METHODS OF SCORING LINKAGE DATA GIVING THE SIMULTANEOUS SEGREGATION OF THREE FACTORS

C. RADHAKRISHNA RAO

King's College, Cambridge and Presidency College, Calcutta

Received 28.v.47

CONTENTS

1. INTRODUCTION

WHEN estimating the values of a number of recombination fractions for various segments of a chromosome, from linkage data, it is often the case that more than one body of data is available. The data may relate to various sources, different types of crosses and consist of parts supplying information for one or more of the segments. In such cases, it is, obviously, desirable to combine the data in order to make joint estimation of the parameters. The estimates, thus obtained, are the most efficient when different parts of the data are homogeneous which can, however, be directly verified by testing whether the estimates closely fit in with the various parts of the data.

Fisher (1935) put forward a method of scoring which seems fitted for general use where combination of data, tests of homogeneity and pooled estimates are considered. This method consists of replacing each body of data by appropriate scores and information which are directly used in arriving at pooled estimates and tests of homogeneity. In a paper Fisher (1946) used this method of scoring in estimating three recombination fractions (arising out of two consecutive segments and subject to Kosambi's (1944) restriction) from data which in parts supply information only for the individual segments. If the data giving the simultaneous segregation of three or more factors are available then two questions arise. What is the most appropriate method of scoring such data and how does it compare with the alternative method of scoring for each recombination fraction by considering the classification with respect to the two relevant factors only and ignoring the rest? The latter method of scoring may be called the *individual segment method.*

When the simultaneous segregation of more than two factors has been recorded, the scores obtained by the individual segment method are less efficient in the sense in which Fisher (1935) defines efficient scores. The estimates to which they lead differ from the maximum likelihood estimates and hence are less efficient than the best estimates. The purpose of this article is to study these problems in the simplest case of scoring data on three factors arranged in eight phenotypical classes.

Firstly, a method appropriate and efficient for such data has been developed for the estimation of recombination fractions in the two different situations, viz. (1) when they are taken as free parameters and (2) when they conform to Kosambi's formula. Secondly, the relative efficiencies of the estimates in each of the two cases obtained by the individual segment method as compared with the above method have been calculated in a particular case to measure the loss in efficiency due to the individual segment method. The method of efficient scores involves a slightly heavier computational procedure and, though undoubtedly more efficient, can be recommended in practice only when even a small gain in efficiency is of considerable importance.

2. THEORETICAL ASPECTS OF THE METHOD OF SCORING AND NOTATIONS

Let there be k sets of data, each part supplying information on one or more of a set of p unknown parameters $\theta_1, \theta_2, \ldots, \theta_p$. If $L_1, L_2, \ldots L_k$ represent the probability densities of the observations corresponding to the k sets of data then L, the likelihood of the parameters, is defined by the product,

$$L = L_1 L_2 \ldots L_k$$

The best estimates of the parameters are those values of $\theta_1, \ldots, \theta_p$ which maximise the above expression. A formal proof of this statement is given elsewhere (Rao, 1947). For convenience the logarithm of L may be maximised. Differentiating log L partially with respect to $\theta_1, \ldots, \theta_p$ and equating to zero we get

$$\frac{\partial \log L}{\partial \theta_i} = \frac{\partial \log L_1}{\partial \theta_i} + \ldots + \frac{\partial \log L_k}{\partial \theta_i} = 0$$
$$i = 1, 2, \ldots, p$$

as the equations giving the desired estimates.

The above equations are usually non-linear and hence the direct evaluation of the estimates is difficult. A general method is to start with trial solutions and derive linear equations giving small corrections to the trial values. The process may have to be repeated until the corrections become negligible. If $d\theta_1, \ldots, d\theta_p$ are additive correc-

tions to trial values $\theta_1^0, \ldots, \theta_p^0$ one derives the equations giving the corrections as

$$\frac{-\partial^2 \log L}{\partial\theta_i\partial\theta_1} d\theta_1 - \frac{\partial^2 \log L}{\partial\theta_i\partial\theta_2} d\theta_2 - \ldots = \frac{\partial \log L}{\partial\theta_i}$$

$$i = 1, 2, \ldots, p$$

where all the derivatives are calculated at the trial values. If the sample is large we may replace $-\partial^2 \log L/\partial\theta_i\partial\theta_j$ by its mean value which is the same as I_{ij} the mean value of $\left(\frac{\partial \log L}{\partial\theta_i} \cdot \frac{\partial \log L}{\partial\theta_j}\right)$ at the trial values. The matrix $I = (I_{ij})$ is called the information matrix for the whole body of data at the assumed values.

The quantity $\partial \log L/\partial\theta_i$ is defined as the i-th efficient score and since

$$\frac{\partial \log L}{\partial\theta_i} = \frac{\partial \log L_1}{\partial\theta_i} + \ldots + \frac{\partial \log L_k}{\partial\theta_i}$$

it follows that the efficient scores for the whole data are the sums of the corresponding scores for the several sets of data. Similarly

$$I_{ij} = I_{ij}^{(1)} + \ldots + I_{ij}^{(k)}$$

where $I_{ij}^{(r)}$ is an element of the information matrix for the r-th part of the data.

Thus the problem of estimation reduces to replacing each part of the data by the efficient scores and information matrix at some trial values and finally obtaining the corresponding quantities for the whole data by simple addition. These supply linear equations in small additive corrections to the trial values.

It has been demonstrated in section 6 that the method of scoring offers a quickly converging process and hence is extremely useful in practice.

The scores for various parts of the data calculated at the best estimates are directly useful in tests of homogeneity as explained in sections 4 and 5 and illustrated in section 6. The theoretical aspects of such tests are fully discussed by the author in (Rao, 1948).

Notations.—

y_1, y_2, y_3 represent the recombination fractions in the segments connecting the second and third, third and first and first and second loci respectively.

$\dot{y}_1, \dot{y}_2, \dot{y}_3$ are the maximum likelihood estimates when they are treated as free parameters.

$\ddot{y}_1, \ddot{y}_2, \ddot{y}_3$ are the corresponding estimates when Kosambi's formula connecting y_1, y_2, y_3 is used.

$\phi_1^{(k)}$, $\phi_2^{(k)}$, $\phi_3^{(k)}$ are the efficient scores for γ_1, γ_2, γ_3 treated as free parameters from the k-th part of the data. When the data is arranged as frequencies in certain classes, it is known that

$$\phi_r^{(k)} = \mathrm{S}\left(\frac{n}{\pi}\frac{\partial \pi}{\partial \gamma_r}\right)$$

where S stands for summation, n for the frequency of a class and π for the corresponding probability. The total scores for the whole body of data are represented by Φ_1, Φ_2, Φ_3.

ψ_1, ψ_3 represent the scores for γ_1 and γ_3 when Kosambi's formula is used.

$(i_{rs}^{(k)})$ and $(\mathrm{I}_{rs}^{(k)})$ represent the information matrix per single observation and for the whole sample in the k-th part of the data. If n_k is the sample size then the relation $(\mathrm{I}_{rs}^{(k)}) = n_k(i_{rs}^{(k)})$ is identically true. It is known that,

$$i_{rs}^{(k)} = \mathrm{S}\left(\frac{1}{\pi}\frac{\partial \pi}{\partial \gamma_r}\right)\frac{\partial \pi}{\partial \gamma_s} = \mathrm{S}\frac{\partial \pi}{\partial \gamma_r}\left(\frac{1}{\pi}\frac{\partial \pi}{\partial \gamma_s}\right)$$

(T_{rs}) or simply denoted by $\mathbf{T}$ stands for the total information matrix $(\mathrm{T}_{rs}) = \mathrm{S}(\mathrm{I}_{rs}^{(k)})$.

