\, \overline{\imath}_{(20,1)} \,) \simeq 1.87$. 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 $E(\, \overline{\imath}_{(20,1)} \,)$ perform? Burrows' approximation (Equation 14.4b) gives

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

Bulmer's approximation (Equation 14.4c) uses

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

which gives $x_{[1-0.075]} \simeq 1.44$. Since $\varphi(1.44) = 0.1415$, $E(\, \overline{\imath}_{(20,1)} \,) \simeq 0.1415/0.075 \simeq 1.89$.

A final correction for finite population size was noted by Rawlings (1976) and Hill (1976, 1977c). If families are sampled, such that n individuals are chosen per family, then the selection intensity is further reduced because there are positive correlations between family members. This effectively lowers the sample size below n — in an extreme case where all individuals are clones with little environmental variance, all have essentially the same value and hence $n \sim 1$. If a total of M individuals are sampled, with n individuals per family, then Burrows' correction (Equation 14.4b) is modified to become

$$\overline{i} - \left[\frac{1-p}{2p(M+1)(1-\tau+\tau/n)}\right]\frac{1}{\overline{i}} \tag{14.5}$$

where τ is the intra-class correlation of family members.

RESPONSE WITH DISCRETE TRAITS: BINARY CHARACTERS

The Threshold/Liability Model

An interesting complication is selection response in **binary traits**, which are characterized simply by presence/absence (such as normal/diseased). The basic trait model to this point assumed a continuous character, which initially seems at odds with a binary trait. However, as discussed in LW Chapters 11 and 25, discrete characters can often be modeled by mapping an underlying (and unobserved) continuous character, the **liability** z, to the observed discrete character states, $y = 0$ or $y = 1$ (Figure 14.4). The assumption is that the breeder's equation holds on the liability scale, and our goal is to predict how changes on this scale map to changes in the frequency of a binary trait. The simplest assumption is a **threshold model**, wherein the character is either present if liability exceeds the threshold value T ($z \geq T$), else it is absent ($z < T$). Roff (1996) reviews a number of examples of such threshold-determined morphological traits in animals. Our analysis is restricted to a single threshold, but extension to multiple thresholds is straightfoward (Lande 1978, Korsgaard et al. 2002).

To predict response, let μ_t be the mean liability and q_t the frequency of individuals displaying the character in generation t, i.e., $q_t = \Pr(y_t = 1)$. If liability is well enough behaved to satisfy the assumptions of the breeder's equation, then $\mu_{t+1} = \mu_t + h^2 S_t$. The problem is to (i) estimate the mean liability μ_t from the observed frequency q_t of the trait, (ii) estimate S on the liability scale given the change in the frequency of the trait following selection, and (iii) translate μ_{t+1} into q_{t+1}. We assume liability is normally distributed on some appropriate scale, in which case we can also 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 14.2b

$$\mu_t = -x_{[1-q_t]} \tag{14.6}$$

where $x_{[p]}$ is the probit transformation of p (Equation 14.2a). For example, if 5% of the population displays the trait, since $\Pr(U \leq 1.65) = 0.95$, $x_{[0.95]} = 1.65$ and the mean on the underlying liability scale is $\mu = -x_{[0.95]} = -1.645$.

The response to selection, as measured by the new frequency q_{t+1} of the trait in the next generation, is

$$\begin{aligned} q_{t+1} &= \Pr(U \geq -\mu_{t+1}) \\ &= \Pr(U \geq -\mu_t - h^2\, S_t\,) \\ &= \Pr(U \geq x_{[1-q_t]} - h^2 S_t\,) \end{aligned} \tag{14.7}$$

