ON TESTING WHETHER ONE LOCUS CAN ACCOUNT FOR THE GENETIC DIFFERENCE IN SUSCEPTIBILITY BETWEEN TWO HOMOZYGOUS LINES

R. C. ELSTON

Department of Biostatistics and the Genetics Curriculum, University of North Carolina, Chapel Hill

Received January 13, 1966

THIS paper develops an appropriate statistical analysis for the following kind of experimental data. We have two isogenic lines (parental lines) which differ markedly in their susceptibilities to a given stimulus; in the extreme case one line is completely susceptible, all individuals in that line responding to the stimulus, while the other line is completely insusceptible. (The word "stimulus" is here being used in a very broad sense, and would include, for example, a chemical compound which is usually toxic to one line and nontoxic to the other. It could even be simply the normal environment, under which a certain phenotype is displayed in markedly different proportions in the two parental lines.) The susceptibilities are further determined in some or all of the following classes of individuals: the F_1 obtained by crossing the two parental lines; the backcrosses of the F_1 to each of the parental lines; and the F_2 . Given a set of data of this kind, we wish to test whether the experimental results are compatible with the hypothesis that the observed differences in susceptibility are controlled solely by alleles at one locus.

Compatibility of data of the type described and a one locus hypothesis is very weak evidence for such a simple hypothesis, as a great many other genetic hypotheses will give rise to exactly the same results. Only further breeding tests can distinguish between the involvement of one and more than one locus (WRIGHT 1934). It is nevertheless useful to have a test that can be applied to the limited data indicated above, which are relatively easy to obtain, for in many cases such data will be sufficient to rule out the possibility of only one locus being involved.

It will be assumed that the locus or loci involved are homozygous in the two parental lines, so that at any one locus only two different alleles can occur. Thus if k loci are involved, there are altogether 3^k different genotypes possible; if only one locus is involved the differences in susceptibility are determined by three different genotypes, while if two or more loci are involved they are determined by nine or more different genotypes. Since we are examining no more than six genotypically different classes, there is no possibility of distinguishing between the involvement of two loci and more than two loci; and to detect a difference between the involvement of one locus and more than one locus it is necessary in general to examine at least four of the six genotypically different classes.

TABLE 1

Genotypic constitution and probability of response on a one locus hypothesis, and observed results

Class	P_1	$P_1 \times F_1$	F_1	F_2	$P_2 \times F_1$	P_2
Genotypic constitution	AA	$\frac{1}{2}AA + \frac{1}{2}AB$	AB	$\frac{1}{4}AA + \frac{1}{2}AB + \frac{1}{4}BB$	$\frac{1}{2}AB + \frac{1}{2}BB$	BB
Probability of response, p_i	α_0	$\frac{1}{2}(\alpha_0 + \alpha_1)$	α_1	$\frac{1}{4}(\alpha_0 + 2\alpha_1 + \alpha_2)$	$\frac{1}{2}(\alpha_1 + \alpha_2)$	α_2
Number examined, n_i	n_0	n_1	n_2	n_3	n_4	n_5
Number responding, x_i	x_0	x_1	x_2	x_3	x_4	x_5

Let the two parental lines be P_1 and P_2 , with the corresponding genotypes—if only one locus is involved— AA and BB . Let the probability that any given individual will respond to the stimulus be α_0 for AA genotypes, α_1 for AB genotypes, and α_2 for BB genotypes. Then the genotypic constitution and probability of response p_i for the six different classes are as given in Table 1. Suppose that in each class we examine n_i individuals, x_i of whom respond and $n_i - x_i$ of whom do not, the subscript i for each class being as indicated in Table 1. Thus under the null hypothesis that only one locus is involved the x_i ($i = 0, 1, \dots, 5$) are independent binomially distributed random variables with parameters p_i and n_i , where p_i is the appropriate function of α_0, α_1 and α_2 . (If we do not examine any individuals in a particular class, then $n_i = x_i = 0$ for that class.)