The inverse of $\mathbf{T}$ is denoted by $\mathbf{T}^{-1}$. The method of finding this is to solve the equations

$$\begin{aligned} p\mathrm{T}_{11}+q\mathrm{T}_{12}+r\mathrm{T}_{13} &= 1, 0, 0 \\ p\mathrm{T}_{12}+q\mathrm{T}_{22}+r\mathrm{T}_{23} &= 0, 1, 0 \\ p\mathrm{T}_{13}+q\mathrm{T}_{23}+r\mathrm{T}_{33} &= 0, 0, 1 \end{aligned}$$

and take the three sets of equations as the three rows of $\mathbf{T}^{-1}$. $\mathbf{j}$ and $\mathbf{J}$ stand for information matrices per single observation and for the whole body of data when Kosambi's formula is used. All the above quantities calculated at $\dot{\gamma}_1$, $\dot{\gamma}_2$, $\dot{\gamma}_3$ and $\ddot{\gamma}_1$, $\ddot{\gamma}_2$, $\ddot{\gamma}_3$ are represented with a single dot and two dots above respectively.

3. SCORING OF DATA FROM BACK CROSSES AND INTERCROSSES

The eight different gametes due to a triply heterozygous parent can be classified into four complementary pairs, the members of each pair having equal chance of being transmitted to an offspring. Let the recombination fractions for the segments connecting the first and second, second and third and third and first loci be represented by γ_3, γ_1 and γ_2. Also let S_0, S_1, S_2 and S_3 represent respectively the frequencies of gametes involving all three recessive genes, only the first, only the second and only the third recessive gene. The S's and γ's are interrelated in the four types of heterozygotes as given in table 1.

TABLE 1

S functions for the four types of triple heterozygotes

Function of y's	Gametic frequencies for heterozygotes			
	$\frac{ABC}{abc}$	$\frac{AbC}{aBc}$	$\frac{ABc}{abC}$	$\frac{aBC}{Abc}$
$\frac{1}{4}(2-y_3-y_1-y_2)$	S_0	S_2	S_3	S_1
$\frac{1}{4}(y_2+y_1-y_3)$	S_3	S_1	S_0	S_2
$\frac{1}{4}(y_3+y_1-y_2)$	S_2	S_0	S_1	S_3
$\frac{1}{4}(y_3+y_2-y_1)$	S_1	S_3	S_2	S_0

For a given set of recombination fractions one can, by using the above table, calculate the S-functions for any type of heterozygote and use them for further calculations as the probabilities, scores and information matrix, in any particular situation directly depend on them.

(a) Triple back cross

The probabilities of the eight phenotypical classes and the appropriate scores, expressed in terms of S-functions, in the case of a back cross of a triple heterozygote with a triple recessive are given in table 2.

TABLE 2

Probabilities and scores in the case of a triple back cross

Phenotype	Probability	Scores for y_1	Scores for y_2	Scores for y_3	Observed frequency
ABC	S_0	$l_0/4S_0$	$m_0/4S_0$	$n_0/4S_0$	n_{111}
AbC	S_2	$l_2/4S_2$	$m_2/4S_2$	$n_2/4S_2$	n_{101}
ABc	S_3	$l_3/4S_3$	$m_3/4S_3$	$n_3/4S_3$	n_{110}
Abc	S_1	$l_1/4S_1$	$m_1/4S_1$	$n_1/4S_1$	n_{100}
aBC	S_1	$l_1/4S_1$	$m_1/4S_1$	$n_1/4S_1$	n_{011}
abC	S_3	$l_3/4S_3$	$m_3/4S_3$	$n_3/4S_3$	n_{001}
aBc	S_2	$l_2/4S_2$	$m_2/4S_2$	$n_2/4S_2$	n_{010}
abc	S_0	$l_0/4S_0$	$m_0/4S_0$	$n_0/4S_0$	n_{000}

The l's, m's and n's of the above table are four times the derivatives of probabilities with respect to y_1, y_2 and y_3 respectively and their values are determined by the following rule. The value of the derivative of the probability for a phenotype with respect to y_i is $-\frac{1}{4}$ or $\frac{1}{4}$ according as the letters other than the i-th in the representation of the phenotype is an old or a new combination, the old combinations being determined by the representation of the triple heterozygote. Thus in deciding the values of l_0, l_2, l_3 and l_1, one need only find which of the combinations BC, bC, Bc, bc, are old and which are new and take the l's corresponding to old combinations as -1 and the rest as $+1$. The values of these coefficients for the various types of heterozygotes are given in table 3 for ready use.

The efficient scores at any given values of y_1, y_2 and y_3 are obtained by summing the products of observed frequencies and scores. The

TABLE 3

Values of the l's, m's *and* n's *of table 2*

Heterozygote	Values of the coefficients
ABC/abc	$l_0 = l_1 = m_0 = m_2 = n_0 = n_3 = -1$
	$l_2 = l_3 = m_1 = m_3 = n_1 = n_2 = 1$
AbC/aBc	$l_2 = l_3 = m_0 = m_2 = n_1 = n_2 = -1$
	$l_0 = l_1 = m_1 = m_3 = n_0 = n_3 = 1$
Abc/aBC	$l_0 = l_1 = m_1 = m_3 = n_1 = n_2 = -1$
	$l_2 = l_3 = m_0 = m_2 = n_0 = n_3 = 1$
ABc/abC	$l_2 = l_3 = m_1 = m_3 = n_0 = n_3 = -1$
	$l_0 = l_1 = m_0 = m_2 = n_1 = n_2 = 1$

elements of the information matrix (i_{rs}) per single observation for given values of y_1, y_2, y_3 are the same for all types of back crosses. They can be simply calculated by using the formulæ

$$\begin{aligned} i_{rr} &= \tfrac{1}{2}(\lambda_0+\lambda_1+\lambda_2+\lambda_3) \quad r = 1, 2, 3 \\ i_{12} &= \tfrac{1}{2}(\lambda_0-\lambda_1-\lambda_2+\lambda_3) \\ i_{23} &= \tfrac{1}{2}(\lambda_0+\lambda_1-\lambda_2-\lambda_3) \\ i_{13} &= \tfrac{1}{2}(\lambda_0-\lambda_1+\lambda_2-\lambda_3) \end{aligned}$$

where $\lambda_0 = 1/(2-y_1-y_2-y_3)$, $\lambda_1 = 1/(y_2+y_3-y_1)$, $\lambda_2 = 1/(y_1+y_3-y_2)$ and $\lambda_3 = 1/(y_1+y_2-y_3)$. It may be noted that the λ's are fixed functions of the recombination fractions and any S-function assumes one of the values of λ for a given heterozygote. These have been introduced merely to simplify computations.

If the whole data consist only of results from back crosses the maximum likelihood estimates can be obtained without the evaluation of the efficient scores. It is easily seen that for such data the equations giving the best estimates are

$$\begin{aligned} (0)\lambda_0+(1)\lambda_1-(12)\lambda_2-(2)\lambda_3 &= 0 \\ (0)\lambda_0-(1)\lambda_1+(12)\lambda_2-(2)\lambda_3 &= 0 \\ (0)\lambda_0-(1)\lambda_1-(12)\lambda_2+(2)\lambda_3 &= 0 \end{aligned}$$

where (0), (1), (2) and (12) are the totals of observed frequencies from all sets of data corresponding to no cross overs, cross overs in the first segment only, cross overs in the second segment only and double cross overs respectively.

These equations yield the solutions

$$(0)\lambda_0 = (1)\lambda_1 = (12)\lambda_2 = (2)\lambda_3$$

or, writing in terms of the y's, they become

$$\frac{(0)}{2-y_1-y_2-y_3}=\frac{(1)}{y_2+y_3-y_1}=\frac{(12)}{y_1+y_3-y_2}=\frac{(2)}{y_1+y_2-y_3}=$$
$$\frac{(1)+(12)}{2y_3}=\frac{(2)+(12)}{2y_1}=\frac{(1)+(2)}{2y_2}=\frac{(0)+(1)+(2)+(12)}{2}=\frac{N}{2}.$$

Hence $\dot{y}_1=\frac{(2)+(12)}{N}$, $\dot{y}_2=\frac{(1)+(2)}{N}$, $\dot{y}_3=\frac{(1)+(12)}{N}$.