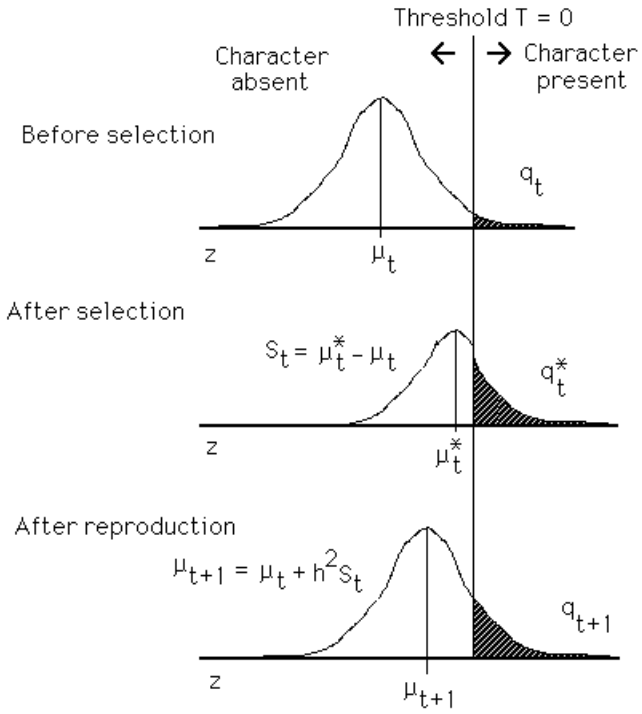


Figure 14.4. Selection response for a binary trait which is presented when the underlying liability z exceeds some threshold value T. We assume that an appropriate scale can be found such that $z \sim N(\mu_t, 1)$, where μ_t is the current mean and $T = 0$. Since z is normal, the probit transform (Equation 14.2b) estimates μ_t from the frequency q_t of individuals displaying the character. We assume that the breeder's equation holds on the liability scale, so that $\mu_{t+1} = \mu_t + S_t h^2$, where $S_t = \mu_t^* - \mu_t$. Using properties of the unit normal allows us to translate the mean libability following selection μ_{t+1} into the new frequency q_{t+1} of the trait (Equation 14.7).

It remains to obtain $S_t = \mu_t^* - \mu_t$, where μ^* is the mean liability value in the selected parents in generation t. While the selected population may consist entirely of adults displaying the trait, 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,

$$\mu_t^* = (1 - p_t)\, E(z|z < 0; \mu_t) + p_t\, E(z|z \geq 0; \mu_t) \tag{14.8a}$$

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

$$E(z|z \geq 0; \mu_t) = \mu_t + \frac{\varphi(\mu_t)}{q_t}, \quad \text{and} \quad E(z|z < 0; \mu_t) = \mu_t - \frac{\varphi(\mu_t)}{1 - q_t}. \tag{14.8b}$$

Substituting into Equation 14.8a gives

$$S_t = \mu_t^* - \mu_t = \frac{\varphi(\mu_t)}{q_t}\,\frac{p_t - q_t}{1 - q_t} = \frac{\varphi(-x_{[1-q_t]})}{q_t}\,\frac{p_t - q_t}{1 - q_t} \tag{14.9}$$

As expected, if $p_t > q_t$, then $S_t > 0$. Maximal selection occurs if only individuals displaying the trait are saved ($p_t = 1$), in which case Equation 14.9 reduces to $S_t = \varphi(\mu_t)/q_t$.

Why we did not simply estimate μ_t^* using $x_{[1-q_t^*]}$, 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 14.8a), 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 14.3. Consider a threshold trait whose liability has heritability $h^2 = 0.25$ (Example 14.4 and especially LW Chapter 25 discuss how h^2 can be estimated). What is the expected response to selection if the initial frequency of individuals displaying the character is 5% and selection is practiced by choosing 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). Only individuals displaying the trait are allowed to reproduce, giving (Equation 14.9) the resulting selection differential on the liability scale as

