THE ESTIMATION OF GENE FREQUENCIES FROM FAMILY RECORDS

I. FACTORS WITHOUT DOMINANCE

D. J. FINNEY

Lecturer in the Design and Analysis of Scientific Experiment, University of Oxford

Received 11.ii.48

1. INTRODUCTION

If a random sample of unrelated individuals can be classified into genotypes for a particular genetic factor, a simple count of genes leads to efficient estimates of the relative frequencies of the two or more allelomorphs. If dominance makes certain genotypes phenotypically indistinguishable, the estimation of gene frequencies may be more troublesome, though, once the dominance relationships are understood, a method can readily be devised by application of the principle of maximum likelihood ; for a pair of allelomorphs, this amounts to no more than a count of recessives, but when several allelomorphic genes are involved, as for the ABO blood groups (Stevens, 1938), greater complexities arise. Often, however, the members of a sample from which gene frequencies are to be estimated are not wholly unrelated. Fisher (1940) has pointed out that the simple method of estimation just described is still consistent, but that it will in general be less precise than if it were based on an " unrelated " sample. If an assessment of precision is required, one of three procedures must then be followed : the sample might be reduced by rejection of all but one from each related group, so as to enable the estimation to be made by the old method ; a theoretical investigation of the precision of the estimate obtained by treating the sample as though its members were unrelated might be undertaken ; or an entirely new method of estimation, with its own assessment of precision, might be developed as particularly appropriate to the types of relationship encountered.

The first of these alternatives is undesirable, since the choice of individuals to be rejected will introduce ambiguity and, in general, the rejects will be capable of giving additional information on the gene frequencies. Fisher considered the third alternative in its application to the particular problem of a simple recessive gene ; he developed maximum likelihood scoring systems for pairs of individuals consisting of parent and child, sibs, or half-sibs. The method is analagous to the systems of scoring for detection of genetic linkage discussed by Fisher (1935 *a* and *b* ; 1946) and Finney (1940, 1941, 1943). Unfortunately, extension of the scoring technique to

larger related groups involves formulæ of rapidly increasing complexity, and appears to be scarcely practicable beyond, say, groups of three relatives. Cotterman (1947) has investigated the second of the three possibilities. He has obtained simple formulæ for the variances of estimates obtained by combining data from all individuals as though they were unrelated : his method, of course, is not fully efficient, but for a number of examples of small families he finds its efficiency to be very high. Careful examination of the principles of scoring suggests that slight modifications might increase still further the efficiency of Cotterman's technique. The present paper describes the modified procedure as applied to a pair of allelomorphic genes showing no dominance, and necessarily recapitulates some of Cotterman's work in order to give a complete description of the method. Similar results for a factor showing dominance will be discussed in a second paper.

2. THE THEORY OF SCORING

The records available for the estimation of a gene frequency may be divided into families. For this purpose, a *family* will be defined as a group of individuals such that any two are either blood relatives (*e.g.* parent and child, sibs, half-sibs) or can be connected by a chain of blood relationships entirely within the records (*e.g.* two parents, themselves unrelated but having one or more children recorded) ; no two members of different families can be blood relatives. The division into families is then unique. In practice, of course, the more distant blood-ties are necessarily ignored.

If the frequency, ν, of a certain gene is to be estimated, a scoring system must be set up for every family type and size in the records. For a family whose members have a particular pattern of relationships to one another, the probability that the phenotypes of the members shall be any one of the possible sets may be expressed as a function of ν, $\mathrm{P}(\nu)$, where

$$\mathrm{S}\,(\mathrm{P}) = 1,$$

the summation being over all possible sets of phenotypes. The probability will usually be based on an assumption of random mating and absence of differential mortality or fertility in the genotypes, and will also take account only of the blood-relationships between individuals. If information were available on the linkage of the factor studied with sex or with some other recorded character, however, allowance for this could be made in the expression for $\mathrm{P}(\nu)$.