Once we have obtained efficient estimates of α_0, α_1 and α_2 , it is a simple matter to compare the observed numbers (x_i and $n_i - x_i$) with the expected numbers in each class to perform a chi-square test of goodness of fit; the number of degrees of freedom associated with the total χ^2 statistic obtained is the number of classes examined less the number of parameters estimated, i.e. the number of classes examined less three if we estimate α_0, α_1 and α_2 . Efficient estimates can be obtained by the method of maximum likelihood, but this is relatively difficult computationally. We shall give this method of estimation for the special case in which $\alpha_0 = 0$ and $\alpha_2 = 1$, i.e. one parental line is completely insusceptible and the other is completely susceptible, so that only the one parameter α_1 need be estimated. For the more general case we shall show how estimates can be easily obtained, without the need for successive approximations by iteration, using what is sometimes known as the method of modified minimum χ^2 (KENDALL and STUART 1961). Such estimates are fully efficient, and are asymptotically the same as maximum likelihood estimates; for finite samples there is no reason to prefer the one type of estimate to the other.

Maximum likelihood estimation of α_1 when $\alpha_0 = 0$ and $\alpha_2 = 1$: When $\alpha_0 = 0$ and $\alpha_2 = 1$ it is clear from Table 1 that the likelihood is proportional to

$$\alpha_1^{x_1} (2 - \alpha_1)^{n_1 - x_1} \alpha_1^{x_2} (1 - \alpha_1)^{n_2 - x_2} (1 + 2\alpha_1)^{x_3} (3 - 2\alpha_1)^{n_3 - x_3} (1 + \alpha_1)^{x_4} (1 - \alpha_1)^{n_4 - x_4}$$

Differentiating the natural logarithm of this expression with respect to α_1 we obtain the score

$$s(\alpha_1) = \frac{x_1 + x_2}{\alpha_1} - \frac{n_1 - x_1}{2 - \alpha_1} - \frac{(n_2 - x_2 + n_4 - x_4)}{1 - \alpha_1} + \frac{2x_3}{1 + 2\alpha_1} - \frac{2(n_3 - x_3)}{3 - 2\alpha_1} + \frac{x_4}{1 + \alpha_1}, \quad (1)$$

and the maximum likelihood estimate of α_1 is a solution to the equation $s(\alpha_1) = 0$.

This equation can be solved by the standard iterative technique, using the result that, where E denotes expectation,

$$-E \frac{\partial s(\alpha_1)}{\partial \alpha_1} = \frac{n_1 + 2n_2}{2\alpha_1} + \frac{n_1}{2(2-\alpha_1)} + \frac{4n_3}{(1+2\alpha_1)(3-2\alpha_1)} + \frac{n_2}{1-\alpha_1} + \frac{n_4}{1-\alpha_1^2} = I(\alpha_1), \quad (2)$$

say. (This quantity is the "information".) Thus if α_1^* is a guessed solution to $s(\alpha_1) = 0$, then a better approximation is $\alpha_1^* + s(\alpha_1^*)/I(\alpha_1^*)$; this is illustrated numerically later.

Modified minimum chi-square estimation for the general case: If the ordinary χ^2 statistic is modified by substituting the observed for the expected numbers in the denominators, we obtain, for the i th class,

$$\frac{(x_i - n_i p_i)^2}{x_i} + \frac{(n_i - x_i - n_i(1-p_i))^2}{n_i - x_i} = \frac{n_i(x_i - n_i p_i)^2}{x_i(n_i - x_i)} \quad (3)$$

The modified minimum χ^2 estimates are those values of the parameters that make this quantity, summed over all classes for which we have data (i.e. for which $n_i \neq 0$), a minimum. Thus the estimating equations are obtained by differentiating (3) with respect to α_j ($j = 0, 1, 2$), summing over i and equating to zero, i.e. we have three equations of the form

$$\sum_i \frac{n_i^2 p'_i (n_i p_i - x_i)}{x_i(n_i - x_i)} = 0, \quad (4)$$

where p'_i is the derivative of p_i with respect to one of the α_j . Since in no case is p'_i a function of α_j , the equations (4) are linear in the α_j , and so easy to solve.

