THE ANALYSIS OF QUANTITATIVE TRAITS FOR SIMPLE GENETIC MODELS FROM PARENTAL, F_1 AND BACKCROSS DATA

R. C. ELSTON

Department of Biostatistics, Genetics Curriculum, and the Biological Sciences Research Center, U.N.C., Chapel Hill, N. C. 27514

AND

JOHN STEWART†

Department of Genetics, Cambridge University

Manuscript received May 4, 1971
Revised copy received November 8, 1972
Transmitted by T. Prout

ABSTRACT

The following models are considered for the genetic determination of quantitative traits: segregation at one locus, at two linked loci, at any number of equal and additive unlinked loci, and at one major locus and an indefinite number of equal and additive loci. In each case an appropriate likelihood is given for data on parental, F_1 and backcross individuals, assuming that the environmental variation is normally distributed. Methods of testing and comparing the various models are presented, and methods are suggested for the simultaneous analysis of two or more traits.

METHODS to determine the number of loci involved in the genetic variation of a quantitative trait have typically involved the solving of moment or cumulant equations, assuming that all the loci have equal and additive effects (see, e.g., Student 1934; Wright 1968; and Falconer 1970). Recently Tan and Chang (1972) have shown how, with the same assumptions and for self-fertilized populations, a maximum likelihood estimate of the number of loci involved can be obtained. Anderson and Kempthorne (1954) devised general models that allow for the estimation of epistatic as well as dominance parameters, but did not consider in any detail the problem of determining which models adequately account for a set of data. Stewart (1969a,b), extending and refining the partitioning method of genetic analysis devised by Powers (1963), gave methods of testing whether one or two loci can adequately account for a given set of backcross data, and of estimating any linkage relationship among such loci. However, these methods assumed that the parental and F_1 distributions, within which the variation is entirely environmentally caused, are completely known.

* This investigation was supported by a Public Health Service Research Career Development Award (1-K3-GM-31, 732) and research grant (GM 16697), and a research grant (NGR 34-003-040) from the National Aeronautics and Space Administration.

† Current address: 1018 HSE, Cardiovascular Research Institute, University of California Medical Center, San Francisco, California 94122.

The main purpose of this paper is to consider, under these and various other simple genetic models, the simultaneous estimation of the parameters of the parental, F_1 and backcross distributions. With this information the observed and expected distributions can be compared, and so a judgment can be made as to which of the models fit the data. Only maximum likelihood estimation is considered, in view of its known superiority over moment estimation (see, e.g., KENDALL and STUART 1961), and its general optimal properties for large samples (RAO 1964). Furthermore, comparison of the likelihoods for several different models can indicate which models are equally compatible with the data, and so is more informative than a simple estimate of the number of loci involved. It should, of course, be clearly understood that data of this type can never prove that only one or two loci are involved; and this would be so even if the data measurements were qualitative and discontinuous in nature. Only further breeding tests can unequivocally distinguish between the involvement of one and more than one locus (WRIGHT 1934). It is nevertheless useful to have methods whereby the maximum amount of genetic information can be gleaned from data limited to parental, F_1 and backcross individuals, which are relatively easy to obtain, in order to decide what further breeding tests are desirable. The extension of the same methods to other types of crosses and further generations is straightforward, but will not be developed here. In an accompanying paper (STEWART and ELSTON 1973), some of the methods presented here are applied to data on physiological traits in mice.

It is assumed throughout that each sample observation is on an individual from one of five classes: the two homozygous parental strains (denoted by the subscripts 1 and 3), the F_1 (denoted by the subscript 2) and the two backcrosses (denoted by the double subscripts 12 and 32, respectively). Thus we assume there are measurements on n_1 individuals from one parental strain and on n_3 from the other, the measurements on the j-th such individuals being x_{1j} and x_{3j} respectively; there are measurements on n_{12} individuals from the backcross to the first parental strain and on n_{32} from the backcross to the other, the measurements on the j-th such individuals being x_{12j} and x_{32j} respectively; and there are measurements on n_2 individuals from the F_1, the measurement on the j-th such individual being x_{2j}. Any of the n's can be zero, but should this be the case the parameters may not all be estimable. We define the total sample size as $N = n_1+n_2+n_3+n_{12}+n_{32}$.

For each genetic model the natural logarithm of the likelihood, denoted by L, is given. Various computer methods can then be used to obtain both the maximum likelihood estimates of the parameters and their variance-covariance matrix, by a search of the log likelihood surface for local maxima and its numerical double differentiation at such maxima (see, e.g., KAPLAN and ELSTON 1972; other methods for finding maxima, or minima, have also been described by POWELL 1964; NELDER and MEAD 1965; and ROSENBROCK 1960). It is necessary to start the search of the likelihood surface at some point, and so reasonable starting values for the parameter estimates are suggested.

The maximization of the log likelihood should be performed under certain constraints; for example a variance estimate should be constrained to be positive, and a recombination fraction should normally be constrained to be between zero and one-half. Certain genetic models can be considered as special cases of more general models, and the maximum likelihood estimates for these models can be easily obtained from the same general likelihoods by imposing one or more functional relationships among the parameters of the model; these functional relationships, or "restrictions," are noted for the most frequently considered models. The maximum likelihood program devivsed by KAPLAN and ELSTON (1972) allows for such constraints and restrictions.

It is reasonable to suppose that for some suitable scale of measurement the data observations, x, are, for each given genotype, normally distributed. It is also reasonable to assume that on the same scale the environmental variances for the different genotypes are all the same: for even if, in fact, they do differ, their estimates in practice will have such large standard errors that they will not be significantly heterogeneous. In fact it has been found empirically, for example, that for many traits similar results are obtained whether the analysis is performed on the original data measurements or on their logarithms. Normality and a common variance will therefore be assumed throughout. COLLINS (1967, 1968) has developed non-parametric methods applicable to cases where the assumptions that the environmental variation is normally distributed, and equal for all groups, are seriously invalid. Even if these assumptions are violated, however, the robustness, efficiency and power of the methods given here make them far preferable to other methods.

For convenience we define

$$f_{ij}(\mu_k) = -(x_{ij}-\mu_k)^2/2\sigma^2$$

where $i = 1, 2, 3, 12$ or 32; μ_k is the mean of the distribution for the particular genotype k, and σ^2 is the common environmental variance.