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

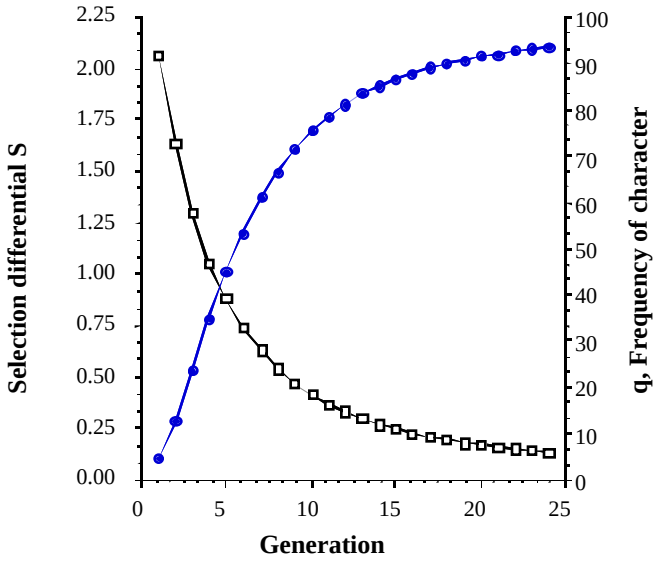
Applying the breeder's equation gives the new mean value of liability as

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

Equation 14.7 translates this new mean into the fraction of the population now above the threshold,

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

Thus, after one generation of selection, the character frequency is expected to increase from 5% 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. Only six generations are required to increase the frequency of the trait to 50% ($\mu = 0$). Note that even though all selected parents show the trait, the selection differential rapidly declines in a nonlinear fashion.



Example 14.4. The effectiveness of selection on wing morphs in females of the white-backed planthopper *Sogatella furcifera* was examined by Matsumura (1996). While this hemipterian is a serious rice pest in Japan, it is unable to overwinter. Rather, each year it migrates from southern China to recolonize Japan. Females exhibit two wing morphs, males just one. *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 a density of one nymph, 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 14.6. The response (on the liability scale) 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 from q and p using Equation 14.9. 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 $\sum_i S_i = 3.91$. One key summary statistic for any selection experiment is the *realized heritability*, the ratio of response to selection differential. As detailed in Chapter 18, there are several ways to compute this for a multi-generation selection experiment. One simple estimate is the total response/total differential ratio,

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

giving an estimated heritability of the underlying liability for macroptery of 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 between phenotypes and their underlying liability is highly non-linear. Even though the parent-offspring regression

on the liability scale is assumed to be linear (and hence liability response is symmetric), the parent-offspring regression on the *phenotypic* level is not linear, resulting in an asymmetric response.

Direct Selection on the Threshold T

It is biologically quite reasonable to imagine that there is variation in T itself (Hazel et al. 1990). Suppose the trait of interest appears when the size of an organism exceeds some critical value, which itself varies over individuals, with certain genotypes and/or environments lowering the value of T, allowing individuals with a lower liability score to display the trait. Decomposing both the liability and threshold in terms of genetic and environmental factors, $z = g_z + e_z$ and $T = g_T + e_T$. The trait appears when $z \geq T$, or

$$g_z + e_z - (g_T + e_T) = (g_z - g_T) + (e_z - e_T) = g + e \geq 0$$

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 and threshold values, and the analysis proceeds as above. If interest is simply on presence/absence of the binary trait, it does not matter as to whether the liability or threshold, or both, show variation. However, as Example 14.4 (below) 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.

The Logistic Regression Model for Binary Traits

The threshold approach offers one model for mapping an underlying continuous liability z into a discrete character space y (which is either zero or one, corresponding to trait absence/presence). This is a deterministic model, with all individuals with $z \geq T$ displaying the trait ($y = 1$), while all those with $z < T$ do not ($y = 0$). A potentially more realistic model is that trait presence is stochastic, with the underlying liability z mapping into a *probability* of displaying the trait, e.g., $p(z) = \text{Prob}(y = 1 \mid z)$. Under the threshold model, this probability is one for $z \geq T$, and zero otherwise. From a biological standpoint, one imagines that $p(z)$ is a monotonically increasing function of z, approaching zero for low values and one for high values. One reasonable candidate that satisfies these requirements is the **logistic function**,

