THEORY AND ANALYSIS OF THRESHOLD CHARACTERS[1]

Daniel Gianola[2]

University of Illinois, Urbana 61801

Summary

This paper deals with theory, methods and problems in the analysis of categorical data in animal breeding when a threshold model is postulated for an underlying normal distribution of phenotypes. Linear and nonlinear statistical models potentially useful for estimating parameters of the underlying and outward distributions are reviewed and extended. Methods for evaluating the genetic value of candidates for selection from single or multiple populations are discussed comparatively.
(Key Words: Threshold Characters, Categorical Data, Linear Models, Nonlinear Models, Sire Evaluation.)

Introduction

Most applications of quantitative genetics theory to animal breeding have been made with respect to characters showing a continuous phenotypic distribution. Multifactorial models based on an underlying Mendelian mechanism usually provide a reasonable fit to continuous data. Another area of animal breeding deals with simple Mendelian models with discrete phenotypic distribution, with attention centered on prevalence and incidence of traits, numbers of alleles and(or) loci, distribution of gene frequencies, map distance, penetrance, fitness and levels of consanguinity (see, e.g., Rasmusen and Lewis, 1973; Rasmusen and Christian, 1976; Haseman and Elston, 1972).

Many traits of importance in animal production such as littler size in sheep, degree of calving difficulty, conformation and type scores, survival or death, or liability to disease, present a discontinuous distribution of phenotypes. In these instances, breeding tests show features that cannot be readily explained by simple strict Mendelian inheritance. Characters of this sort, known as threshold or quasi-continuous (Grüneberg, 1952; Falconer, 1960), can be analyzed by postulating an underlying continuous distribution of phenotypes which maps into the observed distribution via a set of fixed thresholds.

The objective of this paper is to review, characterize and extend models for quasi-continuous variation of possible interest in animal breeding, to outline procedures suitable for estimating parameters of these models, and to discuss methods for estimating the worth of candidates for selection.

Quasi-Continuous Variation

Historical Background. Wright (1934a) analyzed the variability of number of digits between and within strains of guinea pigs. He observed that the character did not occur in grades from which the variances could be calculated directly and that the most significant classification was the dichotomy three-toed or four-toed. Wright (1934a) hypothesized an underlying normal distribution of phenotypes, with mean μ_i in the i^{th} strain and common variance σ^2. In this underlying scale there is a fixed threshold – i.e., if the underlying variate is above the threshold, the four-toed character is expressed. If π_i is the proportion above the threshold (t) in the i^{th} substrain

$$\pi_i = \int_t^\infty (\sqrt{2\pi}\,\sigma)^{-1} e^{-(y-\mu_i)^2/2\sigma^2}\,dy$$
$$= 1 - \Phi(t_i), \qquad (1)$$

where $\Phi(t_i)$ is the cumulative distribution function of a standard normal variate evaluated at

$$t_i = \frac{t - \mu_i}{\sigma}.$$

[1] Invitational paper presented at a Symp. on "Statistical Analysis of Categorical Data in Animal Research", 72nd Annu. Meet. of the ASAS, Cornell University, Ithaca, NY, July 28, 1980.

[2] Dept. of Anim. Sci.

From (1),

$$t_i = \Phi^{-1}(1 - \pi_i), \qquad (2)$$

yields the inverse probability transformation or the point in the underlying scale at which there are probability masses equal to $1 - \pi_i$ and π_i to the left and right of the threshold, respectively. Wright (1934a) calculated the statistics $m_i = -t_i\sigma$, which gives the distance of the mean of the i^{th} strain from the threshold, and

$$\sigma_m^2 = \sum_i (m_i - \bar{m})^2/n,$$

yielding (incorrectly, as pointed out by Wright, 1943) the variance of such means. The proportion of the total variability due to differences among strains was computed as

$$\sigma_m^2/(\sigma_m^2 + \sigma^2),$$

with the within strain standard deviation (σ) taken as unit of measurement. Wright assumed that all strains had the same variance in the underlying scale and that the strains were a random sample from a population of strains. Wright (1934b) extended the model to study the trichotomy three-toed, four-toed (imperfect) and four-toed (good) and attempted to compare means and variabilities of a number of strains. He also used the inverse probability transformation to study flower color in Linanthus parryae (Wright, 1943).

If the distance between thresholds is the same from population to population, the variances of the populations can be compared by taking the difference between thresholds as the unit of measurement; this cannot be done with only two categories of response. The reciprocal of the distance between the thresholds gives the standard deviation on a scale on which the thresholds are separated by a unit distance; the thresholds on this scale can be obtained by multiplying the previous threshold values by the "new" standard deviation. The inverse probability transformation was discovered independently by Bliss (1934a,b) who termed it "probit". Probits are widely used in bioassay (e.g., Finney, 1947).

Generalized Multiple Threshold Model. Consider an animal or experimental unit that is subjected to a set of conditions defining a population in a statistical sense. The expression of a character to a set of factorial combinations is a response in one of m mutally exclusive and exhaustive categories that follow an order in some sense. For example, a ewe of a certain age and breed may produce 0, 1, 2, 3 or more lambs in a given lambing season. The response may be ambiguous and measurement error occurs, e.g., whether a cow has an "extremely difficult" calving or "considerable force is required" (Pollak and Freeman, 1976). Assume an underlying continuous distribution of phenotypes and that this continuous random variable (y) has a joint distribution with the outward discrete variable. In the continuous scale there are $m - 1$ fixed thresholds represented by the vector

$$t' = [t_1, t_2, \ldots, t_{m-1}]$$

corresponding to the m discontinuities in the outward scale. With $t_o = -\infty$ and $t_m = \infty$, if

$$t_{j-1} < y < t_j,$$

for $j = 1, \ldots, m$, the animal is scored as responding in the j^{th} category. The phenotype of the i^{th} animal in the underlying scale is represented by a linear combination of parameters and random variables

$$y_i = x_i'\beta + z_i'u + a_i + e_i, \qquad (3)$$

where β is a $p \times 1$ vector of parameters, x_i is a $p \times 1$ known vector of variables relating y_i to β, $u \sim (0, D)$ is a $q \times 1$ random vector, z_i is a $q \times 1$ known vector of variables relating y_i to u, a_i is the additive genetic value of the i^{th} individual and e_i is an environmental deviation. Location and dispersion assumptions are

$$E\begin{bmatrix} y_i \\ u \\ a_i \\ e_i \end{bmatrix} = \begin{bmatrix} x_i'\beta \\ 0 \\ 0 \\ 0 \end{bmatrix};$$

$$\text{Var}\begin{bmatrix} y_i \\ u \\ a_i \\ e_i \end{bmatrix} = \begin{bmatrix} \sigma^2 & c' & \sigma_a^2 & \sigma_e^2 \\ c & D & 0 & 0 \\ \sigma_a^2 & 0' & \sigma_a^2 & 0 \\ \sigma_e^2 & 0' & 0 & \sigma_e^2 \end{bmatrix}.$$

In addition, u, a_i and e_i are multinormal mu-

tually stochastically independent random variables so that they are also stochastically independent pairwise (Hogg and Craig, 1970). The model can be standardized as

$$y_i^* = (y_i - x_i'\beta)/\sigma = z_i'u/\sigma + a_i/\sigma + e_i/\sigma$$
$$= u_i^* + a_i^* + e_i^*, \quad (4)$$

With $E(y_i^*) = E(u_i^*) = E(a_i^*) = E(e_i^*) = 0$, $Var(u_i^*) = q^2$, $Var(a_i^*) = h^2$, which is heritability in the narrow sense, and $Var(e_i^*) = 1 - h^2 - q^2 = e^2$; all terms in (4) are also mutally stochastically independent. The vector of thresholds becomes

$$t^{*\prime} = [t_{i(1)}^*, t_{i(2)}^*, \ldots, t_{i(m-1)}^*]$$
$$= \left[\frac{t_1 - x_i'\beta}{\sigma}, \ldots, \frac{t_{m-1} - x_i'\beta}{\sigma}\right].$$

The standardized thresholds are now peculiar to the population defined in (3) to which the individual belongs. Each standardized threshold is the distance between the fixed threshold and the mean of said population in standard deviation units.