2. MODELS FOR A SINGLE TRAIT

(i) *One locus:* If the two parental strains differ at one locus only, then only three genotypes are possible. Let the means of the distributions for the parental strains and the F_1 be μ_1, μ_3 and μ_2, respectively. Then the log likelihood of the n_l observations x_{1j} is simply

$$\text{constant } -n_1\, ln\sigma + \sum_{j=1}^{n_1} f_{1j}(\mu_1),$$

where the constant (in this case equal to $-n_1\, ln\sqrt{2\pi}$ can be ignored. Similarly, analogous expressions hold for the other parental and the F_1 observations. The n_{12} backcross observations are distributed as a 1:1 mixture of the two normal distributions $N(\mu_1,\sigma^2)$ and $N(\mu_2,\sigma^2)$, and hence their log likelihood is

$$\text{constant } -n_{12}\, ln\sigma + \sum_{j=1}^{n_{12}} ln(\tfrac{1}{2}e^{f_{12j}(\mu_1)}+\tfrac{1}{2}e^{f_{12j}(\mu_2)});$$

and an analogous expression holds for the n_{32} backcross observations, substituting 3 for 1. Thus, adding together the log likelihoods for the five classes, we have

$$L = K - Nln\sigma + \sum_{i=1}^{3} \sum_{j=1}^{n_i} f_{ij}(\mu_i) + \sum_{j=1}^{n_{12}} ln(\tfrac{1}{2}e^{f_{12j}(\mu_1)} + \tfrac{1}{2}e^{f_{12j}(\mu_2)}) \qquad (1)$$

$$+ \sum_{j=1}^{n_{32}} ln(\tfrac{1}{2}e^{f_{32j}(\mu_3)} + \tfrac{1}{2}e^{f_{32j}(\mu_2)}),$$

where the constant $K(= -Nln\sqrt{2\pi})$ will be kept the same throughout the rest of section 2.

Reasonable starting values for the parameters can be simply obtained by ignoring the backcross data: the three means are taken to be the sample means of the x_{1j}, x_{2j}, and x_{3j}, and σ^2 is taken as the pooled within-class sample variance for these three classes.

(ii) (2) *linked loci:* Provided the number of loci involved is no greater than two, it is not difficult to allow for linkage between the loci. Let the recombination fraction be λ. Then a backcross observation has a probability $(1-\lambda)/2$ of coming from a parental distribution, a probability $(1-\lambda)/2$ of coming from the F_1 distribution, and a probability $\lambda/2$ of coming from each of the two relevant recombinant distributions. Let μ_{12} and μ_{21} now be the means of the two recombinant distributions when the backcross is to parent 1, and μ_{32} and μ_{23} be the means when the backcross is to parent 3. The log likelihood of the whole sample is, then, analogous to (1)

$$L = K - Nln\sigma + \sum_{i=1}^{3} \sum_{j=1}^{n_i} f_{ij}(\mu_i)$$

$$+ \sum_{j=1}^{n_{12}} ln\{(1-\lambda)(e^{f_{12j}(\mu_1)} + e^{f_{12j}(\mu_2)}) + \lambda(e^{f_{12j}(\mu_{12})} + e^{f_{12j}(\mu_{12})})\}$$

$$+ \sum_{j=1}^{n_{32}} ln\{(1-\lambda)(e^{f_{32j}(\mu_3)} + e^{f_{32j}(\mu_2)}) + \lambda(e^{f_{32j}(\mu_{32})} + e^{f_{32j}(\mu_{23})})\}$$

$$-(n_{12} + n_{32})ln2. \qquad (2)$$

The last term in this expression is, of course, a constant, but must be inserted if the values of L—K for the various models are to be comparable.

Starting values of μ_1, μ_2, μ_3, and σ^2 can be the same as before, and a starting value for λ may be arbitrarily taken as 0.25. For the means of the recombinant distributions two sets of starting values should be tried, corresponding to the recombinant means lying between or outside the means of the parental and F_1 distributions:

$$\begin{array}{lcl} \mu_{12} = (2\mu_1 + \mu_2)/3 & & \mu_{12} = 2\mu_1 - \mu_2 \\ \mu_{21} = (\mu_1 + 2\mu_2)/3 & & \mu_{21} = 2\mu_2 - \mu_1 \\ & \text{or} & \\ \mu_{32} = (2\mu_3 + \mu_2)/3 & & \mu_{32} = 2\mu_3 - \mu_2 \\ \mu_{23} = (\mu_3 + 2\mu_2)/3 & & \mu_{23} = 2\mu_2 - \mu_3 \end{array}$$

Provided both these sets of starting values lead to the same local maximum on the likelihood surface, one can be reasonably assured that the maximum is unique.

To fit the model of two linked loci with equal and additive effects we take the likelihood (2) together with the four restrictions

$$\mu_{12}=\mu_{21}=(\mu_1+\mu_2)/2, \quad \mu_{32}=\mu_{23}=(\mu_3+\mu_2)/2. \tag{3}$$

The starting values for all the means in this model are determined by μ_1, μ_2, and μ_3. Similarly we can fit the model of two linked loci with equal and additive genes at each locus by taking (2) together with (3) and the restriction

$$\mu_2 = (\mu_1 + \mu_3)/2, \tag{4}$$

the starting values for all the means now being determined by μ_1 and μ_3.

A model in which equal and additive effects are assumed, either for the two loci or for all four genes involved, may well be unrealistic and too restrictive. On the other hand, maximization of (2) without any restriction whatsoever can lead to meaningless results; for there is then a tendency for the estimates of μ_{12}, μ_{21}, μ_{32}, and μ_{23} to coincide with any outlying observations that are present. A possible compromise, which might approximate reality in many cases, is to use the likelihood (2) together with either of the two "symmetry" restrictions

$$(\mu_{12}-\mu_{21})/(\mu_1-\mu_2) = (\mu_{32}-\mu_{23})/(\mu_3-\mu_2), \tag{5a}$$
$$(\mu_{12}-\mu_{21})^2-(\mu_1-\mu_2)^2 = (\mu_{32}-\mu_{23})^2-(\mu_3-\mu_2)^2. \tag{5b}$$

These are much milder restrictions than (3), but can only be used when data on all five classes of individuals are available. Another possibility, which does not need data on both backcrosses, is to assume that the effects of the two loci are additive but not necessarily equal; this is given by the restrictions

$$\mu_1 + \mu_2 = \mu_{12} + \mu_{21}, \quad \mu_3 + \mu_2 = \mu_{32} + \mu_{23}. \tag{6}$$