$$\ell(z) = \frac{1}{1 + \exp(-z)} \tag{14.10a}$$

with $\ell(z) \simeq 0$ for $z \ll -1$, $\simeq 1$ for $z \gg 1$, and $\ell(0) = 1/2$. A more general version is

$$\ell[\alpha\,(z - m)] = \frac{1}{1 + \exp[-\alpha\,(z - m)]} \tag{14.10b}$$

which has value 0.5 at $z = m$ and a scaling factor α that sets the abruptness of the transition from low to high probability (Figure 14.5). The larger α, the more abrupt the transition, approaching the threshold model for sufficiently large α. Equation 14.10b is often called a **logistic regression**.

The logistic regression and threshold models are essentially identical. To see this, recall that the threshold model very easily extends to the case where T varies over individuals. In such cases, if the liability value of an individual is z, the trait will only be displayed if $T \leq z$. Now consider the logistic regression model where $p(z)$ denotes the probability that an individual with liability value z displays the trait. One source of this stochasticity could simply be population variation in T, so that $p(z)$ can be viewed as the *cumulative distribution*

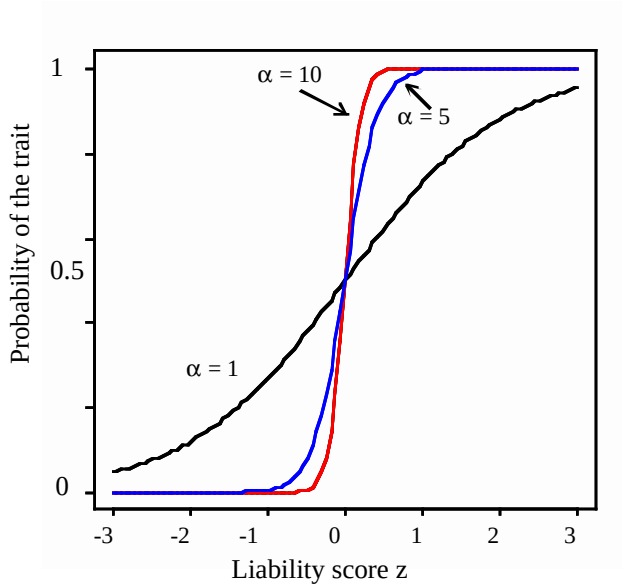


Figure 14.5. A more realistic model of threshold traits is that the liability z (horizontal axis) determines the probability $p(z)$ of displaying the trait (vertical axis). One flexible model is to assume $p(z)$ follows a logistic function (Equation 14.10b) with scale parameter α, plotted here for values of $\alpha = 1$, 5, and 10. For α values in excess of five, the logistic function essentially recovers the discrete threshold model.

function or cdf (LW Chapter 2) for the threshold value T, e.g., $p(z) = \Pr(T \leq z)$. In this case, a fraction $p(z)$ of individuals with liability z that are above the threshold, and hence display the trait.

If the logistic gives the cdf of random threshold values, then the **logistic distribution** $\phi(x, \alpha, m)$ gives the actual distribution of T. From the definition of a cumulative distribution function,

$$\int_{-\infty}^{z} \phi(x, \alpha, m) dx = \frac{1}{1 + \exp[-\alpha\,(z - m)]} \tag{14.11a}$$

Taking derivatives of both sides gives

$$\phi(x, \alpha, m) = \frac{\alpha\, \exp[-\alpha\,(z - m)]}{(1 + \exp[-\alpha\,(z - m)])^2} \tag{14.11b}$$

Johnson and Kotz (1970b) give the first three moments of this distribution as

$$\mu = m, \quad \sigma^2 = \frac{1}{3}\left(\frac{\pi}{\alpha}\right)^2, \quad \mu_3 = 0 \tag{14.11c}$$

As shown in Figure 14.6, the normal and logistic distributions have very similar cumulative distribution functions. Indeed, for a unit normal random variable U,

$$\Pr(U \leq x) \simeq \frac{1}{1 + \exp(-\alpha\, x)}, \quad \text{where} \quad \alpha = \frac{\pi}{\sqrt{3}} \tag{14.12}$$

which is the logistic distribution with variance one (see Equation 14.11c).