The above estimates for y_1, y_2 and y_3 are the same as those obtained by considering the data as classified according two factors each time. Thus $\dot{y}_1$ could be obtained by considering only the second and third factors and ignoring the classification due to the first factor. Thus a complete classification with respect to three factors in the case of back crosses does not supply additional information so far as the problem of estimation of recombination fractions in various segments is concerned. This is true with more than three factors also.

The variances and covariances of the above estimates are

$$V(\dot{y}_1)=\frac{y_1(1-y_1)}{N},\quad V(\dot{y}_2)=\frac{y_2(1-y_2)}{N},\quad V(\dot{y}_3)=\frac{y_3(1-y_3)}{N}$$

$$\text{Cov.}\,(\dot{y}_1\dot{y}_2)=\frac{y_1+y_2-y_3-2y_1y_2}{2N},\qquad \text{Cov.}\,(\dot{y}_2\dot{y}_3)=\frac{y_2+y_3-y_1-2y_2y_3}{2N}$$

$$\text{Cov.}\,(\dot{y}_1\dot{y}_3)=\frac{y_1+y_3-y_2-2y_1y_3}{2N}$$

(b) *Intercross of a triple heterozygote*

The frequencies and their derivatives in terms of S-functions for any type of intercross are given in table 4.

TABLE 4

Frequencies and their derivatives for an intercross

Phenotype	Probability	Derivatives		
	π	$\frac{\partial\pi}{\partial y_1}$	$\frac{\partial\pi}{\partial y_2}$	$\frac{\partial\pi}{\partial y_3}$
ABC	$\frac{1}{4}+S_0+S_1^2+S_2^2+S_3^2$	$l_0(S_0+2S_1)$	$m_0(S_0+2S_2)$	$n_0(S_0+2S_3)$
AbC	$S_2-S_2^2+2S_1S_3$	$l_2(S_0+2S_1)$	m_2S_0	$n_2(S_0+2S_3)$
ABc	$S_3-S_3^2+2S_1S_2$	$l_3(S_0+2S_1)$	$m_3(S_0+2S_2)$	n_3S_0
Abc	$S_1^2+2S_1S_0$	$l_1(S_0+2S_1)$	m_1S_0	n_1S_0
aBC	$S_1-S_1^2+2S_2S_3$	l_1S_0	$m_1(S_0+2S_2)$	$n_1(S_0+2S_3)$
abC	$S_3^2+2S_3S_0$	l_3S_0	m_3S_0	$n_3(S_0+2S_3)$
aBc	$S_2^2+2S_2S_0$	l_2S_0	$m_2(S_0+2S_2)$	n_2S_0
abc	S_0^2	l_0S_0	m_0S_0	n_0S_0

The l's in the column for $\partial\pi/\partial y_1$ are $-\frac{1}{2}$ or $\frac{1}{2}$ according as the factors other than the first in the representation of the corresponding

phenotypes form old or new combinations. These can be readily obtained from table 3 by replacing unity by $\frac{1}{2}$. The appropriate scores, which are used in the calculation of efficient scores, for any class, are calculated by dividing the derivatives by the class probability. The information matrix is evaluated by the formulæ

$$I_{rr} = nS\left(\frac{1}{\pi}\frac{\partial\pi}{\partial y_r}\right)\frac{\partial\pi}{\partial y_r}$$

$$I_{rk} = nS\left(\frac{1}{\pi}\frac{\partial\pi}{\partial y_r}\right)\frac{\partial\pi}{\partial y_k} = nS\,\frac{\partial\pi}{\partial y_r}\left(\frac{1}{\pi}\frac{\partial\pi}{\partial y_k}\right)$$

where n is the total of observed frequencies. The maximum likelihood estimates, in this case, are obtained by successive approximations as explained in the next section.

4. MAXIMUM LIKELIHOOD ESTIMATES OF THE RECOMBINATION FRACTIONS TAKEN AS FREE PARAMETERS

When the data from various sources and different types of crosses giving information on only one or all the three recombination fractions are available, there arise the problems of obtaining the best estimates from the combined data and testing homogeneity of different parts of the data. The numerical computation of the above problems can be arranged as follows. To start with the scores and the information matrix, at the approximate values of y_1, y_2, y_3 are calculated for separate portions of the data as shown in table 5.

TABLE 5

Efficient scores and information matrix at the approximate values

Source and type of cross	Efficient scores			Information matrix					
1	$\phi_1^{(1)}$	$\phi_2^{(1)}$	$\phi_3^{(1)}$	$I_{11}^{(1)}$	$I_{12}^{(1)}$	$I_{13}^{(1)}$	$I_{22}^{(1)}$	$I_{23}^{(1)}$	$I_{33}^{(1)}$
.	.	.	.	.	.	.	.	.	.
k	$\phi_1^{(k)}$	$\phi_2^{(k)}$	$\phi_3^{(k)}$	$I_{11}^{(k)}$	$I_{12}^{(k)}$	$I_{13}^{(k)}$	$I_{22}^{(k)}$	$I_{23}^{(k)}$	$I_{33}^{(k)}$
Total	Φ_1	Φ_2	Φ_3	T_{11}	T_{12}	T_{13}	T_{22}	T_{23}	T_{33}

In the above table, if any part of the data gives information for only one segment, say the first, then only ϕ_3 and I_{33} are present and the rest are zero for that part. The methods of scoring in such cases have been fully discussed by Fisher (1946) and Bhat (1947) and also illustrated in this article in section 7 (*c*). For scoring parts of the data giving the simultaneous segregation of three factors the expressions derived in section 3 may be used.

Using the totals of table 5, the three linear equations in dy_1, dy_2 and dy_3, which are additive corrections to the first approximations of y_1, y_2 and y_3 respectively, can be written

$$\begin{aligned} T_{11}\,dy_1 + T_{12}\,dy_2 + T_{13}\,dy_3 &= \Phi_1 \\ T_{12}\,dy_1 + T_{22}\,dy_2 + T_{23}\,dy_3 &= \Phi_2 \\ T_{13}\,dy_1 + T_{23}\,dy_2 + T_{33}\,dy_3 &= \Phi_3 \end{aligned}$$

These calculations require to be repeated until the values of y_1, y_2 and y_3 for which the total scores are negligible are obtained.

If the scores and information matrix for the i-th part of the data at the estimated values are represented by

$$\dot{\phi}_1^{(i)},\ \dot{\phi}_2^{(i)},\ \dot{\phi}_3^{(i)} \text{ and } (\dot{\mathrm{I}}_{rs}^{(i)})$$

then using the elements of the matrix $(\dot{\mathrm{I}}_{(i)}^{rs})$ reciprocal to $(\dot{\mathrm{I}}_{rs}^{(i)})$ a χ^2 can be calculated by the formula

$$\chi_i^2 = \mathrm{SS}\, \dot{\mathrm{I}}_{(i)}^{rs}\, \dot{\phi}_r^{(i)}\, \dot{\phi}_s^{(i)}$$

If the i-th part of the data concerns only two of the factors, say the first and second, then χ_i^2 is simply $[\dot{\phi}_3^{(i)}]^2/\dot{\mathrm{I}}_{33}^{(i)}$. To test for heterogeneity the total

$$\chi^2 = \chi_1^2 + \ \ldots\ + \chi_k^2$$

can be used as χ^2 with degrees of freedom

$$d_1 + \ \ldots\ + d_k - 3$$

where $d_i = 1$ or 3 according as the i-th part of the data relates to the segregation of only two or three factors. A significant χ^2 at a chosen probability level indicates that different parts of the data are not homogeneous.