Any combination of (4), (5), and (6) can of course be used together, if the particular situation warrants it. If (6) is used, we can represent the genotypic effects in terms of an overall mean m, additive effects a_A and a_B, and dominance effects d_A and d_B, as shown in Table 1. Furthermore, Table 2 shows which genotypes the various μ's correspond to; this depends upon whether, in the parental strains, the two loci are in coupling (AABB and aabb) or repulsion (AAbb and aaBB). From Tables 1 and 2 we can derive the meaning of restrictions (4) and (5), provided they are used in conjunction with (6). It is immediately apparent that (4) is then the same as

$$d_A + d_B = 0,$$

i.e., the average dominance effect of the two loci is zero. Similarly it is found that

$$\frac{d_A}{a_A} = \frac{d_B}{a_B},$$

i.e., the two loci have the same dominance ratio, if, in conjunction with (6), either (5a) is used when the loci are in coupling in the parental strains, or (5b) is used when the loci are in repulsion in the parental strains. If the loci are in coupling, the recombinant means lie between the parental and F_1 means; if in repulsion, the recombinant means lie outside the parental and F_1 means. If (5a)

TABLE 1

Genotypic means due to two loci (A,a and B,b) with additive effects

Locus 2	Locus 1 AA	 Aa	 aa
BB	$m+a_A+a_B$	$m+a_B+d_A$	$m-a_A+a_B$
Bb	$m+a_A+d_B$	$m+d_A+d_B$	$m-a_A+d_B$
bb	$m+a_A-a_B$	$m-a_B+d_A$	$m-a_A-a_B$

TABLE 2

Genotypic means, μ, on the assumption of two loci

Parental strains in coupling				Parental strains in repulsion			
	Locus 1				Locus 1		
Locus 2	AA	Aa	aa	Locus 2	AA	Aa	aa
BB	μ_1	μ_{12}		BB		μ_{21}	μ_1
Bb	μ_{21}	μ_2	μ_{23}	Bb	μ_{23}	μ_2	μ_{21}
bb		μ_{32}	μ_3	bb	μ_3	μ_{32}	

or (5b) is used without (6), the genetic meaning is not so clear; the implied symmetry has, however, intuitive appeal, and in practice these restrictions are found to be useful.

(iii) *Equal and additive unlinked loci:* It is impractical to consider the generalization of (2) to more than two loci, in view of the large number of unknown parameters that would be involved. One way to keep down the number of parameters that need to be estimated is to assume that all the loci are unlinked and have equal and additive effects. This model is quite restrictive, and can only be an approximation to any real situation; nevertheless it is more general than any model examined so far, and so will be considered here in detail.

Suppose the parental strains differ at l equal and additive unlinked loci, in each parent some loci acting in one direction and the remaining ones in the opposite direction. In particular, let $m(<l-m)$ be the smaller number all acting in the same direction. (The case $m=l-m$ is of no practical interest, since it implies $\mu_1=\mu_3$). Thus, since the effects of the loci are equal and additive, $\mu_3-\mu_1$ is the sum of $l-2m$ such effects. It follows that in a backcross individual a locus homozygous as one of the m loci in parent 1 (or as one of the $l-m$ loci in parent 3) contributes

$$-\{m\mu_1-(l-m)\mu_3\}/l(l-2m);$$

a locus homozygous as one of the $l-m$ loci in parent 1 (or as one of the m loci in parent 3) contributes

$$\{(l-m)\mu_1-m\mu_3\}/l(l-2m);$$

and a heterozygous locus contributes μ_2/l, towards the mean of its distribution. (Note that when m=0 the homozygous loci each contribute μ_3/l and μ_1/l, respectively, as we should expect). If one of the n_{12} backcross observations is homozygous at h of the m loci and k of the $l-m$ loci in parent 1, then it comes from a normal distribution with mean

$$\frac{1}{l(l-2m)}[-h\{m\mu_1-(l-m)\mu_3\}+k\{(l-m)\mu_1-m\mu_3\}]+ \\ (l-h-k)\mu_2/l=\mu_{1hk},\text{ say,} \qquad (7)$$

$$h=0,\ 1,\ \ldots.,\ m;\ k=0,\ 1,\ \ldots.,\ l-m.$$

Thus the n_{12} backcross observations x_{12} are distributed as a mixture of $(m+1)$ $(l-\mathrm{m}+1)$ normal distributions: a fraction

$$\binom{m}{h}\binom{l-m}{k}\Big/2^l \text{ is } N(\mu_{1hk},\sigma^2).$$

Interchanging the subscripts 1 and 3 gives the distribution for the n_{32} observations x_{32}, and so for these a fraction

$$\binom{m}{h}\binom{l-m}{k}\Big/2^l \text{ is } N(\mu_{3hk},\sigma^2),$$

where μ_{3hk} is (7) with the subscripts 1 and 3 interchanged. The log likelihood of all the observations under this model thus becomes

$$L=K-N\,ln\sigma+\sum_{i=1}^{3}\sum_{j=1}^{n_i}f_{ij}(\mu_i)+\sum_{j=1}^{n_{12}}ln\left(\sum_{h=0}^{m}\sum_{k=0}^{l-m}\frac{1}{2^l}\binom{m}{h}\binom{l-m}{k}e^{f_{12j}(\mu_{1hk})}\right)+ \\ \sum_{j=1}^{n_{32}}ln\left(\sum_{h=0}^{m}\sum_{k=0}^{l-m}\frac{1}{2^l}\binom{m}{h}\binom{l-m}{k}e^{f_{32j}(\mu_{3hk})}\right). \qquad (8)$$

When $l=1$ we must have $m=0$, and then (8) reduces to (1), as it should. Whatever values are taken for l and m, the same starting values, as given in section (i) for (1), are reasonable.

A special case that may sometimes be of interest is that of equal and additive genes at each of the l loci, i.e., the same model as we have just considered, but with no dominance. The appropriate log likelihood is then also given by (8) provided we add the restriction (4). In this case the number of independent parameters to estimate is reduced by one, and the starting value for μ_2 is determined by the starting values for μ_1 and μ_3.

