A DISCRIMINANT FUNCTION FOR PLANT SELECTION

By H. FAIRFIELD SMITH, M.S.A.

Division of Plant Industry, Council for Scientific and Industrial Research, Canberra, Australia

I. Theory

The characters with which a plant breeder is principally concerned are those known as "quantitative characters". They present particular difficulty because heritable variations are masked by larger non-heritable variations which make it difficult to determine the genotypic values of individual plants or lines unless we have sufficient seed and facilities to grow replicated plots of each line. In the earlier stages of selection breeders try to select plants in the field on the basis of observable characters which they believe may be associated with the desired character or quality (for example, grain and ear sizes as indices to yielding ability, or flintiness of grain as an index of protein content), but the actual worth to be attributed to each character is usually unknown. The problem may be approached by seeking to determine what "discriminant function" (Fisher, 1936) of the observable characters may best indicate the "genetic value" of a plant or line.

Suppose that in a wheat-selection programme we are required to consider n characters, say $x_1, x_2, \dots x_n$. Let us evaluate each in terms of one of them, say x_1. For example, suppose we take x_1 to represent yield of grain; x_2 may represent baking quality and we may consider that an advance of 10 in baking score is equal in value to an advance of 1 bushel per acre in yield; x_3 may represent resistance to flag smut and we may evaluate a decrease of 20 per cent infection as worth 1 bushel of yield;* and so on. Let these values be designated $a_1, a_2, \dots, a_n$. Then taking yield, x_1, as standard and units as indicated, we will have

$$a_1 = 1, \quad a_2 = 0{\cdot}1, \quad a_3 = -0{\cdot}05, \quad \text{etc.}$$

Then, if ξ is the value of x to be expected due to genotype, the genotypic value of a line may be scored as

$$\psi = a_1\xi_1 + a_2\xi_2 + \dots + a_n\xi_n \qquad \dots\dots(1).$$

This, however, cannot be directly evaluated since we can observe only the phenotypes

* Characters such as disease resistance may introduce other considerations which will not be taken up in detail here. For example, we may grow the plants in a disease-infested nursery and suppose that the disease-resisting qualities are sufficiently represented in our measure of yield so that no attention need be given to disease as a separate character. But if we are interested in producing disease-resisting varieties for use as a control measure to reduce the prevalence of disease in a district we may consider disease resistance as a character of value in itself and be willing to sacrifice something in yield to obtain it. Again, we may wish to keep the plant-breeding nursery free of disease but we may take some grain from each plant and test it in the laboratory for disease-resisting properties, which will then require evaluation independently of characters observed in the field.

which, because of non-heritable variations, do not accurately represent their genotypes. Let the phenotypes be scored according to the equation

$$Y = b_1 x_1 + b_2 x_2 + \dots + b_n x_n \qquad \dots\dots(2).$$

Our problem is then to find the values of b such that the function Y may best discriminate those lines which have the greatest genotypic value ψ.

Suppose that we have a large number of plant lines and we have decided to select for further propagation one qth part of them. Assuming that the values of Y for each line are normally distributed with variance V, their frequency distribution may be represented by Fig. 1, and the lines to be selected will be those falling in the shaded area q, that is all those having Y greater than a fixed value Y'. If Y be transformed to a variate, u, with unit variance and mean at zero, that is

$$u = \frac{Y - \bar{Y}}{\sqrt{V}},$$

then the value of $u' = (Y' - \bar{Y})/\sqrt{V}$ corresponding to any given value of q may be ascertained from a table of the normal probability integral.

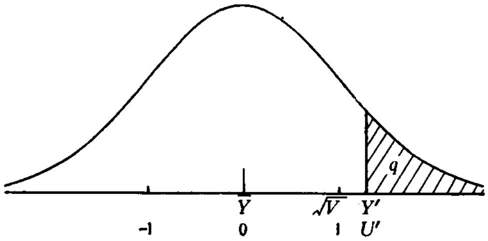

Fig. 1.