Writing $z_i = n_i^2 p'_i / (n_i - x_i)$ and $m_i = n_i z_i / x_i$, (4) becomes

$$\sum_i m_i p'_i p_i = \sum_i z_i p'_i. \quad (5)$$

Taking the p_i from Table 1 it is easily verified that the three simultaneous equations (5) can be written, after collecting terms and multiplying through by 16 to avoid fractions, as follows:

$$(16m_0 + 4m_1 + m_3)\alpha_0 + 2(2m_1 + m_3)\alpha_1 + m_3\alpha_2 = 4(4z_0 + 2z_1 + z_3) \quad (6)$$

$$2(2m_1 + m_3)\alpha_0 + 4(m_1 + 4m_2 + m_3 + m_4)\alpha_1 + 2(m_3 + 2m_4)\alpha_2 = 8(z_1 + 2z_2 + z_3 + z_4) \quad (7)$$

$$m_3\alpha_0 + 2(m_3 + 2m_4)\alpha_1 + (m_3 + 4m_4 + 16m_5)\alpha_2 = 4(z_3 + 2z_4 + 4z_5) \quad (8)$$

Equation (7) can be divided through by two without introducing any fractions, but by leaving it as it is the 3×3 matrix of coefficients of the α_j is symmetric.

For the general case where α_0 , α_1 and α_2 have all to be estimated, equations (6), (7) and (8) are best solved numerically by any of the standard methods. For special cases in which one or two of the α_j are assumed known, simple algebraic solutions can be given. Thus if it is known for example that $\alpha_0 = 0$ (i.e. P_1 is completely insusceptible), we obtain from (7) and (8)

$$\alpha_1 = 2(cd - be)/(ac - b^2), \quad \alpha_2 = 4(ae - bd)/(ac - b^2) \quad (9)$$

where $a = m_1 + 4m_2 + m_3 + m_4$, $b = m_3 + 2m_4$, $c = m_3 + 4m_4 + 16m_5$,
 $d = z_1 + 2z_2 + z_3 + z_4$, and $e = z_3 + 2z_4 + 4z_5$. For the special case $\alpha_0 = 0$

and $\alpha_2 = 1$, for which the maximum likelihood solution is given above, the modified minimum χ^2 estimate of α_1 is obtained from (7) and is

$$(4e - b)/2a, \quad (10)$$

a , b and e being the same as before.

A difficulty arises in the method of estimation just described when, for any class, $x_i = 0$ or $x_i = n_i$; for then m_i or z_i is infinite. It is probably best in such cases to substitute $x_i = 1/2$ for an observed value $x_i = 0$, and $x_i = n_i - 1/2$ for an observed value $x_i = n_i$; this can be justified intuitively by interpreting an observed $x_i = 0$ as being due to the fact that the expected value of x_i lies somewhere between 0 and 1, and analogously for an observed value $x_i = n_i$. This substitution in no way changes the asymptotic properties of the estimates obtained, and has been found by BERKSON (1953) to give good results in another connection.

It should be noted that if all the n_i are the same, and equal to n , say, then the above computations are simplified. For example it suffices, to estimate three parameters, to define $z_i = k/(n_i - x_i)$ and $m_i = z_i/x_i$, where k is some power of ten (chosen so that it is unnecessary to carry a decimal point), and then solve (6), (7) and (8) with these values of z_i and m_i ; the resulting values of α_j will then simply need to be divided by n as a final step.

Numerical example: The methods described in this paper can be illustrated by the data of FULLER, EASLER and SMITH (1950), concerned with audiogenic seizure in the mouse. A small part of their data, sufficient to illustrate the methods, is reproduced here in Table 2; the parental lines are C57BL and DBA mice, and the response that we shall analyse here is convulsion in any one of five trials when an individual is stimulated by being exposed to the sound of a bell.

If we are prepared to assume that C57BL mice never convulse and DBA mice always convulse (at least once in five trials), then there is only the one parameter α_1 to estimate. A first estimate of α_1 can be obtained from the F_1 individuals as $30/38 = 0.79$, and from this the maximum likelihood estimate can be obtained. Using (1) and (2) we have

$$s(0.79) = \frac{81}{0.79} - \frac{100}{1.21} - \frac{18}{0.21} + \frac{156}{2.58} - \frac{68}{1.42} + \frac{181}{1.79} = 47.87$$

$$I(0.79) = \frac{227}{1.58} + \frac{151}{2.42} + \frac{448}{(2.58)(1.42)} + \frac{38}{0.21} + \frac{191}{0.376} = 1017.3.$$