5. MAXIMUM LIKELIHOOD ESTIMATES SUBJECT TO KOSAMBI'S FORMULA

The general formula $\tanh 2x = 2y$ giving the relation between the map distance x and the recombination fraction y is given by Kosambi (1944). In terms of y_1, y_2 and y_3, the relation becomes

$$y_2 = \frac{y_1 + y_3}{1 + 4y_1y_3}$$

There are only two parameters y_1 and y_3 to be estimated, the third one, y_2, being obtainable from the above formula. The scores and information matrix for y_1 and y_3 can be conveniently calculated from the corresponding expressions for y_1, y_2 and y_3 taken as free parameters and in any practical problem the data may be scored for y_1, y_2 and y_3 taken as free parameters and then the appropriate scores for y_1 and y_3, when Kosambi's formula is applicable, can be deduced. If ψ_1 and ψ_3 represent the total scores for y_1 and y_3 and J_{11}, J_{13}, J_{33}, the elements of the information matrix then the formulæ connecting them with the totals of table 5 are

$$\psi_1 = \Phi_1 + \mu_1\Phi_2 \quad , \quad \mu_1 = (1 - 4y_3^2)/(1 + 4y_1y_3)^2$$
$$\psi_3 = \Phi_3 + \mu_3\Phi_2 \quad , \quad \mu_3 = (1 - 4y_1^2)/(1 + 4y_1y_3)^2$$
$$\mathrm{J}_{11} = \mathrm{T}_{11} + 2\mu_1\mathrm{T}_{12} + \mu_1^2\mathrm{T}_{22}$$
$$\mathrm{J}_{13} = \mathrm{T}_{13} + \mu_1\mathrm{T}_{23} + \mu_3\mathrm{T}_{12} + \mu_1\mu_3\mathrm{T}_{22}$$
$$\mathrm{J}_{33} = \mathrm{T}_{33} + 2\mu_3\mathrm{T}_{23} + \mu_3^2\mathrm{T}_{22}$$

The additive corrections dy_1 and dy_3 to approximate values of y_1 and y_3 are obtained from the equations

$$\begin{aligned} J_{11}\ dy_1 + J_{13}\ dy_3 &= \psi_1 \\ J_{13}\ dy_1 + J_{33}\ dy_3 &= \psi_3 \end{aligned}$$

If $\ddot{\Phi}_2$ is the total score for y_2 and $(\ddot{T}^{rs})$ is the inverse of the total information matrix calculated as in table 5 for y_1, y_2, y_3 at the estimated values $\ddot{y}_1, \ddot{y}_2, \ddot{y}_3$, subject to Kosambi's formula, then

$$\chi^2 = \ddot{\Phi}_2^2(\ddot{T}^{22} + \mu_1^2\ddot{T}^{11} + \mu_3^2\ddot{T}^{33} + 2\mu_1\mu_3\ddot{T}^{13} - 2\mu_1\ddot{T}^{12} - 2\mu_3\ddot{T}^{23})$$

can be used as χ^2 with 1 d.f. to test the agreement with Kosambi's formula. The estimates, obtained subject to Kosambi's formula, are valid only when the above test does not show significance.

6. AN ILLUSTRATIVE EXAMPLE

The methods developed in sections 3, 4 and 5 are applied to the linkage data for *Primula sinensis* reproduced from a paper by de Winton and Haldane (1935).

The data chosen consist of two types of back crosses so that exact values of maximum likelihood estimates could be obtained. The method of scoring can also be applied, starting with some trial values, and the first approximations which, in this case, must be identical with the exact values may be obtained.

The trial values chosen are $y_1 = \cdot 35$, $y_2 = \cdot 39$, $y_3 = \cdot 07$ so that,

$$\begin{aligned} \lambda_0 &= 1/(2 - y_1 - y_2 - y_3) = 1/1\cdot 19 = \cdot 840336 \\ \lambda_1 &= 1/(y_2 + y_3 - y_1) = 1/\cdot 11 = 9\cdot 090909 \\ \lambda_2 &= 1/(y_1 + y_3 - y_2) = 1/\cdot 03 = 33\cdot 333333 \\ \lambda_3 &= 1/(y_1 + y_2 - y_3) = 1/\cdot 67 = 1\cdot 492537 \end{aligned}$$

The λ's are one-fourth of the reciprocals of the S-functions so that for any back cross the scores given in table 2 are obtained by multiplying the λ's with the appropriate values of the l's, m's and n's given in table 3.

TABLE 6

Linkage data of SBL in Primula sinensis (♀ *side*)

Set I $\dfrac{\text{sbl}}{\text{sbl}} \times \dfrac{\text{SBL}}{\text{sbl}}$

Phenotype	Observed frequency	Type of cross over	Scores for y_1	Scores for y_2	Scores for y_3	Multiplier
SBL	457	(0)	−	−	−	$\cdot 840336 = 1/4S_0$
SbL	11	(12)	+	−	+	$33\cdot 333333 = 1/4S_2$
SBl	256	(2)	+	+	−	$1\cdot 492537 = 1/4S_3$
Sbl	38	(1)	−	+	+	$9\cdot 090909 = 1/4S_1$
sBL	45	(1)	−	+	+	$9\cdot 090909 = 1/4S_1$
sbL	284	(2)	+	+	−	$1\cdot 492537 = 1/4S_3$
sBl	20	(12)	+	−	+	$33\cdot 333333 = 1/4S_2$
sbl	469	(0)	−	−	−	$\cdot 840336 = 1/4S_0$
	1580					

Set II $\frac{sbl}{sbl} \times \frac{SBl}{sbL}$

Phenotype	Observed frequency	Type of cross over	Scores for y_1	y_2	y_3	Multiplier
SBL	21	(2)	+	+	−	1·492537 = 1/4S_0
SbL	3	(1)	−	+	+	9·090909 = 1/4S_2
SBl	50	(0)	−	−	−	·840336 = 1/4S_3
Sbl	1	(12)	+	−	+	33·333333 = 1/4S_1
sBL	1	(12)	+	−	+	33·333333 = 1/4S_3
sbL	57	(0)	−	−	−	·840336 = 1/4S_3
sBl	4	(1)	−	+	+	9·090909 = 1/4S_2
sbl	26	(2)	+	+	−	1·492537 = 1/4S_0
	163					

The exact values from the combined data are given by

$$\dot{y}_1 = \frac{(2)+(12)}{N} = \frac{(256+284+11+20)+(21+1+1+26)}{1580+163}$$

$$= \frac{620}{1743} = \cdot 355709$$

$$\dot{y}_2 = \frac{(1)+(2)}{N} = \frac{677}{1743} = \cdot 388411$$

$$\dot{y}_3 = \frac{(1)+(12)}{N} = \frac{123}{1743} = \cdot 070568$$

The efficient scores at the trial values.

Set I	306·606720	−250·969032	203·757654
Set II	−16·736410	−22·797016	−29·762162
Total =	289·870310	−273·766048	173·995492

If the data contained some parts giving only two factors segregations then the scores arising from them have to be added to the above values. Such data can supply only one of the efficient scores, the others being considered as zero. In such cases the appropriate scores for intercrosses are given on p. 61 and for back crosses on p. 58 in Mather's book, *The Measurement of Linkage in Heredity* (first edition).

The information matrix per single observation is the same for both the above sets at the trial values.

$$i_{11} = i_{22} = i_{33} = \tfrac{1}{2}(\lambda_0+\lambda_1+\lambda_2+\lambda_3) = 22\cdot 378557$$
$$i_{12} = \tfrac{1}{2}(\lambda_0-\lambda_1-\lambda_2+\lambda_3) = -20\cdot 045685$$
$$i_{23} = \tfrac{1}{2}(\lambda_0+\lambda_1-\lambda_2-\lambda_3) = -12\cdot 447312$$
$$i_{13} = \tfrac{1}{2}(\lambda_0-\lambda_1+\lambda_2-\lambda_3) = 11\cdot 795111$$

The total information matrix T is

$$\mathbf{T} = (1580+163)\begin{pmatrix} 22\cdot 378557 & -20\cdot 045685 & 11\cdot 795111 \\ -20\cdot 045685 & 22\cdot 378557 & -12\cdot 447312 \\ 11\cdot 795111 & -12\cdot 447312 & 22\cdot 378557 \end{pmatrix}$$

and its inverse is

$$\mathbf{T}^{-1} = \frac{1}{1743}\begin{pmatrix} \cdot 227500 & \cdot 198500 & -\cdot 009500 \\ \cdot 198500 & \cdot 237900 & \cdot 027700 \\ -\cdot 009500 & \cdot 027700 & \cdot 065100 \end{pmatrix}$$