Let the regression coefficient of ψ on Y be B. Then the mean value of ψ associated with any given value of Y is given by

$$(\psi - \bar{\psi}) = B\,(Y - \bar{Y}) = B\,(V)^{\frac{1}{2}} u \qquad \dots\dots(3).$$

Summing for all values of Y greater than Y' and dividing by their frequency, q, the mean value or expectation of $(\psi - \bar{\psi})$ to be associated with the selected values of Y is

$$E\,(\psi - \bar{\psi}) = \frac{1}{q}\int_{u=u'}^{\infty} B\sqrt{V}\,.\,u \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}u^2}\, du$$

$$= \frac{z}{q}\, B\,(V)^{\frac{1}{2}} \qquad \dots\dots(4),$$

where z is the ordinate of the unit normal curve at the deviate u', and $E\,(\psi - \bar{\psi})$ may be described as the expectation of genetic advance above the mean of the original population for a given intensity of selection q. The selection intensity having been determined, z is also fixed, and therefore to maximize $E\,(\psi - \bar{\psi})$ we require to maximize $B\,(V)^{\frac{1}{2}}$.

Assume now, by hypothesis, that an observed value of a character x_i may be treated as the sum of two parts, ξ_i due to genotypic and ϵ_i due to environmental factors, so that

$$x_i = \xi_i + \epsilon_i.$$

Further, let us assume that the two parts are independent, that is, the covariance of ξ_i and ϵ_i has an expectation of zero. Then if t_{ii}, g_{ii}, and e_{ii} be the variances of x_i, ξ_i and ϵ_i respectively,

$$t_{ii} = g_{ii} + e_{ii}.$$

Similarly, let the covariances of x_i and x_j be t_{ij}, of ξ_i and ξ_j be g_{ij}, and of ϵ_i and ϵ_j be e_{ij}, with similar additive connexions.

Then the variance of Y is

$$\begin{aligned} V &= E\,(Y - \bar{Y})^2 \\ &= b_1^2 t_{11} + b_2^2 t_{22} + \ldots + 2b_1 b_2 t_{12} + 2b_1 b_3 t_{13} + \ldots \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j t_{ij}, \end{aligned}$$

and the covariance of ψ and Y is

$$\begin{aligned} W &= E\,(\psi - \bar{\psi})\,(Y - \bar{Y}) \\ &= a_1 b_1 g_{11} + a_2 b_2 g_{22} + \ldots + (a_1 b_2 + a_2 b_1)\, g_{12} + (a_1 b_3 + a_3 b_1)\, g_{13} + \ldots \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j g_{ij}, \end{aligned}$$

since the expectations of $(\xi_i \epsilon_j)$ for all values of i and j are zero. The regression of ψ on Y is then

$$B = \frac{W}{V} = \frac{\Sigma\Sigma a_i b_j g_{ij}}{\Sigma\Sigma b_i b_j t_{ij}}$$

and

$$B\,(V)^{\frac{1}{2}} = \frac{\Sigma\Sigma a_i b_j g_{ij}}{\sqrt{\Sigma\Sigma b_i b_j t_{ij}}} \qquad \ldots\ldots(5).$$

Putting

$$\log B\,(V)^{\frac{1}{2}} = \log W - \tfrac{1}{2} \log V$$

it is easy to obtain

$$\frac{\partial \log B\,(V)^{\frac{1}{2}}}{\partial b_1} = \frac{\sum\limits_j a_j g_{1j}}{W} - \frac{\sum\limits_j b_j t_{1j}}{V},$$

and this is equal to zero (that is $\log B\,(V)^{\frac{1}{2}}$—and therefore also $B(V)^{\frac{1}{2}}$—is maximum) when

$$\sum_j b_j t_{1j} = \frac{V}{W} \sum_j a_j g_{1j} = KA_1 \qquad \ldots\ldots(6).$$

We can thus obtain n equations of the general form

$$\begin{aligned} b_1 t_{11} + b_2 t_{12} + \ldots + b_n t_{1n} &= KA_1, \\ b_1 t_{12} + b_2 t_{22} + \ldots + b_n t_{2n} &= KA_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots &, \\ b_1 t_{1n} + b_2 t_{2n} + \ldots + b_n t_{nn} &= KA_n, \end{aligned}$$

which may be written briefly as

$$\sum_{j=1}^{n} b_j t_{ij} = KA_i \qquad \ldots\ldots(7),$$

where i is constant in each equation, $A_i = \sum\limits_j a_j g_{ij}$ and $K = V/W$.