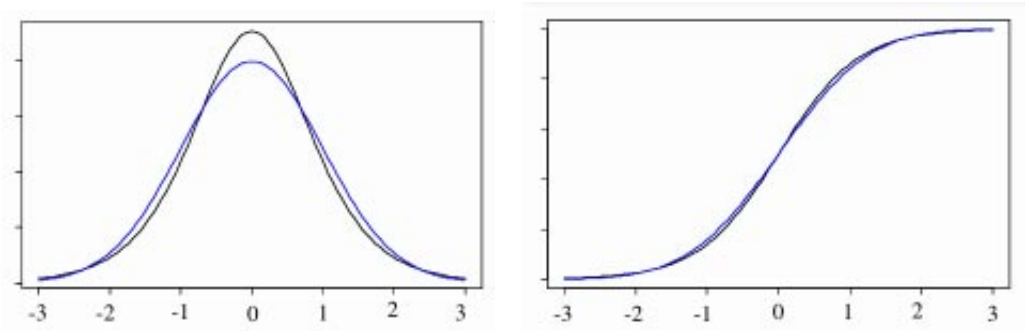


Figure 14.6. A compairson of the unit normal and unit logistic ($\mu = 0$, $\sigma^2 = 1$) distributions, with the horizontal axis the value of z. **Left**: Probability density functions: the logistic is more peaked, with positive kutosis. **Right**: The cumulative distribution functions are very similar.

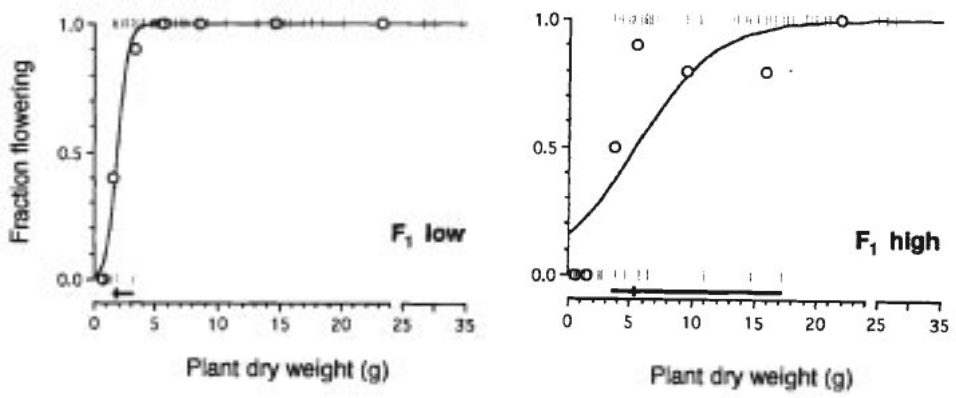
Thus, we have two approaches for mapping liability values into binary traits: the strict threshold approach (a deterministic mapping of liability into the discrete trait) or the logistic regression approach (a stochastic mapping translating a liability value into a probability of observing the trait). Given the very close connection between the threshold and logistic regression models, for most purposes the simple threshold model is a reasonable approach, even if the underlying mapping is stochastic, and as illustrated above can easily be used to predict selection response.

One setting where the logistic regression is more appropriate is in the actual analysis of the behavior of the thresehold when one either knows the liability value or has at least a strong proxy (such as size).