The corrections to trial values are given by

$$dy_i = \Phi_1 T^{i1} + \Phi_2 T^{i2} + \Phi_3 T^{i3}$$

$$dy_1 = \frac{1}{1743}[\cdot 2275(289\cdot 870310) + \cdot 1985(-273\cdot 766048) - \cdot 0095(173\cdot 995492)] = \cdot 00570953$$

$$dy_2 = \frac{1}{1743}[\cdot 1985(\quad) + \cdot 2379(\quad) + \cdot 0277(\quad)] = -\cdot 00158922$$

$$dy_3 = \frac{1}{1743}[-\cdot 0095(\quad) + \cdot 0277(\quad) + \cdot 0651(\quad)] = \cdot 000567990$$

The first approximations to trial values ·35, ·39, ·07, are

·355709, ·388411, ·070568

agreeing accurately up to the number of significant figures maintained in the calculations with the exact estimates

·355709, ·388411, ·070568

obtained earlier. This is not generally true but if the trial values of the cross over percentages can be correctly guessed to the nearest whole numbers the first approximations obtained by the method of scoring are expected to be sufficiently accurate.

The test of the hypothesis that the two sets of data arise from identical values of recombination fractions involves the evaluation of the scores and information matrices at the estimated values. The new scores can, however, be obtained approximately by certain adjustments of the scores at the trial values. The change in the matrix inverse to the information matrix is negligible so that no adjustment is necessary. But in cases where approximations differ considerably from the trial values it is necessary to calculate the scores and information matrix directly.

The adjusted scores for the first set are

$$\dot{\phi}_1^{(1)} = \phi_1^{(1)} - n_1(dy_1 i_{11} + dy_2 i_{12} + dy_3 i_{13})$$
$$= 289\cdot 870310 - 1580[22\cdot 378557(\cdot 005709) - 20\cdot 045685(-\cdot 001589) + 11\cdot 795111(\cdot 000568)]$$
$$= 289\cdot 870310 - 262\cdot 771854 = 27\cdot 098456$$

$$\dot{\phi}_2^{(1)} = -273\cdot 766048 - 1580[-20\cdot 045685(\quad) + 22\cdot 378557(\quad) - 12\cdot 447312(\quad)]$$
$$= -273\cdot 766048 + 248\cdot 171256 = -25\cdot 594792$$

$$\dot{\phi}_3^{(1)} = 173\cdot 995492 - 1580[11\cdot 795111(\quad) - 12\cdot 447312(\quad) + 22\cdot 378557(\quad)]$$
$$= 173\cdot 995492 - 157\cdot 728377 = 16\cdot 267115$$

The value of χ_1^2 is,

$$= \frac{1}{n_1} SSi^{rs}\dot{\phi}_r^{(1)}\dot{\phi}_s^{(1)}$$

$$= \frac{1}{1580}[\cdot 2275(27\cdot 0984)^2 + \cdot 2379(25\cdot 5948)^2 + \cdot 0651(16\cdot 2671)^2$$
$$+2(\ \cdot 1985)(\ 27\cdot 0984)(-25\cdot 5948)$$
$$+2(-\cdot 0095)(\qquad)(\ 16\cdot 2671)$$
$$+2(\ \cdot 0277)(-25\cdot 5948)(\qquad)]$$

$$= \frac{33\cdot 3807}{1580} = \cdot 02113.$$

Similarly

$\dot{\phi}_1^{(2)} = -27\cdot 098456$, $\dot{\phi}_2^{(2)} = 25\cdot 594792$, $\dot{\phi}_3^{(2)} = -16\cdot 267115$ and $\chi_2^2 = \frac{33\cdot 3807}{163} = \cdot 2048.$

The total $\chi^2 = \chi_1^2 + \chi_2^2 = \cdot 0211 + \cdot 2048 = \cdot 2259$ is considerably smaller than its expectation $3+3-3 = 3$, the degrees of freedom. The two sets of data may be regarded as homogeneous.

The variances of the estimates from the combined data are given by the diagonal elements of the matrix $\mathbf{T}^{-1}$.

$$V(\dot{y}_1) = \cdot 2275/1743 = 10^{-3}(\cdot 130522)$$
$$V(\dot{y}_2) = \cdot 2379/1743 = 10^{-3}(\cdot 136489)$$
$$V(\dot{y}_3) = \cdot 0651/1743 = 10^{-3}(\cdot 037349)$$

These are only approximate values, the exact values being obtainable from the inverse of the information matrix calculated at the estimated values.

If Kosambi's formula is assumed, there are only two parameters y_1 and y_3 to be estimated. The appropriate scores and information matrix in this case are calculated from the values obtained before for y_1, y_2 and y_3 considering them as free parameters. The trial values of y_1, y_2 and y_3 are the same as before. To start with calculate

$$\mu_1 = \frac{1-4y_3^2}{(1+4y_1y_3)^2}, \quad \mu_3 = \frac{1-4y_1^2}{(1+4y_1y_3)^2}$$
$$= \frac{\cdot 980400}{1\cdot 205604}, \qquad = \frac{\cdot 510000}{1\cdot 205604}$$
$$= \cdot 813202\ , \qquad = \cdot 423024$$
$$\mu_1^2 = \cdot 661297\ , \quad \mu_1\mu_3 = \cdot 344004\ , \quad \mu_3^2 = \cdot 178949.$$

The efficient scores ψ_1 and ψ_3 are,

$$\psi_1 = \Phi_1 + \mu_1\Phi_2$$
$$= 289\cdot 870310 + \cdot 813202(-273\cdot 766048)$$
$$= 67\cdot 243211$$
$$\psi_3 = \Phi_3 + \mu_3\Phi_2$$
$$= 173\cdot 995492 + \cdot 423024(-273\cdot 766048)$$
$$= 58\cdot 185880$$

Information matrix for y_1, y_3.

j_{rs} =	j_{11}	j_{13}	j_{33}
$+t_{rs}$	22·378557	11·795111	22·378557
	t_{11}	t_{13}	t_{33}
$+\mu_r\mu_s t_{22}$	14·798873	7·698313	4·004620
	$\mu_1^2 t_{22}$	$\mu_1\mu_3 t_{22}$	$\mu_3^2 t_{22}$
$+\mu_r t_{s2}$	−16·301191	−8·479806	−5·265512
	$\mu_1 t_{12}$	$\mu_3 t_{12}$	$\mu_3 t_{23}$
$+\mu_s t_{r2}$	−16·301191	−10·122179	−5·265512
	$\mu_1 t_{12}$	$\mu_1 t_{23}$	$\mu_3 t_{23}$
Total	4·575048	·891439	15·852153

$$J = 1743\begin{pmatrix} 4\cdot575048 & \cdot891439 \\ \cdot891439 & 15\cdot852153 \end{pmatrix}$$

$$J^{-1} = \frac{1}{1743}\begin{pmatrix} \cdot220998 & -\cdot012428 \\ -\cdot012428 & \cdot063782 \end{pmatrix}$$

The additive corrections to y_1 and y_3 are given by

$$dy_1 = \frac{1}{1743}[\cdot220998(67\cdot243211) - \cdot012428(58\cdot185880)]$$
$$= 14\cdot137481/1743 = \cdot00811100$$
$$dy_3 = \frac{1}{1743}[-\cdot012428(\quad) + \cdot063782(\quad)]$$
$$= 2\cdot875513/1743 = \cdot000164975$$

so that the estimates of y_1 and y_3 are

$$\cdot358111 \text{ and } \cdot070165$$

and the estimate of y_2 as obtained from the formula is

$$\ddot{y}_2 = (\ddot{y}_1 + \ddot{y}_3)/(1 + 4\ddot{y}_1\ddot{y}_3)$$
$$= (\cdot428276)/1\cdot100507 = \cdot389162$$

which differ very slightly from the estimates obtained by treating the y's as free parameters. The goodness of fit of Kosambi's formula to the cross over values indicated by the data may be tested as follows.