Another special case is what happens as l and m become large. Rewriting (7)

$$\mu_{1hk}=\mu_2-\frac{h}{l(l-2m)}[m\mu_1+(l-2m)\mu_2-(l-m)\mu_3]+ \\ \frac{k}{l(l-2m)}[(l-m)\mu_1-(l-2m)\mu_2-m\mu_3]\ ;$$

and utilizing the fact that as l and m become large we have, approximately,

$$h \text{ is } N(m/2,\ m/4) \\ \text{and } k \text{ is } N((l-m)/2,\ (l-m)/4),$$

we see that the distribution of the backcross observations x_{12} becomes approximately normal with mean

$$\mu_{1\infty}=\mu_2-\frac{m}{2l(l-2m)}[m\mu_1+(l-2m)\mu_2-(l-m)\mu_3]+$$

$$\frac{l-m}{2l(l-2m)}\left[(l-m)\mu_1-(l-2m(\mu_2-m\mu_3]=(\mu_1+\mu_2)/2\right.$$

and variance

$$\sigma^2_\infty=\sigma^2+\frac{m}{4l^2(l-2m)^2}[m\mu_1+(l-2m)\mu_2-(l-m)\mu_3]^2+$$
$$\frac{l-m}{4l^2(l-2m)^2}[(l-m)\mu_1-(l-2m)\mu_2-m\mu_3]^2. \qquad (9)$$

Now, in the limit as l and m tend to infinity, provided m/l remains at some constant value less than one-half, $\sigma^2{}_\infty$ tends to σ^2. Thus, analogously defining $\mu_{3\infty}=(\mu_3+\mu_2)/2$, the log likelihood for this model is

$$L=K-N\,ln\sigma+\sum_{i=1}^{3}\sum_{j=1}^{n_i}f_{ij}(\mu_i)+\sum_{j=1}^{n_{12}}f_{12j}(\mu_{1\infty})+\sum_{j=1}^{n_{32}}f_{32j}(\mu_{3\infty})\;. \qquad (10)$$

It is interesting to note that this is identical with the result obtained if we impose on (2) the restrictions

$$\lambda=1,\ \mu_{12}=\mu_{21}=(\mu_1+\mu_2)/2,\ \mu_{32}=\mu_{23}=(\mu_3+\mu_2)/2;$$

not only can this be more convenient for programming, but it also shows that, from a statistical point of view, this model is a special case (involving the one restriction $\lambda=1$) of the model implied by (2) and (3) taken together. As will be seen later, this fact can be utilized if we wish to test for a significant difference between the two models. As before, we can also add restriction (4) for the case of equal and additive genes.

Now the model implied by (10) will in practice often be unsatisfactory as an approximation to what happens as l and m become large, since it assumes that the variance in the backcross observations is the same as that in the parental observations. For this reason it is of interest to consider the limiting situation in which the last two terms of (9) are neither zero nor infinite. This will occur if l and m tend to infinity in such a way that $(l-2m)^2/l$ tends to a constant, C, say. These same conditions also imply m/l tends to a half, and in the limit

$$\sigma_\infty=\sigma^2+(\mu_1-\mu_3)^2/16C. \qquad (11)$$

Thus a log likelihood that is appropriate for large l and m, provided m/l is not too far from one-half, is again given by (10) if we redefine

$$f_{12j}(\mu_{1\infty})=-(x_{12j}-\mu_{1\infty})^2/2\sigma^2_\infty\,,$$
$$f_{32j}(\mu_{3\infty})=-(x_{32j}-\mu_{3\infty})^2/2\sigma^2_\infty\,. \qquad (12)$$

The net effect of this is to allow the common backcross variance, $\sigma^2{}_\infty$, to differ from the variance in the parental and F_1 strains; but if $\sigma^2{}_\infty$ is estimated to be smaller than σ^2, (10) without the redefinition (12) is more appropriate.

(iv) *One major locus and an infinite number of equal and additive loci:* If we now suppose there is also, acting additively to these l loci, one locus with a major effect, then the individuals in a given backcross will be either homozygous or heterozygous at the major locus. It follows that each backcross will be a 1:1 mixture of two distributions, each with variance σ^2_∞. If we let the means of these two distributions be μ_{i2} and μ_{2i} for the n_{i2} backcross individuals, i=1,3, then the log likelihood of the whole sample is

$$L = K - N\,ln\sigma + \sum_{i=1}^{3}\sum_{j=1}^{n_i} f_{ij}(\mu_i) + \sum_{j=1}^{n_{12}} \ln(½e^{f_{12j}(\mu_{12})} + ½e^{f_{12j}(\mu_{21})}) + \sum_{j=1}^{n_{13}} ln(½e^{f_{32j}(\mu_{32})} + ½e^{f_{32j}(\mu_{23})}) \,, \qquad (13)$$

provided we use the definitions (12). In the special case that $\sigma^2_\infty = \sigma^2$, this is exactly the same as (2) with $\lambda = 1$, and so here again, for programming and from a testing point of view, it can be useful to consider this log likelihood as a special case of (2).

As with the case of two loci, (13) can be used with any appropriate combination of (4), (5) and (6). Provided (6) can be used, (4) implies the average dominance effect of all loci is zero and (5a) implies that the dominance ratio for the major locus is the same as that for the "infinite" number of equal and additive loci.

3. MODELS FOR TWO TRAITS

If for each of two traits no more than two loci need be postulated to account for the observed genetic variation, then it may be possible to elucidate the linkage relationships among the loci involved; if more than two loci need be postulated it is doubtful whether further numerical analysis, in the absence of more extentive experimental data, will be very helpful. For this reason this section will consider only the model in which two linked loci are assumed for each of the two traits. This, however, will include as special cases models in which one or both of the traits are due to just one locus (by setting the appropriate recombination fraction(s) equal to zero), and models in which one or both of the traits are due to two unlinked loci (by setting the appropriate recombination fraction(s) equal to 0.5).

