GENERALIZATION OF PRODUCT FORMULA*

By K. C. Seal

Calcutta

In all problems of statistical estimation the usual method of Maximum Likelihood is known to provide efficient estimates under quite general conditions. For any given problem of estimation other efficient methods may, and often do, exist. Thus values of recombination fraction p in mendelian genetics can be estimated efficiently by the use of the product formula [Fisher and Balmakund (1928), Immer (1930)]. This formula is simple to apply and possesses the desirable property that it also minimizes certain difficulties encountered in handling data showing poor viability of certain genotypes [Mather (1938)]. It is shown in this paper that there exists a class of estimating equations besides the product formula which leads to asymptotically efficient estimates. Among this class there is one equation which is simpler than the likelihood equation and the product formula. Further, this suggested method leads to an estimate which may be considered to be superior to that of the product formula when disturbed segregations are encountered. Thus it may be concluded that the suggested method of estimation should be preferred to the usual product formula in most of the practical situations.

Instead of confining ourselves to four phenotypic classes as is the case with Double Backcross or Double F_2 data we shall consider the general case of a multinomial distribution with k-frequency classes. Let a_i be the frequency in the i-th class, such that

$$E(a_i) = np_i,\ p_i = g_i(p) > 0,\ a_i > 0,$$

$$\sum_1^k p_i \equiv 1,\ \sum_1^k a_i = n,\ i = 1, 2, \cdots, k. \tag{1}$$

It is easily shown that in this case [see Cramér (1946); Huzurbazar (1948)] the maximum likelihood estimate, *i.e.*, the consistent solution of the likelihood euqation

$$\sum_1^k \frac{\partial p_i}{\partial p}\frac{a_i}{p_i} = 0, \tag{2}$$

* This paper was presented at the Forty-fourth Session of the Indian Science Congress Association, 1957.

exists. Let this estimate of p be denoted by $\hat{p}_n$ which is known to be asymptotically efficient.

Suppose that there exists a function $f(x)$ such that

$$|f(x) - (c_1 + c_2 x)| < \epsilon_1 \text{ when } |x| < \epsilon_2, \tag{3}$$

where $c_1, c_2 \neq 0$, are arbitrary constants, and $\epsilon_1 \to 0$ as $\epsilon_2 \to 0$.

The purpose of this note is to show that there exists one solution $\tilde{p}_n$ of the equation

$$\sum_{i=1}^{k} \frac{\partial p_i}{\partial p} f\left(\frac{a_i - np_i}{np_i}\right) = 0 \tag{4}$$

which is also asymptotically efficient. A particular choice of the function $f(x)$ leads to a generalization of the product formula developed by Fisher and Balmakund (1928) and Immer (1930).

Consider k random variables

$$\frac{a_i - np_i}{np_i}, \quad i = 1, 2, \cdots, k.$$

It can be readily shown that strong law of large number holds for a_i, so that

$$P_r\left[\left|\frac{a_i - np_i}{n}\right| < \epsilon_2, n = N, N+1, \cdots, N+r\right] > 1 - \epsilon_{3i}, \tag{5}$$

where $\epsilon_{3i} \to 0$ as $\epsilon_2 \to 0$, N is a large number and r an arbitrary number. Hence

$$\begin{aligned} P_r\left[\left|\frac{a_i - np_i}{n}\right| < \epsilon_2, i = 1, 2, \cdots, k;\right. \\ \left. n = N, N+1, \cdots, N+r\right] \\ \geq \sum_{i=1}^{k} P_r\left[\left|\frac{a_i - np_i}{n}\right| < \epsilon_2, n = N,\right. \\ \left. N+1, \cdots, N+r\right] - (k-1) \\ > k - \sum_{i=1}^{k} \epsilon_{3i} - (k-1) = 1 - \epsilon_3, \end{aligned} \tag{6}$$

where

$$\epsilon_3 = \sum_{i=1}^{k} \epsilon_{3i} \to 0 \text{ as } \epsilon_2 \to 0.$$

Hence $a_i - np_i/n$, $i = 1, \cdots, k$, jointly converge with probability 1 to 0. This implies that $a_i - np_i/np_i$, $i = 1, \cdots, k$, jointly converge with probability 1 to 0.