This needs the evaluation of the total efficient score $\ddot{\Phi}_2$ at these estimated values. As before, a good approximation to this value can be obtained by using the formula

$$\ddot{\Phi}_2 = \Phi_2 - N[i_{21}(\ddot{y}_1 - \cdot35) + i_{22}(\ddot{y}_2 - \cdot39) + i_{23}(\ddot{y}_3 - \cdot07)$$
$$= -273\cdot766048 - 1743[-20\cdot045685(\cdot00811 1)$$
$$+ 22\cdot378557(-\cdot000838) - 12\cdot447312(\cdot000164975)]$$
$$= -273\cdot766048 + 319\cdot661912$$
$$= \quad 45\cdot895864$$

$$1/V(\ddot{\Phi}_2) = \frac{1}{N}[i^{22} + \mu_1^2 i^{11} + \mu_3^2 i^{33} + 2\mu_1\mu_3 i^{13} - 2\mu_1 i^{12} - 2\mu_3 i^{23}]$$
$$= 10^{-4}(\cdot270694).$$

The quantity

$$\chi^2 = (\ddot{\Phi}_2)^2/V(\ddot{\Phi}_2) = (45{\cdot}895864)^2 \times 10^{-4}({\cdot}270694) = {\cdot}0570$$

can be used as χ^2 with one degree of freedom to test Kosambi's formula. The observed χ^2 is very small, thus indicating fair agreement with Kosambi's formula.

The variances of the estimates obtained by using Kosambi's formula are,

$$V(\ddot{y}_1) = \frac{j^{11}}{N} = \frac{{\cdot}220998}{1743} = 10^{-3}({\cdot}126791)$$

$$V(\ddot{y}_3) = \frac{j^{33}}{N} = \frac{{\cdot}063782}{1743} = 10^{-3}({\cdot}036593)$$

$$V(\ddot{y}_2) = \frac{\mu_1^2 j^{11} + 2\mu_1\mu_3 j^{13} + \mu_3^2 j^{33}}{N} = \frac{{\cdot}149008}{1743} = 10^{-3}({\cdot}085489).$$

By comparing these variances with those obtained before, we find the percentage increase in efficiency by using Kosambi's formula as

$$2{\cdot}94,\ 59{\cdot}66,\ 2{\cdot}06$$

for y_1, y_2 and y_3 respectively.

7. THE INCREASE IN PRECISION BY THE USE OF EFFICIENT SCORES

In the previous sections, methods have been developed for the appropriate scoring of data relating to the simultaneous segregation of three factors. It is, however, of importance to calculate the gain in efficiency by following this method instead of replacing the data by three marginal distributions obtained by ignoring one factor each time and treating the distributions as independent. In such a case the data can be scored by considering only one segment at a time. It is seen in section 3 that this makes no difference in the case of back crosses. The general investigation of this problem in the case of F_2 intercrosses is difficult but an example may be considered to give an idea of the increase in efficiency.

Let the data consist of the results of the F_2 of $\frac{AbC}{aBc}$ classified in eight phenotypical classes. The variances and covariances of the estimates, from the data of the above type of y_1, y_2, y_3 considered as (1) free parameters and (2) subject to Kosambi's formula, can be calculated in each case by using (*a*) the method of efficient scores and (*b*) the individual segment method, and the relative efficiency of the method (*b*) in each case can be found out. The following are the calculations in the particular case wherein the recombination fractions are

$$y_3 = {\cdot}03, y_1 = {\cdot}28 \text{ and } y_2 = {\cdot}30.$$

(a) Method of efficient scores taking y's as free parameters

Using table 4, the appropriate S-functions for the triple heterozygote AbC/aBc are calculated at the above chosen values.

$$\begin{aligned} S_3 &= \tfrac{1}{4}(y_2+y_3-y_1) &&= \cdot 0125 \\ S_2 &= \tfrac{1}{4}(2-y_3-y_2-y_1) &&= \cdot 3475 \\ S_1 &= \tfrac{1}{4}(y_2+y_1-y_3) &&= \cdot 1375 \\ S_0 &= \tfrac{1}{4}(y_3+y_1-y_2) &&= \cdot 0025 \\ & && \cdot 5000 \text{ (check)} \end{aligned}$$

The values of the l's, m's, n's of table 3 with 1 replaced by $\frac{1}{2}$ are

$$\begin{aligned} l_2 = l_3 = m_2 = m_0 = n_2 = n_1 &= -\tfrac{1}{2} \\ l_1 = l_0 = m_1 = m_3 = n_0 = n_3 &= \tfrac{1}{2} \end{aligned}$$

The probabilities, derivatives and scores are given in table 7.

TABLE 7

Probabilities and scores

Phenotype	Probability 16π	Derivatives and scores					
		$16\frac{\partial \pi}{\partial y_1}$	$\frac{1}{\pi}\frac{\partial \pi}{\partial y_1}$	$16\frac{\partial \pi}{\partial y_2}$	$\frac{1}{\pi}\frac{\partial \pi}{\partial y_2}$	$16\frac{\partial \pi}{\partial y_3}$	$\frac{1}{\pi}\frac{\partial \pi}{\partial y_3}$
ABC	6·2771	2·22	·353666	−6·38	−·888944	·22	·025048
AbC	3·6829	−2·22	−·602788	−·02	−·005430	−·22	−·029868
ABc	1·7265	−2·22	−1·285838	6·38	3·231970	·02	·011584
Abc	·3135	2·22	7·081338	·02	·637960	−·02	−·637960
aBC	2·0365	·02	·009820	6·38	2·739920	−·22	−·108028
abC	·0035	−·02	−5·714286	·02	5·714286	·22	62·857140
aBc	1·9599	−·02	−·010204	−6·38	−2·847084	−·02	−·018204
abc	·0001	·02	200·000000	−·02	−200·000000	·02	200·000000
	16·0000	0·00		0·00		0·00	

The information matrix **i**, per single observation, can be calculated by using the formulæ of section 3 (*b*). Thus

$$i_{11} = \frac{1}{16}[2\cdot 22(\ \cdot 353666) + \ldots + \cdot 02(\ 200\cdot 000000)] = 1\cdot 550821,$$

$$i_{12} = \frac{1}{16}[2\cdot 22(-\cdot 888944) + \ldots + \cdot 02(-200\cdot 000000)] = -\cdot 812331,$$

etc.

$$\mathbf{i} = \begin{pmatrix} 1\cdot 550821 & -\cdot 812331 & \cdot 173998 \\ -\cdot 812331 & 3\cdot 642891 & -\cdot 213733 \\ \cdot 173998 & -\cdot 213733 & 1\cdot 117181 \end{pmatrix}$$

The covariance matrix **v**, per single observation, of the maximum likelihood estimates obtained by choosing such values of y_1, y_2, y_3

which make the scores for an observed set of frequencies vanish, is the inverse of the information matrix.

$$\mathbf{v} = \begin{pmatrix} \cdot 737882 & \cdot 159589 & -\cdot 084391 \\ \cdot 159589 & \cdot 312139 & \cdot 034861 \\ -\cdot 084391 & \cdot 034861 & \cdot 914923 \end{pmatrix}$$

(b) *Method of efficient scores : y's being subject to Kosambi's formula*

The information matrix j, per single observation, for y_1, y_3 in this case is obtained from i, calculated above, by using the formula given in section 5 and illustrated in section 6.