With m categories of response, for each $a^* = k$, say, there is a vector

$$G' = [\theta_1, \theta_2, \ldots, \theta_m],$$

with

$$\sum_{j=1}^{m} \theta_j = 1$$

corresponding to the distribution of response probabilities. G is the genotype in the outward scale of an individual with additive genetic value equal to k in the underlying scale, and θ_j is its probability of response in the j^{th} category. Omitting the population subscript in the standardized thresholds

$$\theta_j = \text{Prob}\{t_{j-1}^* < y^* < t_j^* \mid a^* = k\}$$
$$= \text{Prob}\{t_{j-1}^* - a^* < u^* + e^* < t_j - a^* \mid a^* = k\}. \quad (5)$$

The distribution of $s = u^* + e^*$ is normal and independent of a^*. Hence,

$$\theta_j = \int_{t_{j-1}^* - a^*}^{\infty} f(s)\,ds - \int_{t_j^* - a^*}^{\infty} f(s)\,ds$$
$$= \gamma_{j-1} - \gamma_j, \quad (6)$$

where f(s) is the density function of s. The genotype in the outward scale is then

$$G' = [\gamma_o - \gamma_1, \gamma_1 - \gamma_2, \ldots, \gamma_{m-1} - \gamma_m],$$

with $\gamma_o = 1$ and $\gamma_m = 0$, and it should be noted that G is a function of a^*. If in the standardized model of (4) the only sources of variation are a^* and e^*, from (5) and (6) with only two categories of response one obtains

$$\theta = \int_{t^* - a^*}^{\infty} [2\pi(1 - h^2)]^{1/2} \exp^{-e^{*2}/2(1-h^2)}\, de^* \quad (7)$$

which is the outward genotypic value derived by Dempster and Lerner (1950). The genotype in the outward scale can then be regarded as the proportion of environments in which an individual with underlying genetic value $a^* = k$ would exhibit the character. The conditional variance of phenotypes in the outward scale given a fixed genotypic value in such scale, i.e., the environmental variance, is given by

$$\theta_j(1 - \theta_j).$$

Hence, genetic and environmental effects in the outward scale are not statistically independent of each other.

In the general case, the genetic variance for the j^{th} category in the outward scale is

$$E[\theta_j - \bar{\theta}]^2 =$$
$$\int_{-\infty}^{\infty} \left[\int_{t_{j-1}^* - a^*}^{t_j^* - a^*} f(s)\,ds\right]^2 f(a^*)\,da^*$$
$$- \left[\int_{-\infty}^{\infty} \int_{t_{j-1}^* - a^*}^{t_j^* - a^*} f(s) f(a^*)\,ds\,da^*\right]^2, \quad (8)$$

where the expectation (E) is taken with respect to the distribution of a^* and this can be obtained numerically for any desired degree of approximation. While the model postulates that the genetic variance is entirely additive in the underlying scale, there is nonadditive genetic variance in the outward scale. The effect of a gene substitution in the underlying scale on the outward genotype is given by

$$\frac{\partial\theta_j}{\partial a^*} = \frac{\partial(\gamma_{j-1} - \gamma_j)}{\partial(t - a^*)} \frac{\partial(t - a^*)}{\partial a^*}$$
$$= [2\pi(1 - h^2)]^{-1/2} \{\exp^{-(t^*_{j-1} - a^*)^2/2(1-h^2)} - \exp^{-(t^*_j - a^*)^2/2(1-h^2)}\}. \quad (9)$$

Equation (9) illustrates that the outward effect of a gene substitution in the underlying scale is not constant throughout the range of a^*. Dempster and Lerner (1950) showed for binary traits that the relative contribution of nonadditive genetic variance to the total genetic variance in the outward scale increases as heritability increases and as the prevalence of the trait deviates from 50%. For example, if a binary trait has a 20% prevalence and if h^2 in the underlying scale is .36, the nonadditive genetic variance as a proportion of the total genetic variance in the outward scale is about 11 to 12%; if the prevalence is 10%, then this proportion is about 25%. Corresponding figures for a trait with $h^2 = .64$ are 22 and 40%, respectively. This implies that it may be difficult to obtain estimates of additive genetic parameters free of nonadditive biases if the scale of the raw data is used for analysis.

The additive genetic variance in the observed scale can be obtained by postulating a linear relationship between the additive genotype in the outward scale, G^A, and the additive genetic value in the underlying scale (see the Appendix by Robertson in Dempster and Lerner, 1950). Letting θ_j^A be the j^{th} element of G^A, and θ_j^D be a residual genotypic value, one can write

$$\theta_j = \theta_j^A + \theta_j^D = \alpha_j + E\left[\frac{\partial\theta_j}{\partial a^*}\right] a^* + \theta_j^D,$$

where α_j is a location constant. Hence

$$\mathrm{Var}(\theta_j^A) = E^2\left[\frac{\partial\theta_j}{\partial a^*}\right] h^2, \quad (10)$$

where the expectation is taken with respect to the distribution of a^*. Now,

$$E\left(\frac{\partial\theta_j}{\partial a^*}\right) = (2\pi)^{-1/2}[\exp^{-t^{*2}_{j-1}/2} - \exp^{-t^{*2}_j/2}] = z_{j-1} - z_j, \quad (11)$$

following from an extension of results from Robertson (Dempster and Lerner, 1950); z_{j-1} and z_j are ordinates of a standard normal density function corresponding to thresholds between categories $j - 1$ and j, and j and $j + 1$, respectively. From (10) and (11),

$$\mathrm{Var}(\theta_j^A) = (z_{j-1} - z_j)^2 h^2, \quad (12)$$

so the additive genetic variance in the observed scale depends on the distribution of prevalence by category of response in the population in question. Even if h^2 is the same from population to population, differences in prevalence will result in different amounts of additive genetic variance.

The phenotype in the outward scale is an $m \times 1$ random vector, v, which has elements equal to 1 in the position corresponding to the category of response and zero elsewhere. The variance-covariance matrix of v has elements

$$\pi_i(1 - \pi_i),\ i = 1, \ldots, m$$

and

$$-\pi_i\pi_j \text{ for } i \neq j;$$

π_i is the probability of response in the i^{th} category. From (12) the "heritability of the j^{th} category" is

$$h^2_{oj} = h^2(z_{j-1} - z_j)^2/\pi_j(1 - \pi_j), \quad (13)$$

which reduces to $h^2_o = h^2 z^2/\pi(1 - \pi)$ with two categories of response, a well known result due to Dempster and Lerner (1950).

A common practice in animal breeding is to score response categories by a vector of weights (Hazel, 1943; Pollak and Freeman, 1976). Letting

$$\eta = [\eta_1, \eta_2, \ldots, \eta_m]$$

be such an $m \times 1$ vector of scores, the aggregate genetic value in the outward scale is defined as $\eta' G^A$ and the phenotypes are scored as $P = \eta' v$. The heritability of this phenotypic score is then

$$h^2_o = \eta' \mathrm{Var}(G^A)\eta/\eta' \mathrm{Var}(v)\eta.$$

However,

$$G^A = \{\alpha_j + (z_{j-1} - z_j)a^*);\ \text{for } j = 1, \ldots, m.$$

Hence, $\mathrm{Var}(G^A)$ is a symmetric matrix obtained by multiplying h^2 times the vector

$$\{z_{j-1} - z_j\}$$

times its transpose. Thus

$$\eta'\mathrm{Var}(G^A)\eta = h^2\left[\sum_{j=1}^{m} \eta_j(z_{j-1} - z_j)\right]^2 = h^2\left[\sum_{j=1}^{m-1} z_j(\eta_{j+1} - \eta_j)\right]^2 \quad (14)$$

since $z_o = z_m = 0$. Similarly

$$\mathrm{Var}(P) = \eta'\mathrm{Var}(v)\eta = \sum_{j=1}^{m} \eta_j^2 \pi_j - \left(\sum_{j=1}^{m} \eta_j \pi_j\right)^2 \quad (15)$$

The heritability of the scores is given by

$$h_o^2 = h^2\left[\sum_{j=1}^{m-1} z_j(\eta_{j+1} - \eta_j)\right]^2 / \left[\sum_{j=1}^{m} \eta_j^2 \pi_j - \left(\sum_{j=1}^{m} \eta_j \pi_j\right)^2\right] \quad (16)$$

which for two response categories reduces to

$$h^2 z^2/[\pi(1-\pi)].$$

In the case of two categories h_o^2 is invariant to η, this is not so in the general case. Gianola and Norton (1981) describe a scaling procedure with several optimality properties.

Bivariate Aspects of the Threshold Model. In designing animal improvement schemes, information is also needed on genetic and phenotypic associations between pairs of continuous variables, pairs of discrete variables and mixtures of continuous and discrete variables. For example, an index for sheep selection may include records on fleece weight, 90-d body weight, foot-rot classification and litter size. Foot-rot classification and litter size may be regarded as polychotomous variables having an underlying normal distribution. Results presented here have also been obtained in a path analysis framework by Vinson et al. (1976).