In theory, an efficient scoring system for a family type can easily be constructed by application of the principle of maximum likelihood. For, if a first approximation to the required estimate is chosen,

$$\lambda = \frac{1}{\mathrm{P}} \frac{d\mathrm{P}}{d\nu} \quad . \quad . \quad . \quad . \quad (1)$$

evaluated for this approximation gives a score appropriate to any recorded family, with a quantity of information and score divisor

$$\kappa = \mathrm{S}\left\{\frac{1}{\mathrm{P}}\left(\frac{d\mathrm{P}}{d\nu}\right)^2\right\}. \quad . \quad . \quad . \quad . \quad (2)$$

This score provides an adjustment for the first approximation, so as to give a revised estimate

$$\nu + \frac{\Sigma\lambda}{\Sigma\kappa}, \text{ with variance } 1/\Sigma\kappa,$$

where Σ represents summation over all families in the records. A more convenient score to use in practice is obtained as a slight modification of λ, namely

$$x = \mathrm{W_L}\nu + \frac{\nu(1-\nu)}{\mathrm{P}}\frac{d\mathrm{P}}{d\nu}, \quad . \quad . \quad . \quad (3)$$

where $\mathrm{W_L}$ is the *weight* of the maximum likelihood score for a family and is given by

$$\mathrm{W_L} = \nu(1-\nu)\,\mathrm{S}\left\{\frac{1}{\mathrm{P}}\left(\frac{d\mathrm{P}}{d\nu}\right)^2\right\} \quad . \quad . \quad . \quad (4)$$

The revised estimate is now

$$\nu = \frac{\Sigma x}{\Sigma \mathrm{W_L}} \quad . \quad . \quad . \quad . \quad (5)$$

with variance

$$\mathrm{V}(\nu) = \frac{\nu(1-\nu)}{\Sigma \mathrm{W_L}} \quad . \quad . \quad . \quad . \quad (6)$$

Here $\mathrm{W_L}$ may be regarded as the equivalent number of independent genes provided by a family record. In the absence of dominance, a " family " of one—that is to say, a single individual with no recorded relatives—clearly contributes two independent genes, and, as is demonstrated in section 3, evaluation of the weight gives $\mathrm{W_L} = 2$.

As will appear in succeeding sections, maximum likelihood scoring is not always practicable. Many other methods of scoring are possible : none can give more information (or greater weight per family) than the maximum likelihood system, and most will give less. This paper is concerned with the development of scoring systems that shall be simple in application without involving too serious a loss of information. Suppose that z is any function of the phenotypes of a family such that the average value of z, over all possible sets of phenotypes in families with the same relationship-pattern, is dependent upon ν :

$$\mathrm{E}(z) = \zeta = \zeta(\nu). \quad . \quad . \quad . \quad . \quad (7)$$

If ν_0 is a first approximation to the required estimate, the expectation of z on the hypothesis represented by the estimate may be expanded, to the first order, as :

$$\mathrm{E}(z) = \zeta_0 + \zeta_0'\delta\nu_0,$$

where

$$\zeta_0 = \zeta(\nu_0), \quad \zeta' = \frac{d\zeta}{d\nu}.$$

The score

$$y = \left(\nu_0 - \frac{\zeta_0}{\zeta_0'}\right) + \frac{z}{\zeta_0'} \quad . \quad . \quad . \quad . \quad (8)$$

will then have an expectation

$$\mathrm{E}(y) = \nu_0 + \delta\nu_0, \quad . \quad . \quad . \quad . \quad (9)$$

and so an improved approximation to ν can be obtained as an average of y-scores. The variance of y is

$$\mathrm{V}(y) = \frac{\mathrm{SP}z^2 - \zeta^2}{\zeta'^2},$$

so that the weight to be attached to the score from a single family is

$$\mathrm{W} = \frac{\zeta'^2 \nu(1-\nu)}{\mathrm{SP}z^2 - \zeta^2} \quad . \quad . \quad . \quad . \quad (10)$$

The efficiency of the scoring system may be assessed by comparison with the maximum likelihood weight, and is

$$\text{Efficiency } (y) = \frac{\mathrm{W}}{\mathrm{W_L}} \quad . \quad . \quad . \quad . \quad (11)$$