The procedure of computation is as follows :—

$$\mu_1 = \cdot 932671 \quad , \quad \mu_3 = \cdot 642498$$

j_{rs}	j_{11}	j_{13}	j_{33}
i_{rs}	1·550821	·173998	1·117181
$\mu_r\mu_s i_{22}$	3·168852	2·182996	1·503807
$\mu_r i_{s2}$	−·757637	−·521921	−·137323
$\mu_s i_{2r}$	−·757637	−·199342	−·137323
Total	3·204399	1·635701	2·346342

The covariance matrix u, per single observation, of the estimates of y_1 and y_3 is the inverse of j.

$$\mathbf{u} = \begin{pmatrix} \cdot 484471 & -\cdot 337738 \\ -\cdot 337738 & \cdot 661642 \end{pmatrix}$$

If $\ddot{y}_1$ and $\ddot{y}_3$ are the estimates of y_1 and y_3 then, substituting these values in Kosambi's formula, we get $\ddot{y}_2$, the estimate of y_2. The variance of $\ddot{y}_2$ is calculated by the formula,

$$V(\ddot{y}_2) = \cdot 484471(\cdot 932671)^2 - 2(\cdot 337738)(\cdot 932671)(\cdot 642498) + \cdot 661642(\cdot 642498)^2$$

$$= \cdot 289786 \text{ per single observation.}$$

(c) *Individual segment method taking y's as free*

In the individual segment method, only two factors are considered each time and the data are scored for the corresponding recombination fraction. The results of the cross AbC × aBc supply data in repulsion for the estimation of y_3 and y_1 and in coupling for y_2. To score for y_1, it is necessary to consider the scores appropriate to data in repulsion as given in Mather's book, *The Measurement of Linkage in Heredity*, p. 61 (first edition). The observed frequencies in the eight phenotypical classes are represented by n_{ijk} as in table 2. Ignoring the classification with respect to the first factor we get the scores at $y_1 = \cdot 28$

in the four phenotypical classes and observed frequencies as given in table 8.

TABLE 8

Scores in repulsion at $y_1 = \cdot 28$

Phenotype	Observed frequency, ignoring the first factor	Scores
BC	$n_{111}+n_{011}$	·269438
Bc	$n_{110}+n_{010}$	−·607639
bC	$n_{101}+n_{001}$	−·607639
bc	$n_{100}+n_{000}$	7·142857

If the total score for y_1 by this method is represented by Q_1, then

$$Q_1 = \cdot 269438(n_{111}+n_{011}) - \cdot 607639(n_{110}+n_{010}+n_{101}+n_{001}) + 7\cdot 142857(n_{100}+n_{000})$$

The scores for y_2 and y_3 are similarly calculated.

$$Q_2 = -\cdot 562249(n_{111}+n_{101}) + 2\cdot 745098(n_{110}+n_{100}+n_{011}+n_{001}) - 2\cdot 857143(n_{010}+n_{000})$$

$$Q_3 = \cdot 029986(n_{111}+n_{110}) - \cdot 060054(n_{101}+n_{100}+n_{011}+n_{010}) + 66\cdot 666667(n_{001}+n_{000})$$

The estimate of y_i is found, by this method, by choosing that value of y_i which makes the score Q_i zero for the observed set of frequencies. The estimates differ from the true values unless the frequencies are the same as their expected values. If the probabilities of given departures of the Q's from their expected values or simply their sampling errors in large samples are known, then the corresponding errors introduced in the estimate can be calculated. Since the Q's are linear functions of the frequencies in the eight phenotypical classes with probabilities as given in table 7, the covariance m_{ij}, per single observation, between Q_i and Q_j can be calculated (as given in Fisher's *Statistical Methods*, p. 303, ninth edition) by taking the triple products of coefficients of the frequency for any class in Q_i and Q_j and the probability for that class and summing over all the eight classes. Thus

$$m_{11} = \tfrac{1}{16}[(\cdot 269438)^2(6\cdot 2771+2\cdot 0365)+(\cdot 607639)^2(1\cdot 7265+1\cdot 9599+3\cdot 6829+\cdot 0035)+(7\cdot 142857)^2(\cdot 3135+\cdot 0001)]$$
$$= 1\cdot 207860.$$

The matrix $\mathbf{m} = (m_{ij})$, thus calculated, comes out as

$$\mathbf{m} = \begin{pmatrix} 1\cdot 270860 & \cdot 526986 & -\cdot 002269 \\ \cdot 526986 & 3\cdot 118356 & \cdot 046039 \\ -\cdot 002269 & \cdot 046039 & 1\cdot 002248 \end{pmatrix}$$

The covariance of the estimates of y_i and y_j is given by $m_{ij}/m_{ii}m_{jj}$

so that, using the above elements, we can set up the covariance matrix $\mathbf{v}^*$ of the estimates of y_1, y_2, y_3.

$$\mathbf{v}^* = \begin{pmatrix} \cdot 827907 & \cdot 139913 & -\cdot 001874 \\ \cdot 139913 & \cdot 320680 & \cdot 014731 \\ -\cdot 001874 & \cdot 014731 & \cdot 997804 \end{pmatrix}$$

(d) Individual segment method, y's being subject to Kosambi's formula

In this there are two parameters, y_1 and y_3, to be estimated. The scores P_1, P_3 for y_1 and y_3 by this method are,

$$P_1 = Q_1 + \mu_1 Q_2$$
$$P_3 = Q_3 + \mu_3 Q_2$$

so that, knowing Q_1, Q_2, Q_3 considered above, and the differential coefficients μ_1, μ_3, the P's can be calculated. As the covariances of the Q's are known the covariances of P's can be calculated as follows

$$\begin{aligned} V(P_1) &= V(Q_1) + \mu_1^2 V(Q_2) + 2\mu_1 \text{ Cov } (Q_1 Q_2) \\ &= 1 \cdot 207860 + (\cdot 932671)^2 (3 \cdot 118356) + 2(\cdot 932671)(\cdot 526986) \\ &= 4 \cdot 903438. \end{aligned}$$

$$\begin{aligned} \text{Cov } (P_1 P_3) &= \text{Cov } (Q_1 Q_3) + \mu_1 \text{ Cov } (Q_2 Q_3) + \mu_3 \text{ Cov } (Q_1 Q_2) \\ &\qquad + \mu_1 \mu_3 V(Q_2) \\ &= 2 \cdot 247897. \end{aligned}$$

Similarly

$$V(P_3) = 2 \cdot 348684.$$

Thus $\mathbf{d}$, the covariance matrix, per single observation, of P_1, P_3 is

$$\mathbf{d} = \begin{pmatrix} 4 \cdot 903438 & 2 \cdot 247897 \\ 2 \cdot 247897 & 2 \cdot 348684 \end{pmatrix}$$

If the Q's are independent, *i.e.* if three sets of data containing the same number of observations as the total in the case were independently observed for the three segments separately, then the individual segment method is the most efficient method. In this case the covariance matrix $\mathbf{a}$, per single observation, of P_1, P_3 is obtained from the above calculations by dropping the covariance terms. Thus

$$\begin{aligned} V(P_1) &= 1 \cdot 207860 + (\cdot 932671)^2 (3 \cdot 118356) \\ &= 3 \cdot 920430 \end{aligned}$$

and

$$\mathbf{a} = \begin{pmatrix} 3 \cdot 920430 & 1 \cdot 868640 \\ 1 \cdot 868640 & 2 \cdot 289524 \end{pmatrix}$$

Its reciprocal $\mathbf{a}^{-1}$ gives, when the Q's are independent, the covariance matrix of the estimates obtained by choosing values of y_1 and y_3 which make the scores P_1 and P_3 vanish. This, on calculation, comes out as

$$\mathbf{a}^{-1} = \begin{pmatrix} \cdot 417483 & -\cdot 340737 \\ -\cdot 340737 & \cdot 714871 \end{pmatrix}$$

When the Q's are not independent the estimates, obtained by choosing such values of y_1 and y_3 which make the scores P_1 and P_3 for the observed set of frequencies vanish, have the covariance matrix $\mathbf{u}^*$ given by

$$\mathbf{u}^* = \mathbf{a}^{-1}\mathbf{d}\mathbf{a}^{-1}$$

This is a triple product of the matrices $\mathbf{a}^{-1}$, $\mathbf{d}$ and $\mathbf{a}^{-1}$. To evaluate this the product $\mathbf{a}^{-1}\mathbf{d}$ may be found first and then multiplied by $\mathbf{a}^{-1}$. The method of multiplying two matrices is to construct a matrix whose element in the i-th row and the j-th column is the sum of products of the ordered elements in the i-th row of the first matrix and j-th column of the second matrix.