Consider two underlying bivariate standard normal variables (w^* and y^*) represented by a model similar to that of equation (4), except that the subscripts will now indicate the random variable in question:

$$w^* = u_w^* + a_w^* + e_w^*$$

$$y^* = u_y^* + a_y^* + e_y^*.$$

Since the variates are in standard deviation units, the correlation between w and y is

$$\rho_{wy} = \mathrm{Cov}(u_w^*, u_y^*) + \mathrm{Cov}(a_w^*, a_y^*) + \mathrm{Cov}(e_w^*, e_y^*) = \rho_{wy(u)} q_w q_y + \rho_{wy(a)} h_w h_y + \rho_{wy(e)} e_w e_y \quad (17)$$

where $\rho_{wy(u)}$ is the correlation between the u values of the variables w and y, and $\rho_{wy(a)}$ and $\rho_{wy(e)}$ are genetic and environmental correlations.

The phenotypes in the outward scales are n and m polychotomies induced by $n - 1$ and $m - 1$ thresholds in the distributions of w^* and y^*, respectively. Let $W = i$ and $Y = j$ index a response in the i^{th} category of the outward variable W and j^{th} category of the outward variable Y. The joint distribution of W and Y is characterized by the statement

$$\text{Prob}\{W = i, Y = j\} = \text{Prob}\{t^*_{w(i-1)} < w^* \leq t^*_{w(i)}, t^*_{y(j-1)} < y^* < t^*_{y(j)}\} = \int_{t^*_{w(i-1)}}^{t^*_{w(i)}} \int_{t^*_{y(j-1)}}^{t^*_{y(j)}} f(w^*, y^*)\,dw^*\,dy^*, \quad (18)$$

where $f(w^*, y^*)$ is a bivariate normal density function. Now, write

$$\text{Prob}(W = i) = \beta_{oi} + \beta_{1i} w^* + E_i \quad (19)$$

$$\text{Prob}(Y = j) = \beta_{oj} + \beta_{1j} y^* + E_j', \quad (20)$$

where E_i and E_j' are mutually independent residuals which are also independent of other terms in (19) and (20). Thus

$$\text{Cov}[\text{Prob}(W = i), \text{Prob}(Y = j)]$$

$$= \frac{\partial(W = i)}{\partial w^*} \frac{\partial(Y = j)}{\partial y^*} \rho_{wy}$$

$$= (z_{i-1} - z_i)(z'_{j-1} - z'_j)\rho_{wy} \qquad (21)$$

where (21) follows from an argument similar to the one used in (11); z_i is the ordinate of a standard normal variate corresponding to the threshold between categories i and i + 1 of W, and z'_j is its counterpart for Y.

The vector of scores for the categories of W may have the form

$$\eta_1 = [1, 2, \ldots, i, \ldots, n]'.$$

Likewise, the vector of scores for the categories of Y may be

$$\eta_2 = [1, 2, \ldots, j, \ldots, m]'.$$

From (14) and (21)

$$\text{Cov}(W,Y) = \left[\sum_{i=1}^{n-1} z_i\right]\left[\sum_{j=1}^{m-1} z'_j\right]\rho_{wy}. \qquad (22)$$

Hence, the correlation in the underlying scale, ρ_{wy}, can be obtained from the correlation in the outward scale, ρ_{WY}, as follows:

$$\rho_{wy} = \rho_{WY}\,\sigma_W\,\sigma_Y \left[\sum_{i=1}^{n-1} z_i\right]^{-1}\left[\sum_{j=1}^{m-1} z'_j\right]^{-1} \qquad (23)$$

which agrees with the result of Vinson et al. (1976). The additive genetic variance of W and Y and their genetic covariance in the outward scale follow from (14)

$$\text{Var}(W^A) = h_w^2\left[\sum_{i=1}^{n-1} z_i\right]^2 \qquad (24)$$

$$\text{Var}(Y^A) = h_y^2\left[\sum_{j=1}^{m-1} z'_j\right]^2 \qquad (25)$$

$$\text{Cov}(W^A, Y^A) = \rho_{wy(a)}\,h_w\,h_y\left[\sum_{i=1}^{n-1} z_i\right]\left[\sum_{j=1}^{m-1} z'_j\right]. \qquad (26)$$

Hence, $\rho_{wy(a)} = \rho_{WY(a)}$, in agreement with results of Vinson et al. (1976).

A case of interest is the one in which one of the two continuous variables, y*, say, is observable. Then

$$\text{Cov}\{(W = i), y^*\} = (z_{i-1} - z_i)\rho_{wy},$$

$$\text{Cov}(W, y^*) = \left[\sum_{i=1}^{n-1} z_i\right]\rho_{wy},$$

from which

$$\rho_{wy} = \rho_{Wy^*}\,\sigma_W\left[\sum_{i=1}^{n-1} z_i\right]^{-1} = \rho_{Wy}\,\sigma_W\left[\sum_{i=1}^{n-1} z_i\right]^{-1}. \qquad (27)$$

A similar relationship applies to the correlations between additive genetic values, environmental deviations and other random components of the model.

Probability Statements in Threshold Models. It may be of interest to calculate the probability that a certain individual exhibits a response in one of *m* possible categories given information on the distribution of outcomes in a set of relatives. This has received attention and application in genetic counseling in Homo sapiens where the aim has been to assess relative recurrence risks of specific diseases given information on relatives (Curnow, 1972, 1974; Mendell and Elston, 1974; Smith and Mendell, 1974; Curnow and Smith, 1975).

Curnow (1972) proposed a probability model for dichotomous outcomes based on the concept of a risk function. However, as pointed out by Bulmer (1980) this model is equivalent to a model with abrupt thresholds. The method of Curnow (1972) is useful in instances in which multiple integrals involving the multinormal density function can be written as products of single integrals (Curnow and Dunnett, 1962). An extension of the method incorporating information on concomitant continuous variables has been described by Curnow (1974). This could be useful for the analysis of traits such as calving difficulty where the underlying variate may be correlated with birth weight. Information on this trait for a set of relatives could be utilized jointly with calving difficulty scores in the planning of a breeding program aimed to improve calving ease.

Unordered Categories of Response: The External Concept. Polychotomous data also arise in ways in which no explicit ordering of the response categories is possible. For exam-

ple, consider an experiment in which cows of a certain breed are evaluated for their choice of one among m diets at a specific time of the day. If the trial last 2 wk, the data might be summarized as the number of cow-days that a specific diet is chosen. There is a need to distinguish between "design" and "response factors", and both can have a crossed or nested arrangement. In this hypothetical experiment there might have been four diets resulting from the combination of two sources of animal protein and two levels of an additive, thus defining a crossed response structure.

Thurstone (1927) and Bock and Jones (1968) developed the extremal model to account for the mechanism by which an underlying continuous scale maps into the observed scale. In this model, the underlying process is vector valued, i.e., whenever an experimental unit is subject to a treatment combination, a vector $\mathbf{y}$ of order m arises in the underlying scale. The response category corresponds to the largest element of $\mathbf{y}$. The probability that the individual responds in the m^{th} category is

$$\text{Prob}\{y_m > y_1, y_m > y_2, \ldots, y_m > y_{m-1}\} = \text{Prob}\{\bigcap_{i \neq m} (y_m - y_i) > 0\}$$

where the symbol $\cap$ indicates "intersection". The differences

$$y_m - y_i,\ i = 1, \ldots, m-1,$$

can be described by the linear combination

$$\mathbf{y}^* = \mathbf{C}\mathbf{y},$$

where $\mathbf{C}$ is an $(m-1) \times m$ matrix with its j^{th} row having zeroes except for $c_{jj} = -1$ and $c_{jm} = 1, j = 1, \ldots, m-1$. If

$$\mathbf{y} \sim N(\mu, \mathbf{V}),$$

then

$$\mathbf{y}^* \sim N(\mathbf{C}\mu, \mathbf{CVC}')$$

and

$$\text{Prob}\{\bigcap_{i \neq m} (y_m - y_i) 0\} = \int_0^\infty \int_0^\infty \cdots \int_0^\infty k \exp^{-\frac{1}{2}(\mathbf{y}^* - \mathbf{C}\mu)'(\mathbf{CVC}')^{-1}(\mathbf{y}^* - \mathbf{C}\mu)} dy_1^* dy_2^* \cdots dy_{m-1}^* \quad (28)$$

where

$$k = (2\pi)^{-(m-1)/2} |\mathbf{CVC}'|^{-\frac{1}{2}}.$$

If $\mathbf{V}$ is such that all the diagonal elements are equal, say to σ^2, and all off-diagonals are equal to ϕ, then

$$\text{Var}(y_m - y_i) = 2(\sigma^2 - \phi) = \gamma$$

$$\text{Cov}(y_m - y_i, y_m - y_j) = \sigma^2 - \phi = \gamma/2$$

and

$$\text{Corr}(y_m - y_i, y_m - y_j) = 1/2.$$

In this case (Bock and Jones, 1968), the $m-1$ variate logistic distribution can be used to approximate a multivariate normal and

$$\text{Prob}\{\bigcap_{i \neq m} (y_m - y_i) > 0\} \doteq [1 + \exp^{-(\mu_m - \mu_1)} + \cdots + \exp^{-(\mu_m - \mu_{m-1})}]^{-1} = \exp^{\mu_m}/(\exp^{\mu_1} + \exp^{\mu_2} + \cdots + \exp^{\mu_m}). \quad (29)$$

The marginal distributions of (29) are logistic with mean 0 and variance $\pi^2/3$, the covariance for any pair of variables is $\pi^2/6$ and the correlation is 1/2. While application of the multivariate normal model is constrained by the difficulty of integrating (28) when the number of variates exceeds two, the approximation of equation (29) is excellent for moderate values of $t_i = \mu_m - \mu_i$ (Bock, 1975).