In general, then, we can suppose the two traits are x and y, and that there are 18 corresponding parameters; it will be helpful for the sequel to denote the parental and F$_1$ means by the double subscripts 11, 33, and 22, rather than just single subscripts, and so the 18 parameters are:

$$\mu_{x11},\ \mu_{x22},\ \mu_{x33},\ \mu_{x12},\ \mu_{x21},\ \mu_{x32},\ \mu_{x23},\ \sigma^2{}_x,\ \lambda_x \text{ for } x,$$

and

$$\mu_{y11},\ \mu_{y22},\ \mu_{y33},\ \mu_{y12},\ \mu_{y21},\ \mu_{y32},\ \mu_{y23},\ \sigma^2{}_y,\ \lambda_y \text{ for } y.$$

If λ_x=0, then μ_{x12}, μ_{x12}, μ_{x32} and μ_{x23} are non-existent; and analogously if λ_y=0. There are now up to four other parameters that we are interested in, namely each λ_{ij} (i,j=1,2), the recombination fraction between the i-th locus for x and the j-th locus for y. If λ_x orλ_y is zero there are only two such recombination fractions, and if both λ_x and λ_y are zero there is only one.

We now have five sets of vector observations (x,y). If, for each genotype, x and y are uncorrelated then the log likelihood for the n_1 vectors $(x_{1j}\ y_{1j})$, the n_3 vectors (x_{3j}, y_{3j}) and the n_2 vectors (x_{2j}, y_{2j}) is simply the sum of the corresponding log likelihoods for x and y separately; more generally, however, the correlation must be taken into account, as will now be indicated. Specifically, we shall assume that, for each genotype, x and y follow a bivariate normal distribution, and that only the means of this distribution change with genotype, the covariance matrix being the same for all genotypes. Thus we need only introduce one further parameter, the common environmental correlation ρ, to completely specify all of these bivariate distributions. Analogous to the univariate case, we define

$$f_{ij}(\mu_{xg},\mu_{yh}) = -\left\{\frac{(x_{ij}-\mu_{xg})^2}{\sigma_x^2} - \frac{2\rho(x_{ij}-\mu_{xg})(y_{ij}-\mu_{yh})}{\sigma_x\sigma_y} + \frac{(y_{ij}-\mu_{yh})^2}{\sigma_y^2}\right\} / 2(1-\rho^2).$$

It follows immediately that the log likelihood for the two parental and F_1 observations is

$$\text{constant} - (n_1+n_2+n_3)\ ln[\sqrt{(1-\rho^2)}\sigma_x\sigma_y] + \sum_{i=1}^{3}\sum_{j=1}^{n_i} f_{ij}(\mu_{xii},\mu_{yii}). \qquad (14)$$

Each set of backcross observations is distributed as a mixture of 16 bivariate normal distributions; for there are four loci involved, and at each locus an individual may be homozygous or heterozygous. In order to develop the appropriate likelihood, assume for the moment that crossing over between any pair of loci is independent of crossing over between any other pair of loci (even though there may be one locus in common). Let

$$\gamma_{ij}(\lambda) = \begin{cases} 1-\lambda & \text{if } i=j \\ \lambda & \text{if } i\neq j\ . \end{cases}$$

The log likelihood for the n_{12} observations (x_{12j}, y_{12j}) is then

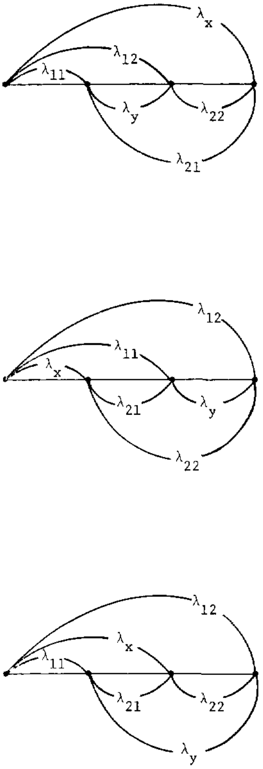
$$\text{constant} - n_{12}\ ln[\sqrt{(1-\rho^2)}\sigma_x\sigma_y]$$

$$+\sum_{j=1}^{n_{12}} ln\left\{\sum_{f=1}^{2}\sum_{g=1}^{2}\sum_{h=1}^{2}\sum_{k=1}^{2}\gamma_{fg}(\lambda_x)\gamma_{hk}(\lambda_y)\gamma_{fh}(\lambda_{11})\gamma_{fk}(\lambda_{12})\gamma_{gh}(\lambda_{21})\gamma_{gk}(\lambda_{22})e^{f_{12j}(\mu_{xfg},\mu_{yhk})}\right\}$$

$$-\, n_{12}\ ln\left\{\sum_{f=1}^{2}\sum_{g=1}^{2}\sum_{h=1}^{2}\sum_{k=1}^{2}\gamma_{fg}(\lambda_x)\gamma_{hk}(\lambda_y)\gamma_{fh}(\lambda_{11})\gamma_{fk}(\lambda_{12})\gamma_{gh}(\lambda_{21})\gamma_{gk}(\lambda_{22})\right\}. \qquad (15)$$

The expression for the n_{32} observations (x_{32j}, y_{32j}) is analogous, 3 replacing 1 appropriately; thus the complete log likelihood is given by the sum of (14), (15), and the corresponding expression with 3 replacing 1.

Now if the crossover frequencies between pairs of loci are not independent, then there are functional relations that hold among the recombination fractions. Rather than trying to develop a likelihood that incorporates these relations directly, it is simpler to obtain maximum likelihood estimates by maximizing the likelihood just given, but under the restrictions implied by these functional relations; this will lead to exactly the same results. Unfortunately, however, there is no certainty as to what functional relations are most appropriate, for two reasons: first, the type of dependency among the crossover frequencies is unknown; and



$$\begin{cases} \lambda_{12} = \lambda_{11} + \lambda_y - 2\lambda_{11}\lambda_y \\ \lambda_{21} = \lambda_y + \lambda_{22} - 2\lambda_y\lambda_{22} \\ \lambda_x = \lambda_{11} + \lambda_y + \lambda_{22} \\ \qquad - 2(\lambda_{11}\lambda_y + \lambda_{11}\lambda_{22} + \lambda_y\lambda_{22}) \\ \qquad + 4\lambda_{11}\lambda_y\lambda_{22} \end{cases}$$

$$\begin{cases} \lambda_{11} = \lambda_x + \lambda_{21} - 2\lambda_x\lambda_{21} \\ \lambda_{22} = \lambda_{21} + \lambda_y - 2\lambda_{21}\lambda_y \\ \lambda_{12} = \lambda_x + \lambda_{21} + \lambda_y \\ \qquad - 2(\lambda_x\lambda_{21} + \lambda_x\lambda_y + \lambda_{21}\lambda_y) \\ \qquad + 4\lambda_x\lambda_{21}\lambda_y \end{cases}$$

$$\begin{cases} \lambda_x = \lambda_{11} + \lambda_{21} - 2\lambda_{11}\lambda_{21} \\ \lambda_y = \lambda_{21} + \lambda_{22} - 2\lambda_{21}\lambda_{22} \\ \lambda_{12} = \lambda_{11} + \lambda_{21} + \lambda_{22} \\ \qquad - 2(\lambda_{11}\lambda_{21} + \lambda_{11}\lambda_{22} + \lambda_{21}\lambda_{22}) \\ \qquad + 4\lambda_{11}\lambda_{21}\lambda_{22} \end{cases}$$