The solutions of these n equations are given by

$$b_j = K\sum_i A_i c_{ij} \qquad \text{......(8)},$$

where j is now constant in each equation and the values of c_{ij} are given by the matrix*

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{12} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{1n} & c_{2n} & \cdots & c_{nn} \end{bmatrix},$$

which is the reciprocal of

$$\begin{vmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{12} & t_{22} & \cdots & t_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{1n} & t_{2n} & \cdots & t_{nn} \end{vmatrix} = |t|.$$

That is,

$$c_{ij} = \frac{T_{ij}}{|t|},$$

where T_{ij} is the co-factor of t_{ij} in $|t|$.

Since we are interested only in the relative values of the b's it is unnecessary to evaluate K which is constant in all equations, and it is sufficient to evaluate simply

$$b_j/K = \sum_i A_i c_{ij} = d_j.$$

From equation (7)

$$\sum_j b_j t_{ij} = K\sum_j a_j g_{ij}.$$

Therefore

$$\sum_i\sum_j b_i b_j t_{ij} = K\sum_i\sum_j b_i a_j g_{ij}$$

and

$$B\,(V)^{\frac{1}{2}} = \sqrt{K\sum_i\sum_j b_i a_j g_{ij}} = \sqrt{K\sum_i b_i A_i} = \sqrt{\sum_i d_i A_i}.$$

Substituting in equation (4) the expectation of genetic advance is therefore

$$(z/q)\sqrt{\Sigma dA} \qquad \text{......(10)},$$

where z/q is a constant depending on the intensity of selection whose value for any given selection intensity q can be ascertained from the Kelley-Wood table of the normal probability integral (Kelley, 1923). The units of this expression are the same as those by which the standard character x_1 has been measured.

The hypothetical premises from which the above deductions proceed are (1) that genotypic and environmental effects are additive to give the observed magnitude of a character, (2) that they are independent, and (3) that Y and ψ are normally distributed. The first two are comparable to the assumptions usually made in applying other statistical

* In practice $c_{11}, c_{12}, c_{13}, \ldots$ are found by solving the equations

$$\sum_j c_{1j} t_{ij} = 0 \qquad \text{......(9)},$$

except that

$$\sum_j c_{1j} t_{1j} = 1.$$

Similarly $c_{12}, c_{22}, c_{23}, \ldots$ are the solutions of the equations

$$\sum_j c_{2j} t_{ij} = 0,$$

except that

$$\sum_j c_{2j} t_{2j} = 1,$$

etc. (Fisher, 1925–34, Sec. 29).

methods which depend upon linear functions. They appear inherently reasonable but perhaps deserve to be the subject of research. With respect to the third, since estimates of regression coefficients are not biassed by considerable departure of the frequency distributions of the variates from normal, such departure would not affect the procedure used to determine the b's; but since the derivation of equation (4) does depend directly on the distribution of Y departure from normality would affect estimates of genetic advance.

II. Examples

(i) *A variety trial.* Smith (1936) reports two variety trials with Australian wheat varieties grown in replicated square yard plots. In addition to yield of grain the components of yield—ear number, grain number and grain size—and weight of straw were observed. Suppose that it is required to discriminate from these observations the genotypic values of the varieties with respect to yield of grain.

The data have been recorded as the logarithms of observed figures so we shall take as our variates

x_1 = logarithm of yield of grain per plant at 0·01 sq. yd. per plant.
x_2 = ,, ,, ear number per plant.
x_3 = ,, ,, average number of grains per ear.
x_4 = ,, ,, average weight per grain.
x_5 = ,, ,, weight of straw per plant.

Obviously $x_1 = x_2 + x_3 + x_4$ ......(11),

and, if we are concerned only with yield of grain, our equation of value is

$\psi = x_1 = x_2 + x_3 + x_4$ ......(12);

that is, we may take either $a_1 = 1$ and $a_2 = a_3 = a_4 = a_5 = 0$ ......(13),

or $a_1 = a_5 = 0$ and $a_2 = a_3 = a_4 = 1$ ......(14).

Both sets of constants lead to precisely the same results throughout.

On account of the identity (11) we can have only three independent equations among the variates x_1 to x_4. Let the discriminant function be

$Y = b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$ ......(15).

For the 1933 experiment analyses of variance and covariance are given in Table XXI of the paper cited. As estimates of t_{ij} we may take one-sixth* of the mean squares and mean products between varieties. This gives a matrix for t_{ij} as in Table I.