In the case of two categories of response, the threshold and extremal concepts are formally equivalent (Bock, 1975) since the vector valued function involved in the extremal model can be reduced to a univariate random variable given by the difference between the two elements of $\mathbf{y}$. If

$$\mu' = [\mu_1, \mu_2],$$

$$\text{Var}(y_1) = \sigma_1^2,$$

$$\text{Var}(y_2) = \sigma_2^2 \text{ and}$$

$$\text{Cov}(y_1, y_2) = \sigma_{12},$$

a response in the first category occurs if $y_1 > y_2$. Letting

$$f = y_1 - y_2,$$

$$\text{Var}(f) = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} = \sigma^2, \text{ and}$$

$$E(f) = f^*,$$

the probability of response in the first category is given by

$$P_1 = (\sqrt{2\pi}\,\sigma)^{-1} \int_0^\infty \exp^{-(f-f^*)^2/2\sigma}\, df$$
$$= (2\pi)^{-1/2} \int_{-f^*/\sigma}^\infty e^{-t^2/2}\, dt = \Phi(f^*/\sigma) \quad (30)$$

and $P_2 = 1 - \Phi(f^*/\sigma)$. By taking $\sigma = 1$ as the unit of measurement the result of equation 1 for the threshold model follows.

Estimation of Heritability

Outward Scale. In view of previous considerations, there is question as to the significance of heritability estimates obtained directly in the outward scale. There may be considerable nonadditive genetic variance present and the assumption of independence of genetic and environmental effects is generally not tenable. For binary traits, Robertson and Lerner (1949) presented formulae for a one-way classification model with two random sources of variability: "among" and "within" families. Their developments are based on an analysis of variance framework with the outward variate scored as zero or one, depending on the absence or presence of the characteristic in the individual in question. Because the variance in the outward scale is binomial

$$\hat{h}^2 = \frac{[SSF/\pi(1-\pi)] - (s-1)}{r(k-s+1)}, \quad (31)$$

where SSF is the corrected sum of squares "due to" families, π is the prevalence of the character in the general population, s is the number of families, r is the additive relationship between family members, and

$$k = \sum_i n_i - \sum_i n_i^2 / \Sigma n_i,$$

with n_i the number of individuals in the i^{th} family. Since $SSF/\pi(1-\pi)$ is chi-square with $s-1$ degrees of freedom, with large n_i's,

$$\text{Var}(\hat{h}^2) = \frac{2(s-1)}{r^2(k-s+1)^2}. \quad (32)$$

An application of this method was presented by Milagres et al. (1979), in which the authors obviated effects of fixed classifications in the model by computing chi-square statistics within levels of fixed effects and pooling them across levels.

The method can be extended to multiple categories of response. Let v_{ij} be an $m \times 1$ vector corresponding to the j^{th} individual in the i^{th} family ($i = 1, \ldots, s$). If the individual responds in the m^{th} category, then v_{ij} has a 1 in the m^{th} position and 0's elsewhere. The ij^{th} individual is scored as $\eta' v_{ij}$, where η is a vector of scores and the model is

$$\eta' v_{ij} = \mu + s_i + w_{ij}, \quad (33)$$

where s_i is the random effect of the i^{th} family and w_{ij} is a residual. Taking into account that

$$\sigma_w^2 = \eta' \text{ var}(v_{ij})\eta - \sigma_s^2,$$

$$\hat{h}^2 = \frac{SSF/\{\eta' \text{Var}(v_{ij})\eta\} - (s-1)}{r(k-s+1)}. \quad (34)$$

Which reduces to equation (31) when $m = 2$;

$$\eta' \text{Var}(v_{ij})\eta$$

can be calculated from (15). Estimates of heritability in the underlying scale can be obtained by the application of (16) to (31) or (34).

Van Vleck (1972) found in a Monte Carlo study that (16) yielded a slight overestimate of heritability in the underlying scale when paternal half-sib correlations were computed from binomial data. However, substantial overestimates resulted when parent-offspring correlations were used; this was particularly true at low or high prevalences and when heritability in the underlying scale was high. These results are consistent with the theory developed by Dempster and Lerner (1950).

Equations (31) and (34) have a multivariate counterpart. Since v_{ij} follows a multinomial distribution with fixed n_i in the i^{th} family, one of the elements in the vector is redundant. Omitting the last element of v_{ij}, the model can be written as

$$\tilde{v}_{ij} = \mu + s + w$$

where $\tilde{v}_{ij}$, μ, s, and w are $(m-1) \times 1$ vectors. The complete data set can be written as

$$Y = \begin{bmatrix} \tilde{v}'_{11} \\ \tilde{v}'_{12} \\ \vdots \\ \tilde{v}'_{sn_s} \end{bmatrix} = [y_1, y_2, \ldots, y_{m-1}]$$

where Y is a matrix of order

$$\sum_i n_i \times (m-1).$$

Further,

$$\mathrm{Var}(y_i) = \sum_{j=1}^{s+} [\pi_i(1-\pi_i) - s_{ii}] I_j + \sum_{j=1}^{s+} J_j s_{ii} = Z_{ii} \text{ and} \quad (35)$$

$$\mathrm{Cov}(y_i, y'_k) = \sum_{j=1}^{s+} [-\pi_i \pi_k - s_{ik}] I_j + \sum_{j=1}^{s+} J_j s_{ik} = Z_{ij}, \quad (36)$$

where Σ^+ denotes direct sum, I_j is an identity matrix of order n_j corresponding to the j^{th} family, J_j is an n_j^2 matrix of 1's, and s_{ik} is the ik^{th} element of S, which is an $(m-1)^2$ symmetric matrix containing "family" variance and covariance components for the $m-1$ categories. With this in mind, one can compute F, the $(m-1)^2$ matrix of "among-families" corrected sums of squares and products corresponding to the $m-1$ independent categories so

$$E(F) = (s-1)V + (k-s+1)S \quad (37)$$

where V has elements $\pi_i(1-\pi_i)$ in the diagonal and $-\pi_i \pi_j$ in the off-diagonals; $i = 1, \ldots, m-1$. The "heritability" matrix $\hat{H}$ is the multivariate version of (31);

$$\hat{H} = [V^{-1}F - (s-1)I]/r(k-s+1) \quad (38)$$

where V^{-1} has diagonal elements equal to

$$\pi_i^{-1} + \pi_m^{-1}, \; i = 1, \ldots, m-1,$$

and off diagonals

$$\pi_m^{-1}$$

(Kendall and Stuart, 1977). The j^{th} element of $\hat{H}$ is the heritability of the j^{th} category and its ij^{th} element is

$$h_i h_j \rho_{ij}$$

where ρ_{ij} is the genetic correlation between the i^{th} and j^{th} category. An approximation to the variance of $\hat{H}$ can be obtained by writing F as $Y'AY$, where A is the matrix of quadratic or bilinear forms arising in the corrected sums of squares and products. Under normality, the variance of the ij^{th} element of F (Searle, 1971) would be

$$\mathrm{Var}(f_{ij}) = \mathrm{Var}(y'_i A y_j) = 2\mathrm{tr}(AZ_{ij})^2 + \mathrm{tr}(AZ_{ii}AZ_{ij}) + j'A[\pi_i^2(1-\pi_i)^2 Z_{jj} + \pi_j^2(1-\pi_j)^2 Z_{ii} + 2\pi_i\pi_j(1-\pi_i)(1-\pi_j)Z_{ij}]Aj, \quad (39)$$

where j is an

$$\sum_i n_i \times 1$$

vector of 1's and *tr* is the trace operation. From (38)

$$\hat{\mathrm{Var}}(H) = [V^{-1}\mathrm{Var}(F)V^{-1}][r(k-s+1)]^{-2} \quad (40)$$

with Var(F) computed from (39). Note from (35), (36) and (39) that Var(F) is a function of S which generally will not be known.

Little work has been done on variance component estimation with categorical data. Landis and Koch (1977) presented a multivariate analysis procedure which yields unbiased estimates for a one-way classification with unbalanced data. The procedure can be extended, at least in theory, to classifications with fixed and additional random factors but comparisons of this method with established procedures in animal breeding such as those of Henderson (1953) or maximum likelihood (Searle, 1979) are lacking. Leonard (1972) has described a Bayesian procedure for the binomial distribution in which the estimates depend little on

prior parameter values. However, Thompson (1979) warns that this method yields zero estimates for variance components when nonzero solutions were "expected".