TABLE 2

Data on audiogenic seizure susceptibility

Class	P_1	$P_1 \times F_1$	F_1	F_2	$P_2 \times F_1$	P_2
n_i	50	151	38*	112	191	96
x_i	0	51	30	78	181	96
$z_i = n_i^2/(n_i - x_i)$	50.51	228.01	180.50	368.94	3648.10	18,432.00
$m_i = n_i z_i / x_i$	5050.51	675.09	228.63	529.76	3849.62	18,528.50

* Only the F_1 "repeat" sample of FULLER *et al.* (1950) is taken for this analysis. They state that the observed percent convulsers in five trials is based on a subsample of 38 animals, and so this is the value of n_i taken here.

Thus a better approximation is $0.79 + 47.87/1017.3 = 0.837$. Taking this value for a second iteration, we find $s(0.837) = 5.95$ and $I(0.837) = 1197.9$, so that the next better approximation is $0.837 + 5.95/1197.9 = 0.842$. A third iteration changes the estimate to 0.843, and this is the maximum likelihood estimate, correct to three decimal places.

If we do not wish to assume that $\alpha_0 = 0$ and $\alpha_2 = 1$, the modified minimum χ^2 estimation equations for all three parameters are easily formed from the last two rows of Table 2 (to compute which it has been assumed that $x_0 = 0.5$ and $x_5 = 95.5$). The equations (6), (7) and (8) are found to be, after rounding,

$$84,038 \alpha_0 + 3,760 \alpha_1 + 530 \alpha_2 = 4,108 \quad (11)$$

$$1,880 \alpha_0 + 11,938 \alpha_1 + 8,229 \alpha_2 = 18,424 \quad (12)$$

$$530 \alpha_0 + 16,458 \alpha_1 + 312,384 \alpha_2 = 325,573 \quad (13)$$

the solution to which is

$$\alpha_0 = 0.004, \quad \alpha_1 = 0.855, \quad \alpha_2 = 0.997.$$

It is clear that in this case the estimates of α_0 and α_2 are virtually 0 and 1. If we wish to assume these values, we can estimate α_1 by the method of modified minimum χ^2 directly from (12) or from (10) as 0.854, which compares well with the maximum likelihood estimate. If we wish to estimate two parameters, for example if we only wish to assume $\alpha_0 = 0$, we find directly from (12) and (13) or from (9) $\alpha_1 = 0.856$ and $\alpha_2 = 0.997$.

Using any of these estimates it is a simple matter to calculate a χ^2 for goodness of fit in the usual way. If we take the three parameters to be estimated, a χ^2 value of 8.4 with three degrees of freedom is obtained. If we take α_0 and α_2 to be known as 0 and 1 respectively, using the maximum likelihood estimate of α_1 we obtain a χ^2 value of 7.3, and using the modified minimum χ^2 estimate the value is 7.6; in both these cases the number of degrees of freedom is still three, since not only have two fewer parameters been estimated, but also there are two fewer binomial distributions (the parental classes) available for the test. The tabulated 95% point of χ^2 with three degrees of freedom is 7.8, so by any of these tests we should be led to believe, on the basis of these data alone, that control at a single locus probably cannot account for the differences in susceptibility observed.

SUMMARY

Suppose we have samples of individuals from two isogenic lines and some or all of the following classes: the F_1 , each backcross and the F_2 . Each individual is tested to determine if he is susceptible or not to a particular stimulus. A statistical procedure is given to test whether the observed differences in susceptibility in the various classes can be accounted for by control at a single locus. Previously published data on audiogenic seizure susceptibility in the mouse are used to illustrate the procedure.

LITERATURE CITED

- BERKSON, J., 1953 A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *J. Am. Stat. Assoc.* **48**: 565-599.

- FULLER, J. L., C. EASLER, and M. E. SMITH, 1950 Inheritance of audiogenic seizure susceptibility in the mouse. *Genetics* **35**: 622-632.
- KENDALL, M. G., and A. STUART, 1961 *The Advanced Theory of Statistics*, Vol. 2, Charles Griffin, London. p. 93.
- WRIGHT, S., 1934 The results of crosses between inbred strains of guinea pigs differing in numbers of digits. *Genetics* **19**: 537-551.