$$\begin{aligned}\mathbf{u}^* &= \begin{pmatrix} \cdot417483 & -\cdot340737 \\ -\cdot340737 & \cdot714871 \end{pmatrix}\begin{pmatrix} 4\cdot903438 & 2\cdot247897 \\ 2\cdot247897 & 2\cdot348684 \end{pmatrix}\mathbf{a}^{-1} \\ &= \begin{pmatrix} 1\cdot281160 & \cdot138175 \\ -\cdot063827 & \cdot913064 \end{pmatrix}\begin{pmatrix} \cdot417483 & -\cdot340737 \\ -\cdot340737 & \cdot714871 \end{pmatrix} \\ &= \begin{pmatrix} \cdot487781 & -\cdot337761 \\ -\cdot337761 & \cdot674472 \end{pmatrix}\end{aligned}$$

The variance, per single observation, of the estimate of y_2 as calculated from the formula given in section 7 (b) is

$$(\cdot487781)(\cdot932671)^2 - 2(\cdot337761)(\cdot932671)(\cdot642498) + (\cdot674472)(\cdot642498)^2 = \cdot297933$$

A comparison is made below of the covariance matrices $\mathbf{v}$, $\mathbf{v}^*$, $\mathbf{u}$, $\mathbf{u}^*$ of the estimates obtained by the four methods discussed above. Choosing the variances alone we have the comparisons as shown in table 9.

TABLE 9

Variances of estimates and relative efficiencies

Method	Variances of estimates of y's treated as						Percentage increase in efficiency due to Kosambi's formula		
	Free parameters			Subject to Kosambi's formula					
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
ι) Efficient scores	·737882	·312139	·914923	·484471	·289786	·661642	52·30	7·71	38·2
ϳ) Individual segment	·827907	·320680	·997804	·487781	·297933	·674472	69·73	7·63	47·9
ι) Relative efficiency $= \frac{(\alpha)}{(\beta)} \times 100$	89·12	97·34	91·69	99·32	97·31	98·10	...	...	...

A measure of overall efficiency can be obtained by comparing the covariance matrices of the estimates by the two methods. Instead

of the simple ratio of the determinants, its s-th root, where s is the number of parameters estimated, may be defined as the relative efficiency of one method as compared with another. Thus, when y's are considered as free parameters the overall efficiency of the individual segment method is

$$\left\{\frac{|v|}{|v^*|}\right\}^{\frac{1}{3}} = \left\{\frac{\cdot 183367}{\cdot 245190}\right\}^{\frac{1}{3}} = \sqrt[3]{\cdot 7478} = \cdot 9077 \text{ or } 90\cdot 77 \text{ per cent.}$$

The corresponding efficiency when y's are subject to Kosambi's formula is

$$\left\{\frac{|u|}{|u^*|}\right\}^{\frac{1}{3}} = \cdot 9801 \text{ or } 98\cdot 01 \text{ per cent.}$$

The overall increase in efficiency by using Kosambi's formula can be calculated by comparing the covariance determinants of the estimates of y_1 and y_3 alone. This increase for the method of efficient scores is

$$\frac{\begin{vmatrix} \cdot 737882 & \cdot 173998 \\ \cdot 173998 & \cdot 914923 \end{vmatrix}^{\frac{1}{2}}}{|u|^{\frac{1}{3}}} - 1 = 79\cdot 89 \text{ per cent.}$$

and for the individual segment method

$$\frac{\begin{vmatrix} \cdot 827909 & -\cdot 001874 \\ -\cdot 001874 & \cdot 997804 \end{vmatrix}^{\frac{1}{2}}}{|u^*|^{\frac{1}{3}}} - 1 = 96\cdot 06 \text{ per cent.}$$

These calculations lead to the following conclusions, (1) The use of Kosambi's formula, when applicable, considerably increases the efficiency of the estimates by either method. In the particular example chosen the overall increase in efficiency is 79·89 per cent. for the most efficient method and 96·06 per cent. for the individual segment method. (2) The loss of efficiency due to the simpler analysis of the individual segment method is smaller when the recombination fractions are estimated with the use of Kosambi's formula. In the above example the overall efficiency of the individual segment method, when y's are considered as free parameters, is 90·77 per cent. which increases to 98·01 per cent. when y's are subject to Kosambi's formula.

8. SUMMARY

The following results have been discussed in this article.

1. The appropriate scoring of data giving the simultaneous segregation of three factors and the method of arriving at the best estimates of recombination fractions from data relating to various sources and types of crosses have been discussed in the two cases (i) when they are taken as free parameters and (ii) when they conform to Kosambi's formula. It has been observed that the scores and information matrix

in the latter case are connected with those in the former by simple relations and in any practical example it is convenient to score the data considering the recombination fractions as free parameters and then deduce the total scores appropriate to the latter case.

It has been found, in the examples discussed in the article, that the estimates found by using Kosambi's formula have considerably smaller variances. This is so, for when this formula is true there are only two parameters to be estimated (the third one being deduced from the formula) and any method of estimation which does not make use of the formula, being different from the maximum likelihood method appropriate to the two parameters, is bound to be inefficient.

The use of such empirical relations as the one considered above among the parameters to be estimated, when known, enhances the precision of the estimates although they may not be strictly accurate. The assumption of a slightly inaccurate relationship may introduce bias in the estimates but such estimates are more useful than the less efficient estimates so long as the bias, in any case, is small in comparison with its standard error. This, in some way, is secured when the test for a hypothesis specifying some restrictions indicates close agreement with the observations. Kosambi's formula is very useful from this point of view, as its use considerably enhances the precision of the estimates. A test has been proposed to judge the validity of this formula in any particular case.

2. In view of the slightly heavier computation involved in the method of efficient scores an investigation has been made to find out the loss in efficiency due to the simpler method of scoring by considering the data as classified with respect to only two factors each time and considering them as independent distributions. The latter method is called the individual segment method.

It has been found that if the data consist of only back crosses both the methods lead to identical estimates, when the recombination fractions are considered as free parameters. This, however, is not true when the estimates are found subject to Kosambi's formula but the loss in efficiency is not expected to be considerable.

In the case of intercrosses a particular example has been chosen to find the relative efficiency of this method in the two cases when the recombination fractions are considered as (i) free parameters and (ii) subject to Kosambi's formula. It is found that the loss of information due to the simpler analysis is negligible when the recombination fractions conform to Kosambi's formula. This may not be generally true, but the loss in any case is not expected to be considerable. Similar results may be expected in the case of data giving the simultaneous segregation of more than three factors. Their exact treatment involves some complications because some extraneous parameters have to be estimated and accounted for in the evaluation of the variances of the estimates of the recombination fractions. This, however, awaits further study.

Finally, I should like to thank Professor R. A. Fisher for his guidance and criticism during the preparation of this paper and Dr K. Mather for his valuable suggestions which led to a clarification of the various steps involved in this article.

9. REFERENCES

BHAT, N. R. 1948.
An improved genetical map of Punnett's B-chromosome in the sweet pea *Lathyrus odoratus* L.
J. Genet., *48*, 3, 343-358.

DE WINTON, D., AND HALDANE, J. B. S. 1935.
The genetics of *Primula sinensis*. III. Linkage in the diploid.
J. Genet., *31*, 67-100.

FISHER, R. A. 1935.
The detection of linkage with dominant abnormalities.
Ann. Eugen., *6*, 187-201.

FISHER, R. A. 1946.
A system of scoring linkage data with special reference to pied factors in mice.
Amer. Nat., *80*, 568-578.

KOSAMBI, D. D. 1944.
The estimation of map distance from recombination values.
Ann. Eugen., *12*, 172-176.

RAO, C. R. 1947.
On minimum variance and the estimation of several parameters.
Proc. Camb. phil. Soc., *43*, 280-283.

RAO, C. R. 1948.
Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation.
Proc. Camb. phil. Soc., *44*, 50-57.