Example 14.5. An interesting analysis of selection on a threshold trait using logistic regressions was given by Wesselingh and de Jong (1995), who studied the connection between plant size and flowering in hound's-tongue (*Cynoglossum officinale*). This species 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 from a 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 that triggers the binary trait (vernalization sensitivity), and in particular whether this size is both variable and heritable. To examine this, they grew plants for different number of days (ranging from 31 to 86) to generate individuals 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. As plotted below, threshold values for resulting F_1 offspring from intercrosses within each set of selected parents were examined.

The data available to the authors were 0/1 (insensitive/sensitive to vernalization) values as a function of size. To estimate the distribution of threshold sizes, they performed a logistic regression on these data, using maximum likelihood (LW Appendix 4) to fit the α

and mean (m) terms of Equation 14.10b. Data for the high and low lines are plotted above along with the ML solution for the logistic regression. 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 progeny from the low and high lines parents and for a control line C contemporaneously-grown with these progeny. The ML estimates of the mean m (which corresponds to the weight yielding 50% flowering) and α for these regressions were

Line	m	α
C	3.30	0.97
Low	1.85	2.58
High	5.41	0.31

Note that the low line not only has a smaller mean size for vernalization (1.85), but also a much larger α value (2.58) and therefore a more abrupt transition between insensitivity and sensitivity. Using these estimates, Equation 14.10b yields the expected percent of vernalization sensitivity (flowering) for a given weight. For a 3 gram plant this is 0.43 in the control line, 0.32 in the high lines, but 0.95 in the low line. Basing estimates of response using the contemporaneously grown control 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 source population), while the smallest flowering high parent was 9.95 grams (corresponding to the upper 12.2% of the founding source population). From Equation 14.3a, these translate into selection intensities of 1.26 and 1.66, respectively. To obtain the selection differentials S for each line, recall that $S = \overline{\imath}\sigma_p$. 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 14.6). Hence, taking the observed 0.25 quartile (in the P lines) of 2.68, and its mean of 5.12, suggests

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

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

Line	$\overline{\imath}$	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 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.

The other setting where the logistic regression model is used is for BLUP selection. Recall that animal breeders routinely use *linear* mixed models to obtain BLUP estimates of breeding values. The standard model for normally-distributed traits is to assume that an observation for individual i can be written as

$$y_i = \mu + \sum \beta_k x_{k,i} + A_i + e_i \tag{14.13}$$

where the β_k are fixed effects (such as adjustments for age and sex), A_i their breeding value, and the residual e_i is normally distributed. In addition to y_i, further information to estimate A_i is borrowed from the y values of relatives through the relationship matrix $\mathbf{A}$, with those individuals with the largest estimated A values are chosen to form the next generation. **Generalized linear mixed models** extend Equation 14.13 to cases where (i) the expected value of y conditioned on the variables of interest is *not* a linear function and (ii) the residuals about this expected value are not necessarily normal. The basic structure of a generalized linear model is that the conditional expectation of y can be expressed as

$$E(y \,|\, z) = g(z) \tag{14.14a}$$

for some montonic function g, with

$$g^{-1}\left[E(y \,|\, z)\right] = z = \mu + \sum \beta_k x_{k,i} + A_i \tag{14.14b}$$

The inverse g^{-1} is called the **link function** as it transforms the conditional expectation into a linear model.

For binary data, a single value of y follows a **Bernoulii distribution** ($y = 0$ or 1) with success parameter $p(z) = \Pr(y = 1 \,|\, z)$. Equation 14.14a becomes $E(y\,|z) = p(z)$, so that $g(z)$ is given by $\ell(z)$, where the simple logistic (Equation 14.10a) is used as one can always work on a scale with $m = 0, \alpha = 1$. The corresponding link function (Equation 14.14b), the inverse of the logistic function, is given by the **logit function** $L(p)$,

$$L(p) = \ln\left(\frac{p}{1-p}\right) \tag{14.15a}$$

which is the log of the odds ratio (probability of the trait divided by probability that the trait is absent). If $\ell(z) = p$, then $L(p) = z$, so that a logit-transformed p value recovers the liability value,

$$L(p|z) = z = \mu + \sum \beta_k x_{k,i} + A_i \tag{14.15b}$$

Under this framework, BLUP selection for individuals with the highest breeding values for a binary trait proceeds by taking the 0/1 binary data from a set of individuals (along with other fixed, and possibly random, effects of interest) and using either maximum likelihood or bayesian approaches to estimate the breeding values (Foulley et al. 1983, Foulley 1992, Vazques et al. 2009). This approach can be extended to $k \geq 2$ thresholds in the mapping of liability into different character states (Korsgaard et al. 2002).