Table I. t_{ij} *for means of* 6 *plots* (1933), × 10^6

i \ j	2	3	4	5
2	3311	−1810	−1391	385
3	—	1535	1299	216
4	—	—	2087	488
5	—	—	—	859

* For immediate purposes we could obviously take simply the intervariety mean squares for t, and intervariety less error mean squares for g; but intervariety sums of squares having been determined from sums of 6 plots, one-sixth of these values is more convenient for further calculations to be described later.

Solving the system of equations (9) gives values of c_{ij} as in Table II.

Table II. c_{ij} *for means of* 6 *plots* (1933)

j i	2	3	4	5
2	1536·432	1862·173	162·390	−1272·259
3	—	3647·819	−702·274	−1397·975
4	—	—	1150·743	− 541·279
5	—	—	—	2411·184

As estimates of e_{ij} we may take one-sixth of the mean squares and mean products for remainder (or error), and by subtracting these from the corresponding values of t_{ij} in Table I we obtain estimates of g_{ij} (Table III) (compare Smith, 1936, Appendix 4).

Table III. g_{ij} *for the* 1933 *experiment*, $\times 10^6$

j i	2	3	4	5	A
1	−48	946	1993	913	—
2	3086	−1743	−1392	252	−48
3	—	1363	1326	171	946
4	—	—	2059	490	1993
5	—	—	—	654	913

Owing to the identity (11) the A factors may be derived either from row 1 of Table III using (13), or from rows 2–5 using (14). Then substituting numerical values from Tables II and III in equations (8) we find

$$Y_1 \propto 0{\cdot}8487x_2+0{\cdot}6839x_3+1{\cdot}1269x_4-0{\cdot}1371x_5,$$

which is the linear function of the characters x_2 to x_5 which may best discriminate the genotypic yield potentialities of the various varieties under the conditions of this experiment.

We may, if desired, take one of our characters, say x_3 (grain number per ear), as having unit value and rewrite the formula as

$$Y_1 \propto 1{\cdot}2411x_2+x_3+1{\cdot}6479x_4-0{\cdot}2005x_5.$$

If we do not wish to evaluate grain number (x_3) but wish to use instead total weight of grain (x_1), the appropriate function is given by replacing x_3 by its equivalent $(x_1-x_2-x_4)$. Thus

$$Y_1 \propto x_1+0{\cdot}2411x_2+0{\cdot}6479x_4-0{\cdot}2005x_5 \qquad \ldots\ldots(16),$$

and this formula is identical with that which would have been obtained had we decided from the first to use these variates in the discriminant function (15). It indicates the extent to which observations of x_2, x_4 and x_5 may contribute to our knowledge of genotypic yielding ability over and above such information as would be given by observing yield of grain alone.

The 1932 experiment (*loc. cit.* Table XIX) with approximately the same group of varieties gave the formulae

$$Y_2 \propto 0{\cdot}784x_2 + 0{\cdot}771x_3 + 0{\cdot}955x_4$$

$$\propto 1{\cdot}017 + x_3 + 1{\cdot}239x_4$$

$$\propto x_1 + 0{\cdot}017x_2 + 0{\cdot}239x_4.$$

Straw weight (x_5) has not been included in these formulae because in 1932 straw production of one or two varieties seemed abnormally erratic. The relative values of the yield components show notable similarity to the 1933 figures for as much as the 1932 experiment was on much poorer and more heterogeneous soil. The principal difference, the lower value of weight per grain (x_4) in 1932, is due to this character having had a higher experimental error in 1932 than in 1933. The eminence of weight per grain in both years is due partly to its having the highest genetic correlation with yield (compare Smith, 1936, III) and partly to its experimental error (e_{44}) being considerably lower than that of the other characters.

The values of $(\Sigma dA)^{\frac{1}{2}}$ are 0·05227 for the 1933 data and 0·05234 for the 1932 data. For a 10 per cent selection intensity, $q = 0{\cdot}1$, $z/q = 1{\cdot}755$; therefore the expectation of genetic advance for a 10 per cent selection among a population of plant lines having statistics similar to the above would be $1{\cdot}755 \times 0{\cdot}0523 = 0{\cdot}0918$. Since the characters have been measured as logarithms and $\log^{-1} 0{\cdot}0918 = 1{\cdot}235$, it is indicated that the average yield of the selected 10 per cent would be about 23·5 per cent greater than the average of the population.