The estimate of ν from a number of families is the weighted mean of the y scores

$$\nu = \frac{\Sigma \mathrm{W}y}{\Sigma \mathrm{W}}, \quad . \quad . \quad . \quad . \quad (12)$$

for which

$$\mathrm{V}(\nu) = \frac{\nu(1-\nu)}{\Sigma \mathrm{W}} \quad . \quad . \quad . \quad . \quad (13)$$

Even though maximum likelihood scoring of all available records may be impracticable, on account of the algebraic complexities, it is always permissible to score some parts of the data, perhaps the smaller families, by maximum likelihood x-scores, and other parts on a simpler system of less than full efficiency. If Σ_1 and Σ_2 represent summations for the two parts,

$$\nu = \frac{\Sigma_1 x + \Sigma_2 \mathrm{W}y}{\Sigma_1 \mathrm{W_L} + \Sigma_2 \mathrm{W}} \quad . \quad . \quad . \quad . \quad (14)$$

with

$$\mathrm{V}(\nu) = \frac{\nu(1-\nu)}{\Sigma_1 \mathrm{W_L} + \Sigma_2 \mathrm{W}} \quad . \quad . \quad . \quad (15)$$

The lack of symmetry in x and y arises because x is already weighted, according to its definition, but the forms of y-scores to be used are

such that less tabulation is required if formulæ are expressed in terms of them rather than of weighted values, Wy.

The provisional value, ν_0, and the final estimate, ν, have not been distinguished throughout this argument. The procedure to be adopted is that a rough estimate, ν_0, is formed by any convenient method, scores and weights are derived using ν_0 for ν in the formulæ, and equations (5), (12), or (14) used to give a revised estimate. If the difference between this estimate and ν_0 is marked, the data may be re-scored using it instead of ν_0, and the process repeated until agreement is judged satisfactory. Equations (6), (13), or (15), using the latest set of weights, will then give the variance.

3. MAXIMUM LIKELIHOOD SCORING

Fisher (1940) has derived formulæ for maximum likelihood scores and weights, corresponding to equations (3) and (4) above, for families consisting of parent and child or two sibs classified in respect of a pair of allelomorphic genes showing dominance. In this section, similar results are obtained for a factor without dominance. The gene frequencies in the population will be taken as μ, ν ($\mu+\nu = 1$), where ν is to be estimated ; it is convenient to write

$$\mu\nu = \gamma. \qquad . \quad . \quad . \quad . \quad (16)$$

The MN blood types are typical of factors of this class.

(i) *Both Parents Recorded*

If two parents and their s children are classified for a factor such as the MN blood types, the children contribute no information additional to that from the parents. Maximum likelihood scoring would demonstrate this fact, which should be obvious since only four independent genes are present and complete knowledge of these

Genotype	P	$\frac{dP}{d\nu}$	$\frac{1}{P}\frac{dP}{d\nu}$	$\frac{1}{P}\left(\frac{dP}{d\nu}\right)^2$
MM . . .	μ^2	-2μ	$-\frac{2}{\mu}$	4
MN . . .	$2\mu\nu$	$2(\mu-\nu)$	$\frac{1}{\nu}-\frac{1}{\mu}$	$\frac{2(\mu-\nu)^2}{\mu\nu}$
NN . . .	ν^2	2ν	$\frac{2}{\nu}$	4

is provided by the parental genotypes. Cotterman (1947) points out that estimation of the gene frequency by count of all genes from the parents and the children will in fact reduce the precision below its value when the children are rejected from the score, since the children introduce additional sampling variation. The right procedure is to ignore the children and to score both parents as unrelated individuals.