RESPONSE WITH DISCRETE TRAITS: POISSON-DISTRIBUTED CHARACTERS

Many discrete characters with multiple states, 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 of 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

$$\Pr(y = k) = e^{-\lambda}\,\frac{\lambda^k}{k!} \tag{14.16}$$

where $\lambda = E(y)$ is the expected value of the trait. Motivated by the above treatment of binary traits, 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, we can take

$$\lambda = \exp(z) \tag{14.17a}$$

ensuring for all z that $\lambda > 0$ and hence a proper expectation for a Poisson. In the context of generalized linear models, $g(z) = \exp(z)$ so that the link function $g^{-1}(z)$ is just $\ln(z)$, with

$$\ln(\lambda) = z = \mu + A + e \tag{14.17b}$$

This is called a **log-linear model**, as the log of the distribution parameter λ is a linear function of the variables of interest (in particular, the breeding value). On this log scale, both the breeding and environmental values are assumed to be normal with mean zero and variances σ_A^2 and σ_e^2. As with binary traits, BLUP selection occurs using this generalized linear model framework to estimate the A_i values (Foulley 1993, Korsgaard et al. 2002, Vazques et al. 2009). Other models are also possible, such as a **zero-inflated Poisson**, which has extra probablity mass at zero relative to a standard Poisson (Rodrigues-Motta et al. 2007).

Under the log-linear model, the liability z of an individual determines $\lambda = \exp(z)$ and then a realization is drawn from a Poisson(λ) to give their observed trait value. The resulting mean trait value in a population becomes

$$\begin{aligned} E(y) &= E(\lambda) = E[\exp(z)] \\ &= E[\,\exp(\mu)\cdot\exp(A)\cdot\exp(e)\,] \\ &= \exp(\mu)\cdot E[\,\exp(A)\,]\cdot E[\,\exp(e)\,] \end{aligned} \tag{14.18}$$

where the last step follows since (by construction) A 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),

$$E(e^{tx}) = \exp\left(\mu t + \frac{\sigma^2}{2}t\right) \tag{14.19a}$$

For a normal random variable x with mean zero and variance σ^2, setting $t = 1$ gives

$$E\left[\exp(x)\right] = \exp\left(\frac{\sigma^2}{2}\right) \tag{14.19b}$$

Substituting into Equation 14.18 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,

$$E(y) = \exp(\mu) \cdot \exp\left(\frac{\sigma_A^2 + \sigma_e^2}{2}\right) = \exp(\mu) \cdot \exp(\sigma_z^2/2) \tag{14.20a}$$

One might initially expect that if A is the breeding value for liability, then its mean phenotype would simply be $\exp(\mu + A)$. However, Equation 14.20a shows that

$$E(y \mid A) = \exp(\mu + A) \cdot \exp(\sigma_e^2/2), \tag{14.20b}$$

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

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

$$\begin{aligned} R &= E(y_{t+1}) - E(y_t) \\ &= \big(\exp(\mu + h^2 S) - \exp(\mu)\big) \cdot \exp(\sigma_z^2/2) \\ &= \big(\exp(h^2 S) - 1\big) \cdot \exp(\mu) \cdot \exp(\sigma_z^2/2) \\ &= \big(\exp(h^2 S) - 1\big) \cdot E(y_t) \end{aligned} \tag{14.21}$$

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

Literature Cited

Bulmer, M. G. 1980. *The mathematical theory of quantitative genetics.* Oxford Univ. Press, NY. [14]