FIGURE 1.—Basic set of three orders for four loci, and the functional relations among the six recombination fractions implied by TROW's formula, if all four loci are linked.

second, the order of the loci is unknown. Although other formulae have been suggested, as a first approximation TROW's (1913) formula is probably the best to use to describe the dependency among crossover frequencies for loci on the same chromosome; this is equivalent to assuming a lack of interference. There are twelve different orders possible for the sequence of four loci (24 if a distinction is made between the start and finish of a sequence), and a basic set of three orders is shown in Figure 1. All the other orders can be derived from these by interchanging λ_x and λ_y, λ_{1j} and λ_{2j} and/or λ_{i1} and λ_{i2}, as necessary. There are only three functionally independent distances, and so, since six recombination fractions are involved, we must have three independent restrictions. The restrictions that follow from the use of TROW's formulae are given in the figure for each of the three basic cases. Which order to assume, and hence which set of restrictions to use, is only a problem when at least three of the loci lie on the same chromosome. It is suggested that in the first place the six recombination fractions should be estimated without any restrictions at all, and the resulting estimates, though incorrect, will be adequate to eliminate all but a few possible orders; these can be tried in turn, and the order that results in the largest log likelihood at its maximum chosen.

Before any pair of traits is analyzed in the manner indicated in this section, each trait should have been analyzed separately as indicated in the previous section. Any restrictions on the means that have been found appropriate when the traits are analyzed separately should be kept when the traits are analyzed together. In fact, to avoid the excessive amount of computational time that may be needed for the simultaneous estimation of twenty or more parameters, the likelihood derived in this section can be used to estimate only the recombination fractions and the environmental correlation (a maximum of four independent parameters); the means and variances may be reasonably fixed equal to the estimates that are obtained from the separate single trait analyses. Terms in σ_x and σ_y are then considered as constant, and so the whole likelihood can be simply expressed as

$$L = \text{constant} - \frac{N}{2} ln\ (1-\rho^2) + \sum_{i=1}^{3} \sum_{j=1}^{h_i} f_{ij}(\mu_{xii},\mu_{yii}) + \sum_{i=1,3} \sum_{j=1}^{n_{i2}}$$

$$ln\left\{\sum_{f=i,2} \sum_{g=i,2} \sum_{h=i,2} \sum_{k=i,2} \gamma_{fg}(\lambda_x)\gamma_{hk}(\lambda_y)\gamma_{fh}(\lambda_{11})\gamma_{fk}(\lambda_{12})\gamma_{gh}(\lambda_{21})\gamma_{gk}(\lambda_{22})e^{f_{12j}(\mu_{xfg},\mu_{yhk})}\right\}$$

$$-(n_{12}+n_{22})\ ln\left\{\sum_{f=i,2} \sum_{g=i,2} \sum_{h=i,2} \sum_{k=i,2} \gamma_{fg}(\lambda_x)\gamma_{hk}(\lambda_y)\gamma_{fh}(\lambda_{11})\gamma_{fk}(\lambda_{12})\gamma_{gh}(\lambda_{21})\gamma_{gk}(\lambda_{22})\right\} \quad (16)$$

4. SIMPLE MULTIVARIATE MODELS

In this section we describe briefly just two ways in which all the models given in section 2 can be adapted for the multivariate examination of p traits. Each observation x_{ij} is now replaced by the $p \times 1$ vector observation $\boldsymbol{x}_{ij}$, which will be assumed to follow a p-variate normal distribution.

The first, and simplest, method is to consider the linear function $\boldsymbol{a}'\boldsymbol{x}_{ij}$ as a new trait, and replace this for x_{ij} in all the log likelihoods given so far. The coefficient vector $\boldsymbol{a}$ may be chosen *a priori* on the basis of what may be biologically meaningful. For example, if the p measures in $\boldsymbol{x}_{ij}$ are all measurements of the same character, but at p different stages in the individual's development, then the coefficient $\boldsymbol{a}$ could be chosen to give the linear (or quadratic, etc.) change in the character with time. On the other hand, $\boldsymbol{a}$ may be left arbitrary in the expression of the log likelihood, its elements being estimated simultaneously with the other parameters of the model. In this way we can estimate that linear function of the p traits that best fits a one-locus model, in which case we would use (1) to estimate μ_1, μ_2, μ_3, σ^2 and $\boldsymbol{a}$ simultaneously; whereas perhaps none of the p measures taken individually fit a one-locus model, a linear function of them may well be found to do so. (Indeed, WEBER (1959) has already demonstrated how, in *Lycopersicum esculentum* Mill, a discriminant that utilizes the quotient of length by width of the cotyledon and the area of the cotelydon can accurately detect the segregation of a major gene.) Since any multiple of $\boldsymbol{a}$ gives the same linear function, apart from a scale factor, only p-1 independent parameters can be estimated; it is convenient to let the sum of the squared elements of $\boldsymbol{a}$ equal unity, but other restrictions on the elements of $\boldsymbol{a}$ are possible. Any linear functions $\boldsymbol{a}'\boldsymbol{x}_{ij}$ can of course be used as one of the two traits in the models discussed in section 3.

The second method for the multivariate examination of the p traits $\boldsymbol{x}_{ij}$ considers a model in which it is the same loci that govern all traits. This is thus a model for pleiotropic action, but it must be remembered that tight linkage would in practice be indistinguishable from this. Let the variance-covariance matrix of $\boldsymbol{x}_{ij}$, for each given genotype, be Σ; also, analogous to the univariate case, let

$$f_{ij}(\mu_k) = -\tfrac{1}{2}(\boldsymbol{x}_{ij}-\mu_k)'\,\Sigma^{-1}(\boldsymbol{x}_{ij}-\mu_k), \qquad (17)$$

where now μ_k is the $p \times 1$ vector mean of the distribution for a particular genotype. Then the multivariate analog of each of the univariate models considered in section 2 is given by substituting $|\Sigma|^{1/2}$ and f_{ij} (μ_k) for σ and f_{ij} (μ_k) respectively; and the appropriate log likelihoods are given, with this substitution, by (1), (2), (8), (10) and (13).