Maximum likelihood scoring is then easy, and may be seen to reduce to the more obvious method of counting genes. For one individual, selected at random from the population, the derivation of the scores is as shown on page 203. Hence

$$S\left\{\frac{1}{P}\left(\frac{dP}{d\nu}\right)^2\right\} = 8 + \frac{2(1-2\nu)^2}{\mu\nu}$$
$$= \frac{2}{\gamma}.$$

Equations (4) and (3) now show that

$$W_L = 2 \quad . \quad . \quad . \quad . \quad (17)$$

and that the score is

$$\left.\begin{aligned} x &= 0 \text{ for MM} \\ x &= 1 \text{ for MN} \\ \text{or } x &= 2 \text{ for NN} \end{aligned}\right\} \quad . \quad . \quad . \quad . \quad (18)$$

Thus the maximum likelihood method is seen to be the same in result as a simple count of N genes divided by the total number of genes, or by twice the number of individuals.

If the symbol $W_{a,\,L}(s)$ represents the weight to be attached to a family for which a parents and s children are recorded, when scored on the maximum likelihood system, the result now obtained is

$$W_{2,\,L}(s) = 4 \text{ for all } s, \quad . \quad . \quad . \quad (19)$$

since the parents only are to be scored, each according to the scores in equation (18). Of course if the data were entirely unrelated individuals and pairs of parents, the problem of estimation would be exceedingly simple. A count of N genes, referred to a binomial distribution, would lead directly to an estimate, ν, and its variance, without any necessity to form a provisional value, ν_0. The reason for discussing this case in so much detail is that in others now to be considered complications arise, and the results for case (i) are required in the form just given for ease of combination.

(ii) One Parent Recorded

The same procedure can be used to give a system of scores for families of which one parent and s children appear in the records. The expression for $S\left\{\frac{1}{P}\left(\frac{dP}{d\nu}\right)^2\right\}$ becomes very complicated, however, for $s>2$, as the summations required for it do not seem to reduce in any simple manner. Cotterman (1947, table III) has shown that for $s = 1$

$$S\left\{\frac{1}{P}\left(\frac{dP}{d\nu}\right)^2\right\} = \frac{3-\gamma}{\gamma},$$

so that the weight per parent-child pair is

$$W_{1,\,L}(1) = 3-\gamma. \quad . \quad . \quad . \quad . \quad (20)$$

Cotterman points out that usually a parent-child pair would give exact information on three independent genes, but that if both members are MN only two genes can be established with certainty, so that on an average the weight is a little less than 3. The adjusted scores are as follows, for different combinations of parent and child genotypes :—

$$\left.\begin{array}{lll} \text{Parent MM, child MM} & & x = -\mu\nu^2 \\ \text{,, MM, ,, MN} & \left.\begin{array}{l}\\\end{array}\right\} & x = 1-\mu\nu^2 \\ \text{,, MN, ,, MM} & & \\ \text{,, MN, ,, MN} & & x = 1+\nu-\mu\nu^2 \\ \text{,, MN, ,, NN} & \left.\begin{array}{l}\\\end{array}\right\} & x = 2-\mu\nu^2 \\ \text{,, NN, ,, MN} & & \\ \text{,, NN, ,, NN} & & x = 3-\mu\nu^2 \end{array}\right\} \quad . \quad (21)$$

The score may be regarded as $-\mu\nu^2$ increased by the number of independent N genes demonstrated to be present, with a further addition of ν for the pair MN, MN. Table 1 contains numerical values of scores and weights.