From condition (3) it follows that when

$$\left| \frac{a_i - np_i}{np_i} \right| < \epsilon_2,$$

$$\left| f\left(\frac{a_i - np_i}{np_i}\right) - \left(c_1 + c_2 \frac{a_i - np_i}{np_i}\right) \right| < \epsilon_1, \tag{7}$$

where

$$\epsilon_1 \to 0 \text{ as } \epsilon_2 \to 0.$$

Hence

$$P_r\left[\left| f\left(\frac{a_i - np_i}{np_i}\right) - \left(c_1 + c_2 \frac{a_i - np_i}{np_i}\right) \right| < \epsilon_1, \right.$$

$$\left. i = 1, \cdots, k;\ n = N, N+1, \cdots, N+r \right]$$

$$\geq P_r\left[\left| \frac{a_i - np_i}{np_i} \right| < \epsilon_2,\ i = 1, \cdots, k;\ n = N, \right.$$

$$\left. N+1, \cdots, N+r \right] > 1 - \epsilon_3, \tag{8}$$

where $\epsilon_1 \to 0$, $\epsilon_3 \to 0$ for sufficiently large N. Hence $f(a_i - np_i/np_i)$, $i = 1, \cdots, k$, for sufficiently large n, jointly converge with probability 1 to

$$c_1 + c_2 \frac{a_i - np_i}{np_i},\ i = 1, \cdots, k.$$

Consider now the estimating equation for p,

$$\sum_{i=1}^{k} \frac{\partial p_i}{\partial p} f\left(\frac{a_i - np_i}{np_i}\right) = 0. \tag{9}$$

The L.H.S. of the above equation (9) strongly converges to

$$\sum_{i=1}^{k} \frac{\partial p_i}{\partial p}\left(c_1 + c_2 \frac{a_i - np_i}{np_i}\right)$$

$$= \frac{c_2}{n} \sum_{i=1}^{k} \frac{\partial p_i}{\partial p} \frac{a_i}{p_i}, \text{ since } \sum_{1}^{k} \frac{\partial p_i}{\partial p} = 0.$$

Further $c_2 \neq 0$, hence, there exists a solution $\tilde{p}_n$ of equation (9) which is identical with the solution $\hat{p}_n$ of the likelihood equation

$$\sum_{i=1}^{k} \frac{\partial p_i}{\partial p} \frac{a_i}{p_i} = 0$$

except for a set of measure zero, when n is chosen sufficiently large. Thus $\tilde{p}_n = \hat{p}_n$ almost everywhere for sufficiently large n. This implies that $\tilde{p}_n$, *i.e.*, a solution of equation (9), is an asymptotically efficient estimate of p, since the asymptotic moments of $\tilde{p}_n$ and $\hat{p}_n$ are identical so that

$$\text{Var } (\tilde{p}_n) = \text{Var } (\hat{p}_n) = \text{minimum.}$$

Corollary.—If $f_1(x)$ be such that

(i) $f_1'(x)$ exists and is continuous in $[-\epsilon_2, \epsilon_2]$ and $f_1'(0) \neq 0$ for an arbitrarily small $\epsilon_2 > 0$,

(ii) $f_1''(x)$ exists in $(-\epsilon_2, \epsilon_2)$,

then $f_1(x)$ may be taken as the function $f(x)$ considered above.

A few examples of such $f(x)$ are now given below:—

$f(x)$	c_1	c_2
(i) $\log(1+x)$ ..	0	1
(ii) $(1+x)^{\alpha}$, $\alpha \neq 0$ ..	1	α
(iii) $e^{\pm\alpha(1+x)}$, $\alpha \neq 0$ ..	$e^{\pm\alpha}$	$\pm \alpha e^{\pm\alpha}$
(iv) $\sin x$	0	1
(v) $\tan x$	0	1

If $f(x) \equiv \log(1 + x)$ be considered, the equation (4) will reduce to

$$\sum_{1}^{k} \frac{\partial p_i}{\partial p} \log \frac{a_i}{np_i} = 0, \tag{4 a}$$

which implies

$$\sum_{1}^{k} \frac{\partial p_i}{\partial p} \log a_i = \sum_{1}^{k} \frac{\partial p_i}{\partial p} \log p_i, \tag{4 b}$$

that is,

$$\prod_{i=1}^{k} a_i^{\frac{\partial p_i}{\partial p}} = \prod_{i=1}^{k} p_i^{\frac{\partial p_i}{\partial p}}. \tag{4 c}$$

When double backcross or single backcross families are considered we have $k = 4$ and $p_i = \alpha_i p + \beta_i$, $p_i > 0$, $\sum_{1}^{4} p_i = 1$, p being the recombination factor, so that the above estimating equation reduces to the well-known product formula

$$\frac{a_1 a_4}{a_2 a_3} = \frac{p_1 p_4}{p_2 p_3}. \tag{10}$$

Likewise for double F_2, with four phenotypic classes equation (4 c) reduces to the product formula by taking p_i as linear function of P, where $P = (1 - p)^2$ or p^2 according as coupling or repulsion data are considered. The estimate of p obtained by such product formula is known to be asymptotically efficient; the above estimating equation (4 c) is therefore a generalization of the above formula.