Regression of Relatives on Propositi. Falconer (1965) developed a method for estimating heritability of liability to disease from prevalence in relatives. This procedure has received little attention in animal breeding but could be used, for example, for estimating the heritability of propensity to twinning. Assumptions of the method are: 1) an underlying normal distribution of liabilities in the general population with the individuals showing the disease (propositi) being those whose liabilities exceed a certain fixed threshold; and 2) the distribution of liabilities in relatives of affected individuals is also normal with variance equal to that in the general population. The method applies to situations in which the genetic component is multifactorial and excludes cases in which the variation of liability is discontinuous such as would happen with a gene of major effect.

Consider a population with a prevalence, $\bar{q}$, of a certain disease. The mean liability for affected individuals is thus

$$a = z/\bar{q},$$

where z is the ordinate of the truncation point,

$$x = \Phi^{-1}(1 - \bar{q}),$$

in a standard normal distribution. Since the mean liability in the general population is 0, then *a* can be regarded as a selection differential for liability and x is the distance between the threshold and the mean of the population in standard deviation units. In relatives of the propositi, the prevalence is $\bar{q}_R$ and the deviation of the threshold from the mean liability of the relatives is

$$x_R = \Phi^{-1}(1 - \bar{q}_R).$$

By using the concept of response to selection, the regression of relatives on propositi is

$$b = (x - x_R)/a$$

and because the variance of liabilities in propositi and relatives is the same, b yields the correlation between relatives. Hence, in the absence of or elimination of common familial environmental influences,

$$\hat{h}^2 = (x - x_R)/a\, r, \tag{41}$$

where r is the additive relationship between relatives. Asymptotically (Falconer, 1965),

$$\mathrm{Var}(\hat{h})^2 = r^{-2}\left\{[1/a - b(a - x)]^2 \frac{(1 - \bar{q})}{a^2 N} + \frac{(1 - \bar{q}_R)}{a^4 N_R}\right\} \tag{42}$$

where N and N_R are the sample sizes from which q and $\bar{q}_R$ are calculated, respectively.

If the prevalence of a character differs between sexes, interesting results are obtained. For example, the prevalence of congenital pyloric stenosis is lower in females than in males, but the prevalence in relatives of female propositi is higher than is found in relatives of male propositi (Carter, 1961). This can be explained by the threshold model, as follows: the females have a lower mean liability than males and, hence, a larger "selection differential", and if liability is inherited, the expected "response to selection" will be larger in females than in males (Falconer, 1965). If data on the two sexes are available, four separate regressions can be calculated, two for like-sexed and two for unlike-sexed relatives. The first two estimate h_m^2 and h_f^2 (where m = males and f = females) and the second two both estimate $h_m h_f r_g$ (Falconer, 1965), where r_g is the genetic correlation between sexes. The two estimates from unlike-sexed relatives should be the same and if all four estimates are the same, within the limits of sampling errors, then the genetic correlation does not differ from unity. In this latter case the four estimates should be combined statistically. Falconer (1965) pointed out that if the propositi and their relatives belong to different generations, the estimate of heritability would be valid only if the variance of liability remained unchanged from generation to generation. However, it was shown later (Falconer, 1967) that the method adjusts automatically for this source of bias. Two sources of bias are present in Falconer's (1965) method. Since the propositi form a truncated group with a skewed distribution, the variance of liabilities among relatives may be smaller than in the general population and the distribution of liabilities in relatives may also be skewed (Smith, 1970). Numerical analyses by

Smith (1970) indicate that Falconer's (1965) method underestimates the correlation between relatives by about 10% of the true value, so the bias is only important when the correlation is high. Refinements of Falconer's method are given by Reich et al. (1972) and Mendell and Elston (1974).

The frequency in relatives can be calculated from a method presented by James (1971). Suppose individuals are taken at random from the whole population and given the score X = 1 if they have the attribute and X = 0 if they do not. Relatives of these individuals are scored in the same way, with the variable Y denoting their scores. Since X has only two values, the conditional expectation of Y given X = 1 is linear in X and given by

$$\bar{q}_R = \bar{q} + \frac{\mathrm{Cov}(Y,X)}{\bar{q}(1-\bar{q})}(1-\bar{q})$$

$$= \bar{q} + \frac{\mathrm{Cov}(Y,X)}{\bar{q}}, \qquad (43)$$

which yields the expected frequency in relatives of individuals showing the attribute. Cov(Y,X) is the covariance between relatives which can be obtained by standard procedures.

Underlying Scale. Tallis (1962) presented an application of maximum likelihood to the estimation of correlation between relatives and repeatability of records. The data are arranged into a p × q contingency table, with one of the dimensions the ordered categories of response in the parent (first record in the repeatability model) and the other the categories of response in the progeny (or second record). Hence, n_{ij} would be the number of observations in the i^{th} category of parental responses and j^{th} category of progeny responses, with

$$\sum_i \sum_j n_{ij} = n$$

taken as fixed. If the underlying variables in the two dimensions follow a bivariate normal distribution, with thresholds in each of the axes given by

$$\mathbf{a}' = (a_1, a_2, \ldots, a_{p-1})$$

and

$$\mathbf{b}' = (b_1, b_2, \ldots, b_{q-1}),$$

then

$$P_{ij} = \int_{a_{i-1}}^{a_i} \int_{b_{j-1}}^{b_j} \phi(u,v;\rho)\, du\, dv$$

$$= \Phi(a_i,b_j;\rho) - \Phi(a_{i-1},b_j;\rho)$$

$$- \Phi(a_i,b_{j-1};\rho) + \Phi(a_{i-1},b_{j-1};\rho) \qquad (44)$$

where $\phi(u,v;\rho)$ is a bivariate standard normal density function with correlation ρ, and Φ is the cumulative distribution function. An extension of results of Tallis (1962) yields

$$\frac{\partial \Phi(a_i,b_j;\rho)}{\partial \rho} = \phi(a_i,b_j;\rho),$$

$$\frac{\partial \Phi(a_i,b_j;\rho)}{\partial a_i} = \phi(a_j)\,\Phi\left(\frac{b_j - \rho a_i}{\sqrt{1-\rho^2}}\right)$$

$$= \phi(a_i)\Phi(B_{ij}) \text{ and}$$

$$\frac{\partial \Phi(a_i,b_j;\rho)}{\partial b_j} = \phi(b_j)\Phi\left(\frac{a_i - \rho b_j}{\sqrt{1-\rho^2}}\right)$$

$$= \phi(b_j)\,\Phi(A_{ij}),$$

where $\phi(a_i)$ and $\phi(b_j)$ are standard normal densities evaluated at a_i and b_j, respectively,

$$B_{ij} = \frac{b_j - \rho a_i}{\sqrt{1-\rho^2}}, \text{ and}$$

$$A_{ij} = \frac{a_i - \rho b_j}{\sqrt{1-\rho^2}}.$$

The log-likelihood equations are

$$\frac{\partial \ln L}{\partial \rho} = \sum_i \sum_j \frac{n_{ij}}{P_{ij}} \frac{\partial P_{ij}}{\partial \rho}$$

$$= \sum_i \sum_j \frac{n_{ij}}{P_{ij}} [\phi(a_i,b_j;\rho)$$

$$- \phi(a_{i-1},b_j;\rho) - \phi(a_i,b_{j-1};\rho)$$

$$+ \phi(a_{i-1},b_{j-1};\rho)], \qquad (45)$$

$$\frac{\partial \ln L}{\partial a_i} = \phi(a_i) \sum_j \left\{\frac{n_{ij}}{P_{ij}} - \frac{n_{(i+1)j}}{P_{(i+1)j}}\right\}$$

$$\{\Phi(B_{ij}) - \Phi(B_{i(j-1)})\} \qquad (46)$$

for i = 1, . . . , p−1, and

$$\frac{\partial \ell n\, L}{\partial b_j} = \phi(b_j) \sum_i \left\{ \frac{n_{ij}}{P_{ij}} - \frac{n_{i(j+1)}}{P_{i(j+1)}} \right\} \{\Phi(A_{ij}) - \Phi(A_{(i-1)j})\} \quad (47)$$

$$\text{for } j = 1, \ldots, q-1.$$

These equations are nonlinear and must be solved by iterative procedures such as the Newton-Raphson method (see, e.g., Dahlquist and Björck, 1974). This method requires evaluation of the inverse of the matrix of second partial derivatives of the log-likelihood with respect to the parameters. An additional computational difficulty is that bivariate volumes need to be calculated. The large-sample variance-covariance matrix of the maximum likelihood estimates can be obtained as usual by evaluation of the negative inverse of the matrix of second partials at the solution.