TABLE 1

Maximum likelihood score and weight for one parent and one child

ν (provisional)	Scores for pairs of genotypes			W_L
	MM, MM	MM, MN	MN, MN	
0·00 . . .	0·0000	1·0000	1·0000	3·0000
0·05 . . .	−0·0024	0·9976	1·0476	2·9525
0·10 . . .	−0·0090	0·9910	1·0910	2·9100
0·15 . . .	−0·0191	0·9809	1·1309	2·8725
0·20 . . .	−0·0320	0·9680	1·1680	2·8400
0·25 . . .	−0·0469	0·9531	1·2031	2·8125
0·30 . . .	−0·0630	0·9370	1·2370	2·7900
0·35 . . .	−0·0796	0·9204	1·2704	2·7725
0·40 . . .	−0·0960	0·9040	1·3040	2·7600
0·45 . . .	−0·1114	0·8886	1·3386	2·7525
0·50 . . .	−0·1250	0·8750	1·3750	2·7500
0·55 . . .	−0·1361	0·8639	1·4139	2·7525
0·60 . . .	−0·1440	0·8560	1·4560	2·7600
0·65 . . .	−0·1479	0·8521	1·5021	2·7725
0·70 . . .	−0·1470	0·8530	1·5530	2·7900
0·75 . . .	−0·1406	0·8594	1·6094	2·8125
0·80 . . .	−0·1280	0·8720	1·6720	2·8400
0·85 . . .	−0·1084	0·8916	1·7416	2·8725
0·90 . . .	−0·0810	0·9190	1·8190	2·9100
0·95 . . .	−0·0451	0·9549	1·9049	2·9525
1·00 . . .	0·0000	1·0000	2·0000	3·0000

The score is unaffected by interchange of parent and child genotypes.
For MN, NN, add 1·0000 to the score for MM, MN.
For NN, NN, add 2·0000 to the score for MM, MN.

For $s = 2$, similar calculations lead to

$$W_{1,\,L}(2) = \frac{14-\gamma+2\gamma^2}{2(2+\gamma)} \quad . \quad . \quad . \quad (22)$$

an expression which is always larger than that for $s = 1$, since 2 children on an average give more information about the unrecorded parents. The scores for $s = 2$, however, are much more complicated ; x is no longer unaffected by interchange of parent and child genotypes, and twelve different cases have to be distinguished. Neither the formulæ nor a table will be given here, as replacement of the method by that of section 4 is recommended. For any higher value of s, of course, the complexities are still greater. For very large s, the proportions of the three genotypes found amongst the sibs will be sufficient to indicate almost with certainty the genotype of the unrecorded parent, so that the family may be scored as though both parents were recorded. This implies that

$$W_{1,\,L}(s) \to 4 \quad \text{as} \quad s \to \infty, \quad . \quad . \quad . \quad (23)$$

for all values of ν.

(iii) Neither Parent Recorded

The scoring of a family of s sibs for which neither parent has been recorded may be developed on the same lines, but the algebra is more troublesome. For $s = 1$, the one child is an " unrelated individual " and should be scored as such. For $s = 2$, the weight per sibship is

$$W_{0,\,L}(2) = \frac{6+5\gamma+4\gamma^2}{(1+\gamma)(2+\gamma)} \quad . \quad . \quad . \quad (24)$$

a result which has also been obtained by Cotterman. The weight is rather less than for a parent-child pair. The adjusted scores are obtained from one of the following formulæ :—

$$\left.\begin{array}{lll}
\text{Children MM, MM} & x = W_\nu - \dfrac{2\nu(3-2\nu)}{2-\nu} \\
\text{,, MM, MN} & x = W_\nu + \dfrac{2(1-4\nu+2\nu^2)}{2-\nu} \\
\text{,, MM, NN} & x = W_\nu + 2(1-2\nu) \\
\text{,, MN, MN} & x = W_\nu + \dfrac{(1-2\nu)(1+2\gamma)}{1+\gamma} \\
\text{,, MN, NN} & x = W_\nu + \dfrac{2(1-2\nu^2)}{1+\nu} \\
\text{,, NN, NN} & x = W_\nu + \dfrac{2(1+2\nu)(1-\nu)}{1+\nu}
\end{array}\right\} \quad (25)$$

Numerical values of the scores and weights are given in table 2. For $s > 2$, the enumeration of all cases and evaluation of W_L would be very laborious, and in practice Cotterman's method (section 4)

is to be preferred. For large s, the proportions of the three genotypes among the sibs will uniquely specify the parental genotypes, and the family may be scored as though the parents were recorded ;

$$W_{0,\,L}(s) \to 4 \quad \text{as} \quad s \to \infty \qquad . \qquad . \qquad . \qquad (26)$$

for all ν.