(ii) *Effect of varying experimental errors* (*or quantity of material averaged*). It is of some interest to consider how the discriminant functions would be affected if for reasons of labour, space or quantity of seed we could sow only a small area with each variety or line. Suppose we had only one square yard plot of each and we had 50 or 60 lines so that the total area of the experiment was unaltered. Suppose also that the genetic variability were the same as between the 10 varieties considered above, that is, with respect to the 1933 conditions, Table III still provides the estimates of g. Estimates of e will be given by mean squares and products in the row of the analyses entitled "total soil effects", Table XXI, and are reproduced in Table IV.

Table IV. e_{ij} *for single plots* (1933), $\times 10^6$

j i	2	3	4	5
2	1694	−7641	−17	953
3	—	1570	−93	353
4	—	—	195	60
5	—	—	—	1512

Adding Tables III and IV we obtain estimates of t_{ij} as in Table V.

Table V. t_{ij} *for single plots* (1933), $\times 10^6$

i \ j	2	3	4	5
2	4780	−9384	−1409	1205
3	—	2933	1233	524
4	—	—	2254	550
5	—	—	—	2166

Proceeding as before, the discriminant function is found to be

$$Y_3 \propto -0{\cdot}0038x_2 - 0{\cdot}1005x_3 + 0{\cdot}8822x_4 + 0{\cdot}2240x_5$$
$$\propto -0{\cdot}0379x_2 - x_3 + 8{\cdot}7755x_4 + 2{\cdot}2278x_5$$
$$\propto -x_1 + 0{\cdot}9621x_2 + 9{\cdot}7755x_4 + 2{\cdot}2278x_5.$$

Similarly for 1932 we find

$$Y_4 \propto 0{\cdot}184x_2 + 0{\cdot}347x_3 + 0{\cdot}574x_4$$
$$\propto 0{\cdot}530x_2 + x_3 + 1{\cdot}656x_4$$
$$\propto x_1 - 0{\cdot}470x_2 + 0{\cdot}656x_4.$$

The outstanding feature of these formulae is the relatively enhanced value of weight per grain due to the relative accuracy with which it may be measured even in small samples. The difference between the two experiments in this respect becomes particularly accentuated.

The values of $(\Sigma dA)^{\frac{1}{2}}$ are 0·04320 for 1933 and 0·04284 for 1932, indicating about 19 per cent as the expectation of genetic advance for a selection intensity of 10 per cent. But for comparison with the former figure (23·5 per cent) for a 10 per cent selection intensity among means of five plots it must be noted that if strains are grown in only one plot each, then five times as many strains can be grown for equal ground and labour and selection may be five times as intense as formerly. For $q = 0{\cdot}02$, $z/q = 2{\cdot}4209$, and the expectation of genetic advance $= 0{\cdot}043 \times 2{\cdot}421 = 0{\cdot}1041$ in logarithms $= 27{\cdot}1$ per cent.

Since the data in the above examples were in logarithmic form the coefficients there obtained refer to variations in *relative* size of the variates. We could derive a system of equations appropriate to arithmetical values but the above simple transformations between coefficients appropriate to the whole and component parts of a composite character would no longer be available. It would then probably be best to consider yield and its components as four individual variates. The system would be to some degree artificial because, although yield is uniquely determined when values have been given to all its components (say n in number), an $(n+1)$th equation is provided only by the discrepancy in attempting to describe a product function by a linear equation.

III. Ratios as selective factors

It is a common practice among breeders to seek to use a ratio between two plant characters as a selective index. Particular attention has been given to the ear : tiller ratio (the survival rate of tillers) and to the grain : straw ratio (the so-called "migration coefficient" when considered as the ratio of grain to total produce).