The method of Tallis (1962) as extended here addresses only a subset of problems encountered in animal breeding: there can be only two populations of thresholds corresponding to the parental and progeny generations or first and second records in the repeatability model. In the case of sib-data in which families are independent, the likelihood function is the product of the likelihood of the families and Thompson (1972) has presented a completely general solution for independent families in which each individual can belong to a different population. The computations are formidable and require the evaluation of k-dimensional integrals where k is the family size. In some instances it is possible to write these multiple integrals as products of one-dimensional normal integrals (Curnow and Dunnett, 1962). If the families are not independent, computations involved in maximum likelihood are generally prohibitive.

Plackett (1965) has presented a simple method of estimating the correlation between two underlying variates each of which is dichotomous in the outward scale. The data consist of a fourfold table with entries n_{11}, n_{12}, n_{21} and n_{22}, where n_{ij} is the number of records in the i^{th} category of the parents (or first record in the repeatability model) and j^{th} category of the progeny (or second record). The estimator is

$$\hat{\rho} = -\cos\left[\Psi^{1/2}\pi/(1+\Psi^{1/2})\right] \quad (48)$$

where

$$\Psi = n_{11}n_{22}/n_{12}n_{21}.$$

Further,

$$\hat{\mathrm{Var}}(\hat{\rho}) = \pi^2 \sin^2\left\{\frac{\Psi\pi^{1/2}}{1+\Psi^{1/2}}\right\} \Psi^2\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right) [4\Psi(1+\Psi^{1/2})^4]^{-1} \quad (49)$$

gives a consistent estimator of $\mathrm{Var}(\hat{\rho})$. Rutledge (1977), in a Monte Carlo study, found that $\hat{\rho}$ had much less bias than a statistic proposed by Lush (1956) for estimating the correlation in the outward scale. However, the bias of Plackett's statistic seemed to be affected by the threshold location and by the value of ρ. It would be interesting to compare the maximum likelihood estimator of ρ with Plackett's (1965) estimator since the latter is much easier to compute. For a data set found in Plackett's paper, his estimator had a larger estimated sampling variance than the maximum likelihood estimator of ρ. Wahrendorf (1980) has extended Plackett's (1965) statistic to p × q contingency tables.

Linear Models for Categorical Traits

Fixed Effects. Grizzle et al. (1969) developed a linear model approach to the analysis of categorical data. The procedure is available in a release of the Statistical Analysis System (Helwig and Council, 1979) under the FUNCAT procedure. The data are arranged into a $s \times r$ contingency table where the s rows represent factorial combinations and the r columns are mutually exclusive and exhaustive response categories. The entries in the table,

$$n_{ij},\ i = 1, \ldots, s, \text{ and } j = 1, \ldots, r,$$

are counts and

$$n_{i.} = \sum_j n_{ij}$$

indicates the marginal total for the i^{th} factorial combination. Let π_{ij} be the expected cell probabilities such that

$$\sum_j \pi_{ij} = 1,$$

and

$$p_{ij} = n_{ij}/n_{i.}$$

be the observed probabilities. On defining

$$\pi_i' = [\pi_{i1}, \pi_{i2}, \ldots, \pi_{ir}];$$

$$p_i' = [p_{i1}, p_{i2}, \ldots, p_{ir}]';$$

$$\pi' = [\pi_1', \pi_2', \ldots, \pi_s'];$$

$$p' = [p_1', p_2', \ldots, p_s']$$

the variance-covariance matrix of p_i is symmetric with elements

$$\pi_{ij}(1 - \pi_{ij})/n_{i.},$$

for $j = 1, \ldots, r$, and

$$-\pi_{ij}\pi_{ik}$$

for $j \neq k$. Hence,

$$\mathrm{Var}(\mathbf{p}) = \sum_{i=1}^{s+} \mathrm{Var}(\mathbf{p}_i)$$

is block diagonal of order rs × rs.

In view of the restrictions

$$\pi_{i.} = 1, \text{ for } i = 1, \ldots, s$$

up to $s(r - 1)$ independent functions of cell probabilities may be selected to describe the contingency table. These functions,

$$f_m(\pi), \text{ with } m = 1, \ldots, u < s(r - 1),$$

can be of any form provided that they have partial derivatives up to second order with respect to π_{ij}. Compactly

$$\mathbf{F}'(\pi) = [f_1(\pi), f_2(\pi), \ldots, f_u(\pi)] \tag{50}$$

$$\mathbf{F}' = [f_1(\mathbf{p}), f_2(\mathbf{p}), \ldots, f_u(\mathbf{p})] \tag{51}$$

describe the contingency table where

$$\mathbf{F} = \mathbf{F}(\pi = \mathbf{p}).$$

The asymptotic variance-covariance matrix of $\mathbf{F}$ is estimated by $\mathbf{S}$, a matrix of order u × u

$$\mathbf{S} = \hat{\mathrm{V}}\mathrm{ar}(\mathbf{F}) = \left[\frac{\partial \mathbf{F}}{\partial \pi} \mid \pi = \mathbf{p}\right] \left[\sum_{i=1}^{s+} \mathrm{Var}(\mathbf{p}_i)\right] \left[\frac{\partial \mathbf{F}}{\partial \pi} \mid \pi = \mathbf{p}\right]' \tag{52}$$

When $\mathbf{F}$ is a linear function of $\mathbf{p}$, $\mathbf{S}$ is the exact variance-covariance matrix of $\mathbf{F}$. The next step is to describe the family of functions by

$$\mathbf{F}(\pi) = \mathbf{X}\beta \tag{53}$$

where $\mathbf{X}$ is a u × v matrix of rank $r < v < u$ and β is a v × 1 fixed vector of unknown parameters. Minimizing

$$\mathrm{SS}[\mathbf{F}(\pi) = \mathbf{X}\beta] = \mathbf{F}'\mathbf{S}^{-1}\mathbf{F} - \tilde{\beta}'(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})\tilde{\beta} \tag{54}$$

with respect to $\tilde{\beta}$ yields

$$\tilde{\beta} = (\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-} \mathbf{X}'\mathbf{S}^{-1}\mathbf{F} \tag{55}$$

from which it is possible to obtain best asymptotic normal estimators (BAN) of functions of β which are invariant to the generalized inverse

$$(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-}$$

when evaluated at $\beta = \tilde{\beta}$. BAN estimators have the same asymptotic properties as maximum likelihood estimators, are much easier to compute, and in some instances (Berkson, 1968) yield estimates and test statistics numerically similar to maximum likelihood. The form in (54) has an asymptotic chi-square distribution with u − r degrees of freedom so the model can be tested for fit. Given the model, the hypothesis

$$\mathrm{H}: \mathbf{C}\beta = \mathbf{0}$$

of rank d can be tested by referring

$$\mathrm{SS}[\mathbf{C}\beta = \mathbf{0}] = \tilde{\beta}'\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-}\mathbf{C}']^{-1}\mathbf{C}\tilde{\beta} \tag{56}$$

to a chi-square distribution with d degrees of freedom. Confidence intervals for $\mathbf{C}\beta$ can be obtained from

$$\mathbf{C}\tilde{\beta} \pm [\chi_d^2 \mathbf{C}(\mathbf{X}'\mathbf{S}^{-1}\mathbf{X})^{-}\mathbf{C}']^{1/2}$$

where χ^2_d is the appropriate percentage point of the χ^2 distribution with d degrees of freedom.

The method assumes that all cells in the table are filled but singularity in **S** can result if some of the n_{ij}'s are zero. The empty cell problem may require collapsing the table or replacing the empty cell by 1/r (Berkson, 1955). Additional limitations relate to the assignment of scores to the elements of π (see the examples in Grizzle et al., 1969). It may be of interest to obtain fitted probabilities by inverting (53) so

$$\hat{\pi} = \mathbf{F}^{-1}(\mathbf{X}\tilde{\beta})$$

yields an estimate of π if **F** is invertible. Even in this case, the resulting probability estimates for some classes of functions may not be in the range 0 to 1. This illustrates that although BAN estimators have the same asymptotic properties of maximum likelihood estimators, not all of them can be maximum likelihood. Some invertible functions of response probabilities such as logits provide estimates of π within the parametric range. When $\mathbf{F}(\pi)$ defines a family of logit transforms and r = 2, the method of Grizzle et al. (1969) reduces to Berkson's (1955) "minimum logit chi-square", widely used in bioassay. Koch et al. (1972) have generalized the method of Grizzle et al. (1969) for incompletely classified data.

Mixed Models. The multivariate general linear model is the main tool used for analysis of animal breeding data. When response variables are categorical, tests of hypotheses based on the assumption of normality are approximate and may be inadequate. In mixed models in the sense of Henderson (1973), where the aim is to estimate conditional means, additional problems arise when the data are categorical.