TABLE 2

Maximum likelihood score and weight for two sibs

ν (provisional)	Scores for pairs of genotypes						
	MM, MM	MM, MN	MN, NN	MN, MN	MN, NN	NN, NN	W_L
0·00 . .	0·0000	1·0000	2·0000	1·0000	2·0000	2·0000	3·0000
0·05 . .	−0·0031	0·9713	1·9456	1·0864	2·0409	2·1361	2·9125
0·10 . .	−0·0102	0·9372	1·8846	1·1506	2·0664	2·2482	2·8455
0·15 . .	−0·0187	0·9002	1·8191	1·1983	2·0800	2·3409	2·7942
0·20 . .	−0·0268	0·8621	1·7510	1·2337	2·0843	2·4176	2·7548
0·25 . .	−0·0331	0·8241	1·6812	1·2602	2·0812	2·4812	2·7248
0·30 . .	−0·0363	0·7872	1·6107	1·2801	2·0722	2·5338	2·7024
0·35 . .	−0·0356	0·7523	1·5401	1·2957	2·0587	2·5772	2·6861
0·40 . .	−0·0300	0·7200	1·4700	1·3088	2·0415	2·6129	2·6751
0·45 . .	−0·0184	0·6913	1·4009	1·3208	2·0216	2·6423	2·6688
0·50 . .	0·0000	0·6667	1·3333	1·3333	2·0000	2·6667	2·6667
0·55 . .	0·0264	0·6471	1·2678	1·3480	1·9775	2·6872	2·6688
0·60 . .	0·0622	0·6336	1·2051	1·3664	1·9551	2·7051	2·6751
0·65 . .	0·1089	0·6275	1·1460	1·3904	1·9339	2·7217	2·6861
0·70 . .	0·1686	0·6301	1·0917	1·4222	1·9152	2·7387	2·7024
0·75 . .	0·2436	0·6436	1·0436	1·4647	1·9008	2·7579	2·7248
0·80 . .	0·3372	0·6705	1·0038	1·5211	1·8927	2·7816	2·7548
0·85 . .	0·4533	0·7142	0·9750	1·5959	1·8940	2·8129	2·7942
0·90 . .	0·5973	0·7792	0·9610	1·6949	1·9083	2·8557	2·8455
0·95 . .	0·7764	0·8716	0·9668	1·8260	1·9412	2·9156	2·9125
1·00 . .	1·0000	1·0000	1·0000	2·0000	2·0000	3·0000	3·0000

4. SIMPLIFIED SCORING SYSTEMS

If the records consisted only of unrelated individuals, pairs of parents with any number of children, parent-child pairs and sib-pairs, the maximum likelihood scoring systems described and tabulated in section 3 would suffice for the estimation of ν. For larger families, the complexity of maximum likelihood scores makes it unlikely that analagous tables will be prepared, and unless an investigator chooses to derive maximum likelihood equations specially for his data he will have to be content with some alternative method of estimation.

Cotterman (1947) has shown that a score consisting of the total number of N genes recorded is of high efficiency for the estimation of ν, at least in those cases for which the maximum likelihood weight is available for comparison. The weight to be attached to this score is, of course, less than the total number of genes recorded, on account of the non-independence of records from related persons. Cotterman

has obtained and tabulated the expressions for weights corresponding to complete or partial records of parents and children. In the absence of tables of maximum likelihood scores for different values of s, when the non-statistician is obliged to use a method that is not fully efficient, the simplicity and high efficiency of Cotterman's method are strong recommendations for its general use. By a small modification, which necessitates only a slight elaboration of the instructions for scoring, the scoring of records of one parent with s children can be made of still higher efficiency, and details are given later in this section.

In order to show the relationship between the modified scores and the maximum likelihood and Cotterman systems, certain of Cotterman's results will be repeated here. The symbol $W_{a,\,C}(s)$ will be used to denote the total weight in Cotterman's method of a family having a parents and s children recorded, and $W_{a,\,M}(s)$ as the corresponding weight for the modified, or most efficient linear, scoring system. In practice, weights per gene are more convenient than total family weights, as the calculations for combination of scores are more easily performed in terms of them, but for assessment of efficiency the total weight is also required.