This multivariate analysis will usually be performed after the traits have been analyzed separately and in pairs, using the methods in sections 2 and 3. Then, if several traits seem to have the same underlying genetic mechanism, this multivariate analysis can be performed to obtain estimates on the assumption that it is the same loci that are involved for all traits. Two traits that enter such an analysis must necessarily have approximately the same correlation in the parental and backcross classes; and analysis by the methods of section 3 should indicate no significant departure from the case $\lambda_{11} = 0 = \lambda_{22}$ and $\lambda_{12} = \lambda_{21} = \lambda_x = \lambda_y$. For each trait, that restriction on the means that has been found to be most appropriate in the univariate analysis is automatically retained in the multivariate analysis: it is not necessary to have the same restriction for all p traits. Thus the univariate analyses supply all the starting estimates required for the multivariate analysis, except for the off-diagonal elements of Σ; starting estimates for these can be obtained as the pooled within groups covariances from the two parental classes and the F_1 alone, since these do not involve mixtures of distributions. In fact it is suggested, in order to keep down the number of parameters that need be estimated, that the (i,j)-th element of Σ be set equal to $r_{ij}\sigma_i\sigma_j$, where r_{ij} is the pooled sample correlation within the parental classes and the F_1; σ_i and σ_j are the standard deviations of the i-th and j-th traits respectively, estimated jointly with the means and the recombination fraction. This procedure reduces the number of parameters that need to be estimated iteratively by $p(p\text{-}1)/2$.

5. TESTING MODEL FIT

Finally, we discuss some methods of determining which models can, and which cannot, be excluded on the basis of a given set of data of the type we have been discussing. We restrict our attention here to a univariate trait, though some of the methods could just as easily be extended to the multivariate situation.

A simple pictorial method is to plot, separately for each of the five classes of individuals, the empirical distribution together with one or more theoretical distributions, the latter being represented by the appropriate model with maximum likelihood estimates in place of the unknown parameters. Cumulative plots should be used, so that each theoretical distribution is either a cumulative normal or a mixture of cumulative normal distributions. The empirical distribution is plotted as a series of n points, one for each of the n individuals in the class; in each case the ordinate is $r(x)/(n+1)$, where $r(x)$ is the rank of the individual with measurement x (the individual with smallest x has rank 1, the one with the largest x has rank n). By a comparison of the plots a rough idea of how well each model fits the data points is obtained at a glance.

A simple quantitative measure of how the different models compare with each other, but not of how they compare with the data points, is obtained by comparing the log likelihoods for the various models: a difference of D in the log likelihood indicates that under one model the data are (antilog D)-fold more likely than under another. Furthermore if the two models being compared are such that one is a special case of the other, involving d independent restrictions, we can use the likelihood ratio criterion to test whether the more general model fits significantly better than the restricted model. The test statistic to take is twice the difference between the two corresponding log likelihoods, and this should be compared with a chi-square distribution with d degrees of freedom (KENDALL and STUART 1961). In this way, for example, a chi-square with one degree of freedom can be obtained to determine if restriction (4) leads to a significantly worse fit, or if the model implied by (10) is significantly worse than two linked loci with equal and additive effects.

In order to assess how well a set of data fit any given model, we recommend the use of four different statistics. These were chosen from among over twenty test statistics, including all those discussed by PYKE (1965), on the basis of how they performed empirically on a fair-sized body of data. Some of the test statistics never gave rise to significant results whatever genetic model was assumed, and so lacked power; others always gave rise to significant results, due to the fact that an excessively large sample size is necessary before the assumed asymptotic sampling distribution of the test statistic is approximated. The following four statistics, however, were both robust and powerful against some of the alternative hypotheses of interest, and so can be recommended to judge whether, on the basis of the data available, a given genetic model is acceptable or not. Two of the test statistics are based directly on NEYMAN's smooth test (NEYMAN 1937; results for small sample sizes are given by BARTON 1953a and 1953b), one is based indirectly on NEYMAN's smooth test, and the last is based on the modified mean test proposed by LEWIS (1965).

Let the cumulative distribution for a particular class of individuals, assuming as a null hypothesis the genetic model that we wish to test, be $F(x)$; for the parental classes this will be a cumulative normal distribution, and for the backcross classes this will be a mixture of such distributions. Then under the null hypothesis $F(x_j)$, where x_j is the j-th observed measurement for a particular class of individuals, will be uniformly distributed on the unit interval. We therefore calculate $F(x_j)$ for each observed measurement in the class, assuming the unknown parameters in F are equal to their maximum likelihood estimates, and (ignoring the fact that the parameters have been estimated) test whether the resulting quantities could in fact be a sample of independently and uniformly distributed random variables. In each of the four tests this is done by means of a different chi-square statistic with one degree of freedom, as will be explained below; this statistic is then summed over the five classes (or fewer classes, if some of the n's are zero) to yield a chi-square with five (or fewer) degrees of freedom. Provided the number of parameters estimated is small compared with the total sample size N, and provided no class (unless it is non-existent) has fewer than five individuals in it, the assumption of a chi-square distribution under the null hypothesis will be accurate enough for all practical purposes.