Cochran (1951) showed that the conditional expectation of a predictand given the data is the best selection rule under certain idealized conditions since it 1) minimizes mean square error of prediction among all predictors; 2) is unbiased; 3) maximizes the expected value of the predictand in a group that has been selected by truncation; 4) maximizes the correlation between predictor and predictand; and 5) if the predictand and the data follow a multivariate normal distribution, genetic progress is maximized.

In the multivariate normal distribution, the conditional expectation of the predictand given the data is linear in the observations. This gives a strong justification for using linear models and predictors in animal breeding. The optimal properties of best linear unbiased predictors (BLUP) of realized values of random variables (Henderson et al., 1959; Henderson, 1973, 1975) have given impetus to research in sire evaluation and these methods have been applied in the U.S. and elsewhere. However, if the data are categorical, the best linear predictor may not be a good approximation of the best selection rule. Suppose that in the underlying scale the additive genetic value (a) and the phenotype of an individual (y) follow a bivariate normal distribution and y and a are defined such that

$$E(a) = E(y) = 0,$$

$$\sigma_y = 1, \text{ and}$$

$$\text{Corr}(a,y) = h$$

where h is the square root of the heritability of the trait. Now, y is polychotomized in the outward scale by a set of thresholds

$$t_1, t_2, \ldots, t_{m-1},$$

defining an outward variable Y, i.e., if

$$t_{i-1} < y < t_i,$$

then Y is a score in the i^{th} category of response, with m mutually exclusive and exhaustive such categories. The joint distribution of a and Y is defined by

$$\begin{aligned} f_i(a) &= \text{Prob}\{Y = i,\ a < A < a + da\};\\ &\qquad i = 1, \ldots, m\\ &= \text{Prob}\{t_{i-1} < y < t_i,\\ &\qquad a < A < a + da\};\\ &\qquad i = 1, \ldots, m. \end{aligned}$$

Hence, the conditional distribution of a given Y = i is

$$f(a|Y = i) = \int_{t_{i-1}}^{t_i} \phi(a,y)\, da\, dy / \int_{t_{i-1}}^{t_i} \phi(y)\, dy;\ i = 1, \ldots, m$$

where $\phi(a,y)$ is the bivariate normal density. The conditional expectation or best predictor of a is

$$E(a|Y = i) = \int_{t_{i-1}}^{t_i} h^2 y\, \phi(y)dy / \int_{t_{i-1}}^{t_i} \phi(y)dy$$

$$= \frac{h^2}{\sqrt{2\pi}} (e^{-t_{i-1}^2/2} - e^{-t_i^2/2}) / [\Phi(t_i) - \Phi(t_{i-1})]. \quad (57)$$

If the threshold model is postulated as in this example, the best linear predictor of additive genetic value (BLP) can be calculated from equations (22) through (27) as

$$BLP(a) = \frac{h^2 \left(\sum_{i=1}^{m-1} e^{-t_i^2/2} \right)}{\sqrt{2\pi}\, \sigma_Y^2} (Y - \bar{Y}) \quad (58)$$

where $\bar{Y}$ and σ_Y^2 are the mean and variance of Y, respectively. Clearly, equations (57) and (58) differ.

Shaeffer and Wilton (1976) studied the method of Grizzle et al. (1969) for possible application to sire evaluation for calving ease, a categorically scored trait. They pointed out that the number of observations per subclass did not allow for estimation of π_i in each population. They suggested each subclass be considered a random sample from the same population i.e.,

$$\pi' = (\pi_1, \pi_2, \ldots, \pi_r),$$

where π_i is estimated from the overall total in each response category. With this assumption and when the same function of probabilities is calculated in each subclass

$$\mathbf{Ap} = [\mathbf{a}'\mathbf{p}_1, \mathbf{a}'\mathbf{p}_2, \ldots, \mathbf{a}'\mathbf{p}]'$$

the method of Grizzle et al. (1969) is equivalent to least-squares applied to the scores of each experimental unit. They concluded that since BLUP provides a modification of least-squares to accommodate random effects in the model, it would also yield an extension of the approach of Grizzle et al. (1969) to mixed models under the previous conditions. However, regarding all subclasses as random samples from the same population for the purpose of estimating π was not consistent with the model, where several populations were defined by specific combinations of fixed effects. There is also difficulty in assigning meaningful scores to the elements of **a**. Schaeffer and Wilton (1976) combined three categories of calving difficulty with two categories of calf condition and assigned scores from 1 (alive, normal or unobserved) to 6 (dead, extremel difficulty) to the six resulting subclasses. It is difficult to agree or disagree with the proposition that an "alive, extremely difficult birth" should receive a score of 3 while a "dead, slight difficult birth" would be scored as 5. Further, in combining calving difficulty with calf condition as a single variable, information provided by the association between these two response variables may be lost. Pollak and Freeman (1976) and Berger and Freeman (1978) presented applications of BLUP to sire evaluation for calving difficulty in dairy cattle.

The problem of assigning scores to response categories is not a trivial one. If the response process is related to an underlying normal model with thresholds, (16) shows that heritability in the outward scale is generally not invariant to the scores. In fact, it is possible to develop a set of scores "more heritable" than any other set of scores (Gianola and Norton, 1981). Since heritability in the outward scale is always smaller than the underlying heritability, maximizing (16) with respect to the scores would yield a scoring criterion that would mimic a transformation from the outward to the underlying scale. Fisher (1938) developed a comparable scoring procedure by maximizing the ratio of treatment sum of squares to the total sum of squares. Snell (1964) has presented a method for determining numerical scores for ordinal categorical data arising from an underlying normal distribution. The procedure consists of estimating the thresholds of a logistic distribution (which resembles the normal distribution throughout the real line) by maximum likelihood and then developing scores by taking the mid-points of neighboring estimated thresholds except in the extreme categories where a different calculation is needed. Since the logistic distribution has variance $\pi^2/3$, one may wish to multiply the estimated thresholds by $\sqrt{3}/\pi$ in order to approximate more closely the spread of a standard normal distribution. Tong et al. (1977) applied Snell's scoring procedure to type classification records and calving ease data. In the case of highly skewed calving ease data, they found that the heritability of calving ease calculated from Snell's scores was larger than

that of equally spaced scores but the rank correlation between sire evaluations obtained from the two sets of scores was .98. Their results are data specific and no generalization can be made.

The method of sire evaluation for calving difficulty presented by Berger and Freeman (1978) consists of applying BLUP to raw scores with the variance-covariance matrix of the residuals taken as block diagonal with blocks corresponding to parity classes, and maintaining a single variance-covariance matrix for the random effects. In the context of a threshold model, as pointed out before, the conditional variance of phenotypes in the outward scale given a fixed genotypic value, i.e., the environmental variance is not constant throughout the range of genotypic values. Hence, independence between residuals and predictands cannot be assumed in the observed scale. Further, the additive genetic variance in the outward scale depends on the position of the thresholds and applications of mixed linear models consistent with the threshold concept should attempt to consider this problem. Presence of nonadditive genetic variance in the observed scale is another complicating factor in evaluations from the raw scores. With skewed distributions it may be difficult to obtain estimates of heritability free of dominance or epistatic biases and mixed model predictors under these circumstances will yield unnecessarily large prediction error variances.

The preceding methods of sire evaluation for categorical traits have not taken advantage of the probability structure of the data. If the sum of response probabilities must equal 1, this should be taken into account in the estimation procedure since it would be possible to derive predictors with smaller sampling variance than those obtained without restrictions in the estimation space. This is well known in linear estimation (Searle, 1971; Urquhart and Weeks, 1978). Quaas and Van Vleck (1980) have presented a "score free" multivariate BLUP predictor that takes into account the probabilistic aspects of the data. The method is useful for unordered categories of response and does not assume an underlying continuous distribution of liabilities as a point of departure. Under certain conditions which simplify computations but reduce the flexibility of the method, particular applications of the method of Quaas and Van Vleck (1980) yield results equivalent to BLUP of raw scores. Unfortunately, the method can be applied to one-way random models only.

Gianola (1980a,b) described a method of sire evaluation for unordered categorical data based on the logistic distribution. Transformations of counts from m categories of response are described by an $m - 1$ variate mixed model linear in the logarithms of the transformations. The residual variance of the log transformation can be obtained from asymptotic theory considerations and estimated consistently. The method is conditional on the variance-covariance matrix of random effects and, in fixed effects models, it yields BAN estimators of functions estimable in the log-linear scale. In other words, the solutions are identical to those obtained by Grizzle et al. (1969), and when m = 2, to Berkson's (1955) minimum logit chi-square. The properties of this proposed predictor are: consistency, asymptotic normality and asymptotic efficiency, i.e., it is impossible to obtain a predictor with smaller asymptotic variance.