(i) *Both Parents Recorded*

Cotterman has shown that a score based on the total number of N genes in the $(s+2)$ individuals carries a weight

$$W_{2,\,C}(s) = \frac{4(s+2)^2}{(s+1)(s+4)} \quad . \quad . \quad . \quad . \quad (27)$$

This quantity is less than 4 (except for $s = 0$), so that scoring of the children has had the effect of reducing the information obtained by introducing irrelevant variation. The better procedure (as Cotterman states) is to score the parents only, as in section 3 (i), a fully efficient method which can quite correctly be combined with non-efficient scoring for other families.

(ii) *One Parent Recorded*

Cotterman scores this type of family by counting the N genes in the $(s+1)$ individuals and assigning to the total a divisor $2(s+1)$. He obtains a result equivalent to

$$W_{1,\,C}(s) = \frac{4(s+1)}{s+2} \quad . \quad . \quad . \quad . \quad (28)$$

The efficiency of this score for two special cases can be obtained by comparison of equation (28) with equations (20) and (22), giving

$$\text{Efficiency} = \frac{8}{3(3-\gamma)} \text{ for } s = 1, \quad . \quad . \quad . \quad (29)$$

and

$$\text{Efficiency} = \frac{6(2+\gamma)}{14-\gamma+2\gamma^2} \text{ for } s = 2. \quad . \quad . \quad . \quad (30)$$

These quantities are both large, the first being always at least 0·89 and the second always at least 0·86 ; they are shown graphically in fig. 1. As s increases without limit, $W_{1,\,C}(s)$ tends to 4, so that the score approaches full efficiency.

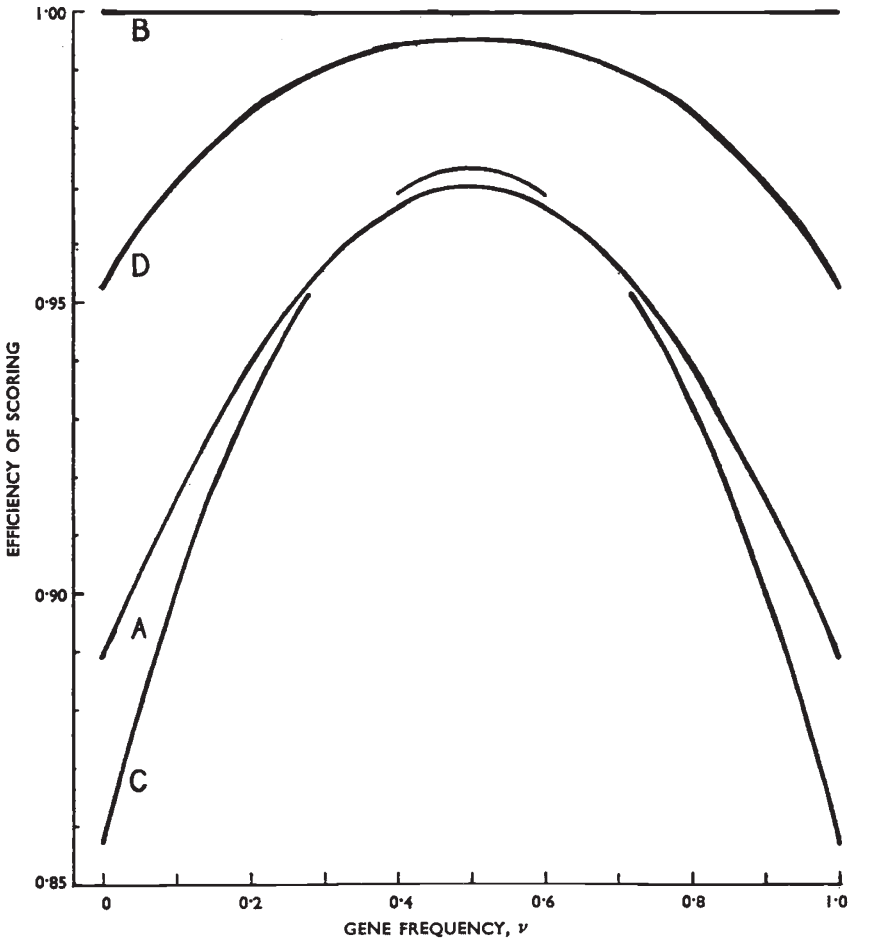

FIG. 1.—Efficiencies of Cotterman's scores and most efficient linear scores for families with one recorded parent, type (ii).