Suppose we take the logarithm of grain : straw ratio as a sixth variate (x_6), then we may write

$$x_6 = x_1 - x_5,$$

and if we choose to replace observations on straw weight by observations of grain : straw ratio the discriminant function for means of six plots in 1933 (16) would become

$$Y_1 \propto 0{\cdot}7995x_1 + 0{\cdot}2411x_2 + 0{\cdot}6479x_4 + 0{\cdot}2005x_6.$$

Or, if we wished to consider only total weights of grain and of straw, we would have, from Table XXI (*loc. cit.*):

Table VI. *t_{ij} and g_{ij} for grain and straw weights* (1933), × 10^6

j \ i	t_{ij}		g_{ij}		A
	1	5	1	5	
1	3128	1099	2891	913	2891
5	—	859	—	654	913

whence

$$Y_5 \propto x_1 - 0{\cdot}2164x_5$$
$$\propto 0{\cdot}7836x_1 + 0{\cdot}2164x_6.$$

Both formulae show that, for these data, grain : straw ratio alone can give only a small part of the available information on genotypic yield value.

In general the derivation of these functions demonstrates that grain : straw ratio can never supply more information than can be given by a consideration of both grain and straw separately and can give as much information only if the coefficients of x_1 and x_5 are numerically equal and opposite in sign. It seems that the latter condition may arise only under exceptional circumstances. Since both grain and straw yields must be determined before their ratio can be evaluated, and since all the available information can be obtained from the two primary variates, it seems an unnecessary step to evaluate the ratio.

One condition has so far been ignored. As the goodness of fit of a regression may be improved by considering higher powers and products of the variates, so here the precision of the selection function might be increased by considering terms of the second and higher degree. In similar manner the addition of grain : straw ratio to a function using arithmetical values for weights of grain and straw (grain, straw and their ratio being no longer connected by a simple linear function as above) might increase its accuracy. But if this were so it seems likely, although not inevitable, that an equation using logarithmic values, as in the above example, would give equally good results with less labour. Only extensive data can decide the respective merits of such slightly different procedures. Meantime it may be concluded that, while such ratios may occasionally provide supplementary knowledge, any attempt to use them to replace consideration of the primary variates will, except in very unusual circumstances, result in serious loss of information.

IV. Practical application

For present application to selection it is clearly necessary that we should have some estimate of the genetic and total variances and covariances of all characters in the breeding programme. This given, the appropriate function may be determined as shown, and then applied to discriminate in a logical manner the lines to be retained. For best results it will always be desirable to be able to determine, from internal evidence, the discriminant most suitable for each particular case just as we seek to determine an experimental error for each separate experiment. But just as from past experience we can anticipate the approximate magnitude of the error which is likely to occur under given conditions, so we may hope that further research will enable us to anticipate the variances and covariances likely to occur among a given group of characters in segregates from some specified system of crosses. We mȧy then construct a preliminary discriminant function which will serve for initial field selections. If for example we had a number of lines derived from a "composite hybrid mixture" (Harlan & Martini, 1929), we had reason to believe that the variances and covariances in the culture would be approximately as in the group of varieties studied above, and we had only small plots of less than 100 plants of each line, then, with respect to yield, initial field selection in such a culture would be concentrated almost entirely on size of grain. Definite suggestions must however await research on variability of characters in groups of segregates. At different stages in the selection process conditions may vary considerably from those indicated above for a group of established varieties.

V. Summary

The object of this paper is to suggest how a method for selecting plant lines may be worked out in a logical and systematic manner. The value of a plant may be expressed as a linear function of its characters, then, using Fisher's concept of "discriminant functions", we may derive that linear function of observable characters which will be the best available guide to the genetic value of each line.

The expectation of "genetic advance" over the mean of the unselected population for any given selection intensity may also be estimated and used to compare the relative efficiencies of various breeding programmes.

It is shown further that arbitrary ratios, such as the "migration coefficient" or the "tiller survival rate", are likely to be inefficient as indices to the genetic value of either of the characters whose ratio is observed.

Acknowledgements

A formal note of acknowledgement is scarcely adequate to express my indebtedness to Prof. R. A. Fisher for guidance and inspiration. In particular, section I of the above paper is little more than a transcription of his suggestions.

REFERENCES

Fisher, R. A. (1925–34). *Statistical Methods for Research Workers*. Edinburgh.

—— (1936). "The use of multiple measurements in taxonomic problems." *Ann. Eugen.* **7**, 179–89.

Harlan, H. V. & Martini, M. L. (1929). "A composite hybrid mixture." *J. Amer. Soc. Agron.* **21**, 487–90.

Kelley, T. L. (1923). *Statistical Method*, Appendix C. New York.

Smith, H. Fairfield (1936). "Investigations on analysing yield of wheat varieties. I–III." MS. in the custody of C.S. and I.R., Australia.