Unfortunately, no critical comparisons of the proposed methods of sire evaluation for categorical data have been made. This is an area of considerable importance.

Literature Cited

Berger, P. J. and A. E. Freeman. 1978. Prediction of sire merit for calving difficulty. J. Dairy Sci. 61: 1146.

Berkson, J. 1955. Maximum likelihood and minimum χ^2 estimates of the logistic function. J. Amer. Statist. Assoc. 50:130.

Berkson, J. 1968. Application of minimum logit χ^2 estimate to a problem of Grizzle with a notation on the problem of no interaction. Biometrics 24: 75.

Bliss, C. I. 1934a. The method of probits. Science 79: 38.

Bliss, C. I. 1934b. The method of probits — a correction. Science 79:409.

Bock, R. D. 1975. Multivariate Statistical Methods in Behavioral Research. McGraw-Hill Book Co., New York.

Bock, R. D. and L. V. Jones. 1968. The Measurement and Prediction of Judgement and Choice. Holden-Day, San Francisco.

Bulmer, M. G. 1980. The Mathematical Theory of Quantitative Genetics. Clarendon Press, Oxford.

Carter, C. O. 1961. The inheritance of congenital pyloric stenosis. Brit. Med. Bull. 17:251.

Cochran, W. G. 1951. Improvement by means of selection. Proc. 2nd Berkeley Symp. Math. Stat. and Prob., pp 449–470.

Curnow, R. N. 1972. The multifactorial model for the inheritance of liability to disease and its implications for relatives at risk. Biometrics 28:931.

Curnow, R. N. 1974. The use of additional informa-

tion in estimating disease risks from family histories. Biometrics 30:655.
Curnow, R. N. and C. W. Dunnett. 1962. The numerical evaluation of certain multivariate normal integrals. Ann. Math. Stat. 33:571.
Curnow, R. N. and C. Smith. 1975. Multifactorial models for familial diseases in man. J. Roy. Statist. Soc., Ser. A. 138:131.
Dahlquist, G. and A. Björck. 1974. Numerical Methods. Prentice Hall, Englewood Cliffs.
Dempster, E. R. and I. M. Lerner. 1950. Heritability of threshold characters. Genetics 35:212.
Falconer, D. S. 1960. Introduction to Quantitative Genetics. Ronald, New York.
Falconer, D. S. 1965. The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann. Hum. Genet. 29:51.
Falconer, D. S. 1967. The inheritance of liability to diseases with variable age of onset with particular reference to diabetes mellitus. Ann. Hum. Genet. 31:1.
Finney, D. J. 1947. Probit Analysis: A Statistical Treatment of The Sigmoid Response Curve. University Press, Cambridge, England.
Fisher, R. A. 1938. Statistical Methods for Research Workers (7th Ed.). Oliver and Boyd, Edinburgh.
Gianola, D. 1980a. A method of sire evaluation for dichotomies. J. Anim. Sci. 51:1266.
Gianola, D. 1980b. Genetic evaluation of animals for traits with categorical responses. J. Anim. Sci. 51:1272.
Gianola, D. and H. W. Norton. 1981. Scaling threshold characters. Genetics (In Press).
Grizzle, J. E., C. F. Starmer and G. G. Koch. 1969. Analysis of categorical data by linear models. Biometrics 25:489.
Grüneberg, H. 1952. Genetical studies on the skeleton of mouse. IV. Quasi-continuous variations. J. Genet. 51:95.
Haseman, J. K. and R. C. Elston. 1972. The investigation of linkage between a qualitative trait and a marker locus. Behav. Genet. 2:3.
Hazel, L. N. 1943. The genetic basis for constructing selection indexes. Genetics 28:476.
Helwig, J. F. and K. A. Council. 1979. SAS User's Guide. SAS Institute, Inc., Raleigh.
Henderson, C. R. 1953. Estimation of variance and covariance components. Biometrics 9:226.
Henderson, C. R. 1973. Sire evaluation and genetic trends. Proc. of Anim. Breeding and Genetics Symp. in honor of Dr. J. L. Lush, Amer. Soc. of Anim. Sci. and Amer. Dairy Sci. Assoc., Champaign, IL.
Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423.
Henderson, C. R., O. Kempthorne, S. R. Searle and C. M. von Krosigk. 1959. The estimation of environmental trends from records subject to culling. Biometrics 15:192.
Hogg, R. V. and A. T. Craig. 1970. Introduction to Mathematical Statistics. McMillan, New York.
James, J. W. 1971. Frequency in relatives for an all-or-none trait. Ann. Hum. Genet. 35:47.
Kendall, M. and A. Stuart. 1977. The Advanced Theory of Statistics. Vol. 1. MacMillan, New York.
Koch, G. G., P. B. Imrey and D. W. Reinfurt. 1972. Linear model analysis of categorical data with incomplete reference vectors. Biometrics 28:663.
Landis, J. R. and G. G. Koch. 1977. A one-way components of variance model for categorical data. Biometrics 33:671.
Leonard, T. 1972. Bayesian methods for binomial data. Biometrika 59:581.
Lush, J. L. 1956. Answer to Query. Biometrics 12:84.
Mendell, N. R. and R. C. Elston. 1974. Multifactorial and qualitative traits: genetic analysis and prediction of recurrence risks. Biometrics 30:41.
Milagres, C., E. U. Dillard and O. W. Robison. 1979. Heritability estimates for some measures of reproduction in Hereford heifers. J. Anim. Sci. 49:668.
Plackett, R. L. 1965. A class of bivariate distributions. J. Amer. Statist. Assoc. 60:516.
Pollak, E. J. and A. E. Freeman. 1976. Parameter estimation and sire evaluation for dystocia and calf size in Holsteins. J. Dairy Sci. 59:1817.
Quaas, R. L. and L. D. Van Vleck. 1980. Categorical trait sire evaluation by best linear unbiased prediction of future progeny category frequencies. Biometrics 36:117.
Rasmusen, B. A. and L. L. Christian. 1976. H blood types in pigs as predictors of stress susceptibility. Science 191:947.
Rasmusen, B. A. and J. M. Lewis. 1973. The M-L blood-group system and survival of Suffolk and Targhee lambs. Anim. Blood Group and Biochem. Genet. 4:55.
Reich, T., J. W. James and C. A. Morris. 1972. The use of multiple thresholds in determining the mode of transmission of semi-continuous traits. Ann. Hum. Genet. 36:163.
Robertson, A. and I. M. Lerner. 1949. The heritability of all-or-none traits: viability of poultry. Genetics 34:395.
Rutledge, J. J. 1977. Repeatability of threshold traits. Biometrics 33:395.
Schaeffer, L. R. and J. W. Wilton. 1976. Methods of sire evaluation for calving ease. J. Dairy Sci. 59:544.
Searle, S. R. 1971. Linear Models. John Wiley and Sons, New York.
Searle, S. R. 1979. Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology. Biometrics Unit, Cornell Univ., Ithaca, NY.
Smith, C. 1970. Heritability of liability and concordance in monozygous twins. Ann. Hum. Genet. 34:85.
Smith, C. and N. R. Mendell. 1974. Recurrence risks from family history and metric traits. Ann. Hum. Genet. 37:275.
Snell, E. J. 1964. A scaling procedure for ordered categorical data. Biometrics 20:592.
Tallis, G. M. 1962. The maximum likelihood estimation of correlation from contingency tables. Biometrics 18:342.
Thompson, R. 1972. The maximum likelihood approach to the estimate of liability. Ann. Hum. Genet. 36:221.
Thompson, R. 1979. Sire evaluation. Biometrics 35:339.

Thurstone, L. L. 1927. Psychophysical analysis. Amer. J. Psychol. 38:368.

Tong, A.K.W., J. W. Wilton and L. R. Schaeffer. 1977. Application of a scoring procedure and transformations to dairy type classifications and beef ease of calving categorical data. Can. J. Anim. Sci. 57:1.

Urquhart, N. S. and D. L. Weeks. 1978. Linear models of messy data: some problems and alternatives. Biometrics 34:696.

Van Vleck, L. D. 1972. Estimation of heritability of threshold characters. J. Dairy Sci. 55: 218.

Vinson, W. E., J. M. White and R. H. Kliewer. 1976. Overall classification as a selection criterion for improving categorically scored components of type in Holsteins. J. Dairy Sci. 59:2104.

Wahrendorf, J. 1980. Inference in contingency tables with ordered categories using Plackett's coefficient of association for bivariate distributions. Biometrika 67:15.

Wright, S. 1934a. An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19:506.

Wright, S. 1934b. The results of crosses between inbred strains of guinea pigs differing in number of digits. Genetics 19:537.

Wright, S. 1943. An analysis of local variability of flower color in Linanthus parryae. Genetics 28: 139.