From the preceding arguments it follows that there exists many other estimating equations besides the product formula which have the same desirable property. Among these estimating equations the product formula is seen to possess the special property that the corresponding estimating equation is expressible in a simple compact form.

The product formula has another advantage in the estimation of recombination fractions, namely that by its use certain difficulties encountered in handling data showing poor viability of certain genotypes are minimized [see Mather (1938)]. It will now be shown that among the class of estimating equations suggested above there is one equation which is very easy to solve and is also superior to the product formula

in minimizing errors of estimation of recombination fraction in disturbed segregations.

Taking $f(x) = (1 + x)^{-1}$ it is easily seen, the estimating equation is of the form

$$\sum_{i=1}^{k} \frac{\partial p_i}{\partial p} \frac{p_i}{a_i} = 0 \tag{11}$$

Thus for the Double Backcross repulsion families the equation (11) reduces to

$$p\left(\frac{1}{a_1} + \frac{1}{a_4}\right) - \left(1 - p\right)\left(\frac{1}{a_2} + \frac{1}{a_3}\right) = 0, \tag{11 a}$$

where a_1, a_2, a_3, a_4 represent the observed frequencies in AB, Ab, aB and ab phenotypic classes. Likewise for Double F_2 data the estimating equation is

$$\frac{2 + P}{a_1} - \frac{1 - P}{a_2} - \frac{1 - P}{a_3} + \frac{P}{a_4} = 0, \tag{11 b}$$

where $P = (1 - p)^2$ or p^2, according as coupling or repulsion data are considered.

We shall now consider the relative performance of the maximum likelihood estimate, estimate from the product formula and the suggested estimate from equation (11) in controlling the errors in a few types of disturbed segregation, where product formula is preferable to the usual maximum likelihood estimate without taking into account the disturbed segregation. In the following the maximum likelihood estimate, estimate from the product formula and the estimate of p from (11) will be denoted by p_m, p_l and p_{-1} respectively.

Suppose at first that 100 α% of the Ab and ab classes are misclassified as AB and aB respectively owing to the failure of the recessive bb to manifest itself in 100 α% cases. Then we expect to find an observable segregation of

$$p + \alpha(1 + p)\,AB : (1 + \alpha)(1 + p)\,Ab : (1 - p) + \alpha p\,aB : (1 - \alpha)\,p\,ab \tag{I}$$

in Double Backcross repulsion data,

$$(2 + P) + \alpha(1 - P)\,AB : (1 - \alpha)(1 - P)\,Ab : (1 - P) + \alpha P\,aB : (1 - \alpha)\,P\,ab \tag{II}$$

in Double F_2 data.

It is easy to see that any selection among p_m, p_l and p_{-1} made from their relative performance in repulsion data will also hold good for coupling data. Hence these estimates are compared with the exact values in repulsion data only for $\alpha = \cdot10, \cdot40, \cdot80$ and $p = \cdot10, \cdot25, \cdot45$ as shown in Table I for Double Backcross and in Table II for Double F_2 data.

TABLE I

Values of p_m, p_l *and* p_{-1} *in Double Backcross repulsion data with* 100 $\alpha\%$ *cases of* ab *and* Ab *phenotypic classes are misclassified into* aB *and* AB

p \ α	·10	·40	·80	
·10	·140	·260	·420	p_m
	·132	·189	·236	p_l
	·125	·134	·114	p_{-1}
·25	·275	·350	·450	p_m
	·272	·308	·353	p_l
	·269	·286	·267	p_{-1}
·45	·455	·470	·487	p_m
	·454	·464	·472	p_l
	·454	·458	·457	p_{-1}

It will be seen that for both these cases p_{-1} is nearest the true value and is better than either p_m or p_l by a considerable margin for large α and small p. It may be mentioned here that it is not true that p_{-1} is the 'best' estimate among the entire class of estimating equations suggested above in the sense that it always leads to a value nearest the correct value of p. For instance, by taking $f(x) = (1 + x)^{-10}$, the corresponding estimate p_{-10} (say) in Double Backcross repulsion data for

(i) $\alpha = \cdot40$, $p = \cdot10$ is $\cdot100$,

(ii) $\alpha = \cdot40$, $p = \cdot25$ is $\cdot250$,

(iii) $\alpha = \cdot40$, $p = \cdot45$ is $\cdot450$,

(iv) $\alpha = \cdot80$, $p = \cdot25$ is $\cdot250$.