Burrows, P. M. 1972. Expected selection differentials for directional selection. *Biomet.* 28: 1091–1100. [14]

Burrows, P. M. 1975. Variance of selection differentials in normal samples. *Biometrics* 31: 125–133. [14]

Crow, J. F., and M. Kimura. 1979. Efficiency of truncation selection. *Proc. Natl. Acad. Sci. USA* 76: 396–299. [14]

David, F. N. 1981. *Order statistics*, 2nd Ed. Wiley, New York. [14]

Foulley, J. L. 1992. Prediction of selection response from threshold dichotomous traits. *Genetics* 132: 1187–1194. [14]

Foulley, J. L. 1993. Prediction of selection response for Poisson distributed traits. *Genet. Sel. Evol* 25: 297–303. [14]

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. [14]

Harter, H. L. 1961. Expected values of normal order statistics. *Biometrika* 48: 151–166. [14]

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. [14]

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. [14]

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. [14]

Hill, W. G. 1976. Order statistics of correlated variables and implications in genetic selection programmes. *Biometrics* 32: 889–902. [14]

Hill, W. G. 1977c. Order statistics of correlated variables and implications in genetic selection programmes. II. Response to selection *Biometrics* 33: 703–712. [14]

Johnson, N. L., and S. Kotz. 1970a. *Continuous univariate distributions – 1.* John Wiley & Sons, NY. [14]

Johnson, N. L., and S. Kotz. 1970b. *Continuous univariate distributions – 2.* John Wiley & Sons, NY. [14]

Kendall, M., and A. Stuart. 1977. *The advanced theory of statistics. Vol. 1. Distribution theory.* 4th Ed. Macmillan, NY. [14]

Kimura, M., and J. F. Crow. 1978. Effect of overall phenotypic selection on genetic change at individual loci. *Prod. Natl. Acad. Sci. USA* 75: 6168–6171. [14]

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. [14]

Lande, R. 1978. Evolutionary mechanisms of limb loss in tetrapods. *Evolution* 32: 73–92. [14]

Lindgren, D. and J.-E. Nilsson 1985. Calculations concerning selection intensity. Report 5, Swedish University of Agricultural Sciences, Ulmea. [14]

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. [14]

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. [14]

Nordskog, A. W., and A. J. Wyatt. 1952. Genetic improvement as related to size of breeding operations. *Poulty Sci* 31: 1062–1066. [14]

Rawlings, J. O. 1976. Order statistics for a special class of unequally correlated multinormal variates. *Biometrics* 32: 875–887. [14]

Rodrigues-Motta, M., D. Gianola, B. Heringstad, G. J. M. Roza, and Y. M. Chang. 2007. A zero-inflated Poisson model for genetic analysis of number of mastitis cases in Norwegian Red cows. *J. Dairy Sci.* 90: 5306–5315. [14]

Roff, D. A. 1996. The evolution of threshold traits in animals. *Quarterly Review of Biology* 71: 3–35. [14]

Sarhan, A. E. and B. G. Greenberg. 1962. *Contributions to order statistics*. Wiley, New York. [14]

Saxton, A. M. 1988. Further approximations for selection intensity. *Theor. Appl. Genet* 76: 465–466. [14]

Simmonds, N. W. 1977. Approximations for *i*, intensity of selection. *Heredity* 38: 413–414. [14]

Smith, C. 1969. Optimum selection procedures in animal breeding. *Anim. Prod.* 11: 433–442. [14]

Tukey, J. W. 1962. The future of data analysis. *Annals of Math.l Stats* 33: 1-67. [14]

Vazquez, A. I., D. Gianola, D. Bates, K. A. Weigel, and B. Heringstad. 2009. Assessment of Poisson, logit, and linear models for genetic analysis of clinical mastitis in Norwegian Red cows. *J. Dairy Sci.* 92: 739–748. [14]

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. [14]