The first chi-square statistic with one degree of freedom tests whether the mean value of $F(x_j)$ for each class is equal to one-half. If n measures are observed in the class, the statistic is

$$u_1^2 = 12\left[\sum_{j=1}^{n} \left(F(x_j) - \tfrac{1}{2}\right)\right]^2 / n. \tag{18}$$

The second of the chi-square statistics with one degree of freedom tests whether the variance of $F(x_j)$ for each class is equal to $1/12$. It is thus

$$u_2^2 = 180\left[\sum_{j=1}^{n} \left(F(x_j) - \tfrac{1}{2}\right)^2 - n/12\right]^2 / n. \tag{19}$$

The third and fourth tests consider the spacings between the $F(x_j)$. Let $x_{(j)}$ denote the ranked observations, so that

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Then the $n+1$ spacings are given by

$$D_j = F(x_{(j)}) - F(x_{(j-1)}),\ j = 1, 2, \ldots, n+1,$$

where we define $F(x_{(0)}) = 0$ and $F(x_{(n+1)}) = 1$. Under the null hypothesis $1-(1-D_j)^n$ is uniformly distributed on the unit interval, and the third chi-square statistic tests whether the variance of this for each class is equal to one-twelfth (ignoring the small correlation among these $n+1$ quantities). Thus analogous to (19), the statistic is

$$u_2'^2 = 180\left[\sum_{j=1}^{n+1} \left(\tfrac{1}{2}-(1-D_j)^n\right)^2 - (n+1)/12\right]^2 / (n+1). \tag{20}$$

The last statistic is based on the statistic S' proposed by LEWIS (1965). Denoting the ranked spacings by $D_{(j)}$, so that

$$D_{(1)} \leq D_{(2)} \leq \ldots \leq D_{(n+1)},$$

the chi-square statistic with one degree of freedom is

$$l^2 = 144\left[2(n+1) - 2\sum_{i=1}^{n+1} i\,D_{(i)} - \frac{n}{2}\right]^2 / n^2. \tag{21}$$

As explained above, in each case the statistic (18), (19), (20) or (21) is calculater for each class, and the resulting sums over all classes, which we can conveniently denote by $U^2{}_1$, $U^2{}_2$, $U'{}_2{}^2$, or L^2 respectively, are compared with the chi-square distribution with five (or possibly fewer) degrees of freedom. It is of course never possible to prove a null hypothesis; but if, for a particular genetic model, none of these four statistics is significant, we can be reasonably sure that the data adequately fit the model in question. On the other hand if any one of the four statistics shows significance, the model is thrown into doubt and the data should be examined to determine the reason for the significance. It should be noted that these four statistics between them will detect many kinds of departure from the null hypothesis, including departure from a normal distribution with constant environmental variance for each genotype. Unequal variances in the parental distributions will often cause $U^2{}_2$ to be significant, and non-normality (especially tied values, which occur with probability zero in truly normal distributions) tend to cause $U'_2{}^2$ and L^2 to be significant. Thus by examining the original data, and by comparing the parental class means and variances with the maximum likelihood estimates, it is possible to distinguish whether a test statistic is significant because of the genetic or the environmental part of the assumed model. If the latter, then the genetic model may nevertheless be acceptable.

Finally we wish to note the empirical finding that even though the environmental part of the model may be seriously violated, and this detected by one or more of the test statistics, the maximum likelihood procedure described in this paper (which assumes normality and a common variance) is nevertheless very robust. A model which *genetically* fits the data always leads to a greater likelihood than one that does not; and the maximum likelihood estimates of the class means and common variance so obtained are always in good agreement with the empirical class means and average parental variance. This fact, which is illustrated in an accompanying paper (STEWART and ELSTON 1973), lends support to the general utility of the methods presented here.

LITERATURE CITED

ANDERSON, V. L. and O. KEMPTHORNE, 1954 A model for the study of quantitative inheritance. Genetics **39**: 883–888.

BARTON, D. E., 1953a On Neyman's smooth test of goodness of fit and its power with respect to a particular system of alternatives. Skand. Aktuar. **36**: 24–63. ——, 1953b The probability distribution function of a sum of squares. Trabajos de Estadistica **4**: 199–207.

COLLINS, R. L. 1967 A general non-parametric theory of genetic analysis I. Application to the classical cross. Genetics **56**: 551. ——, 1968 A general non-parametric theory of genetic analysis II. Digenic models with linkage for the classical cross. Genetics **60**: 169–170.

FALCONER, D. S., 1970 *Introduction to quantitative genetics.* The Ronald Press Co., New York.

KAPLAN, E. B. and R. C. ELSTON, 1972 A subroutine package for maximum likelihood estimation (MAXLIK). University of North Carolina Institute of Statistics Mimeo Series, No. **823**.

KENDALL, M. G. and A. STUART, 1961 *The advanced theory of statistics.* Volume 2. Charles Griffin and Co. Ltd., London.

LEWIS, P. A. W., 1965 Some results on tests for Poisson processes. Biometrika **52**: 67–77.

NELDER, J. A. and R. MEAD, 1965 A simplex method for function minimization. Computer J. **7**: 308–313.

NEYMAN, J., 1937 "Smooth test" for goodness of fit. Skand. Aktuar. **20**: 150–199.

POWELL, M. J. D., 1964 An efficient method for finding the minimum of a function of several variables without calculating derivatives. Computer J. **7**: 155–162.

POWERS, L., 1963 The partitioning method of genetic analysis and some aspects of its application to plant breeding. pp. 280–318. In: *Statistical Genetics and Plant Breeding.* Edited by W. D. HANSON and H. F. ROBINSON. National Academy of Sciences, Washington.

PYKE, R., 1965 Spacings. J. Roy. Statist. Soc. B. **27**: 395–449.

RAO, C. R., 1964 Criteria of estimation in large samples. pp. 345–362. In: *Contribution to statistics.* Edited by C. R. RAO. Pergamon Press, New York.

ROSENBROCK, H. H., 1960 An automatic method for finding the greatest or least value of a function. Computer J. **3**: 175–184.

STEWART, J., 1969a Biometrical genetics with one or two loci. I. The choice of a specific genetic model. Heredity **24**: 211–224. ——, 1969b B:ometrical genetics with one or two loci. II. The estimation of linkage. Heredity **24**: 225–238.

STEWART, J. and R. C. ELSTON, 1973 Biometrical genetics with one or two loci: the inheritance of physiological characters in mice. Genetics (this issue).

STUDENT, 1934 A calculation of the minimum number of genes in Winter's selection experiment. Ann. Eug (London) **6**: 77–82.

TAN, W. Y. and W. C. CHANG, 1972 Convolution approach to the genetic analysis of quantitative characters of self-fertilized populations. Biometrics **28**: 1073–1090.

TROW, A. H., 1913 Forms of reduplication: primary and secondary. J. Genet. **2**: 313–324.

WEBER, E., 1959 The genetic analysis of characters with continuous variability on a Mendelian basis. I. Monohybrid segregation. Genetics **44**: 1131–1139.

WRIGHT, S., 1934 The results of crosses between inbred strains of guinea pigs differing in numbers of digits. Genetics **19**: 537–551. ——, 1968 *Evolution and the genetics of populations.* Volume 1. University of Chicago Press, Chicago.