TABLE II

Values of p_m, p_l *and* p_{-1} *in Double* F_2 *repulsion data with* 100 α% *cases of* ab *and* Ab *phenotypic classes misclassified into* aB *and* AB

p \ α	·10	·40	·80	
·10	·103	·120	·333	p_m
	·101	·109	·117	p_l
	·102	·105	·102	p_{-1}
·25	·262	·281	·392	p_m
	·262	·265	·279	p_l
	·262	·260	·259	p_{-1}
·45	·452	·459	·478	p_m
	·451	·455	·460	p_l
	·451	·454	·452	p_{-1}

Thus p_{-10} practically eliminates the error due to such disturbed segregation. However, evaluation of p_{-10} is not so simple as that of p_{-1}.

There may be other kinds of disturbed segregations where p_{-1} may not lead to a value so near the true value for all α and p. For example, if 100 α% of AB and aB are misclassified as Ab and ab respectively owing to some disturbed segregation, then the observable segregation of

$$(1-\alpha)\,p\,AB : (1-p)+\alpha p\,Ab : (1-\alpha)(1-p)\,aB : p+\alpha(1-p)\,ab \qquad (\text{I}\,a)$$

in Double Backcross repulsion data, and

$$(1-\alpha)(2+P)\,AB : (1-P)+\alpha(2+P)\,Ab : (1-\alpha)(1-P)\,aB : P+\alpha(1-P)\,ab \qquad (\text{II}\,a)$$

in Double F_2 data, are expected. It is easy to see that the estimates p_m, p_l and p_{-1} in the case Ia are identical with those of case I as given in Table I, so that same conclusion regarding p_{-1} remains valid. However for Double F_2 repulsion data the estimates p_m, p_l and p_{-1} are different from those of Table II and these are shown in Table III.

TABLE III

Values of p_m, p_l *and* p_{-1} *in Double* F_2 *repulsion data with* 100 α% *cases of* aB *and* AB *phenotypic classes misclassified into* ab *and* Ab

p \ α	·10	·40	·80	
·10	·280	·417	·482	p_m
	·276	·391	·434	p_l
	·271	·324	·224	p_{-1}
·25	·338	·436	·486	p_m
	·336	·415	·448	p_l
	·330	·367	·305	p_{-1}
·45	·463	·484	·496	p_m
	·462	·478	·482	p_l
	·461	·468	·458	p_{-1}

It will be noticed that although p_{-1} is nearest the true value even in such cases, yet the departure from the correct value is not negligible for usual values of α and p. In any case p_{-1} may be considered definitely superior to p_l, *i.e.*, the estimate obtained from well-known product formula.

Further it is readily verified that when the disturbed segregation is solely due to the poor viability of the recessive *bb* classes the estimate p_{-1} will give an absolutely correct estimate similar to the estimates obtained from likelihood equation or the product formula.

Thus it is concluded that in most of the practical situations where disturbed segregation is expected it is preferable to use the simple estimate p_{-1} in place of the estimate derived from the commonly used product formula. One point worth mentioning here is that the fact that the so-called maximum likelihood estimate p_m is found to be worse than either p_l or p_{-1} does not mean that the method of maximum likelihood is inapplicable here. A correct application of maximum likelihood by allowing for the disturbances in the segregation will always lead to a correct estimate. However, in cases where a knowledge of the cause of disturbance is not possessed, thus necessitating the employment of an approximate method, the simple estimating equation (11) is to be preferred to p_m or p_l.

SUMMARY

It is shown that there exists a class of estimating equations besides the product formula which leads to asymptotically efficient estimates of the recombination fraction p used in mendelian genetics. Among these estimating equations the product formula has the special property that the corresponding estimating equation is expressible in a simple compact form. However, there is another equation within the suggested class which is very easy to solve and is also superior to the product formula in minimizing errors of estimation of recombination fraction in disturbed segregations.

REFERENCES

Cramér, H. .. *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

Fisher, R. A. and Balmakund, B. "The estimation of linkage from the offspring of selfed heterozygotes," *Jour. Genet.*, 1928, **20,** 79.

Huzurbazar, V. S. .. "The likelihood equation, consistency and the maxima of the likelihood function," *Ann. Eug.*, 1948, **14,** 185.

Immer, F. R. .. "Formulæ and tables for calculating linkage intensities," *Genetics*, 1930, **16,** 26.

Mather, K. .. *The Measurement of Linkage in Heredity*, Methuen & Co. Ltd., London, 1938.