Curve A : Cotterman, $s = 1$
Curve B : Most efficient linear, $s = 1$
Curve C : Cotterman, $s = 2$
Curve D : Most efficient linear, $s = 2$

Nevertheless, it is not difficult to construct a score of still higher efficiency. Just as scoring of the children when both parents are recorded serves only to increase the variance of the estimate, so here, when only one parent is recorded, one gene in each child can add nothing to the information and may be expected to reduce precision if scored. If the recorded parent is MM, every N gene found in the children must come from the unrecorded parent. If the recorded

parent is MN, the N genes amongst the children will on an average contain $\frac{1}{2}s$ from this parent, and if the recorded parent is NN, the N genes in the children must include s from this source. Elimination of these N genes from the score should increase the weight of a family record.

The actual scoring system for maximum precision, subject only to the restriction that any score shall be a linear function of the observed genotype frequencies, is not quite so simple as the heuristic argument of the last paragraph might suggest. The required formulæ can be derived easily by separate examination of the possibilities for each genotype of the recorded parent. If the recorded parent is MM, the probabilities for numbers of MM and MN children in the total of s are :—

Other parent	No. of children		P	No. of N genes in children (z)
	MM	MN		
MM . . .	s	0	μ^2	0
MN . . .	m	$s-m$	$2\mu\nu\binom{s}{m}2^{-s}$	$s-m$
NN . . .	0	s	ν^2	s

If z is defined as the number of N genes amongst the children, summations of z and z^2 over all values of m (including $m = 0$, $m = s$) lead to formulæ for the expectation and variance :

$$\mathrm{E}(z) = s\nu, \quad . \quad . \quad . \quad . \quad . \quad (31)$$
$$\mathrm{V}(z) = \tfrac{1}{2}\gamma s(s+1) \quad . \quad . \quad . \quad . \quad (32)$$

Hence the information on ν given by the score is

$$i(\nu) = \frac{2s}{\gamma(s+1)} \quad . \quad . \quad . \quad . \quad (33)$$

When the recorded parent is MN, a slightly more complicated analysis is required. The probabilities are :—

Other parent	No. of children			P	z (see below)
	MM	MN	NN		
MM . .	m	$s-m$	0	$\mu^2\binom{s}{m}2^{-s}$	$m\alpha$
MN . .	m	$s-m-n$	n	$2\mu\nu\binom{s}{m+n}\binom{m+n}{m}2^{-(s+m+n)}$	$m\alpha+n\beta$
NN . .	0	$s-n$	n	$\nu^2\binom{s}{n}2^{-s}$	$n\beta$

The most general score which is a linear combination of genotype frequencies can be obtained by adding α for each MM child, β for each NN child, where α, β are arbitrarily chosen numbers. Summations for z and z^2 show that

$$\mathrm{E}(z) = \tfrac{1}{2}s\{\alpha+\nu(\beta-\alpha)\}, \quad . \quad . \quad . \quad . \quad . \quad (34)$$
$$\mathrm{V}(z) = \tfrac{1}{8}\gamma s(s+1)(\beta-\alpha)^2+\tfrac{1}{4}s(\mu\alpha^2+\nu\beta^2), \quad . \quad . \quad (35)$$

whence the information on ν is found to be

$$i(\nu) = \left\{\frac{\gamma(s+1)}{2s} + \frac{\mu\alpha^2+\nu\beta^2}{s(\beta-\alpha)^2}\right\}^{-1}.$$

This expression is maximised by taking

$$\alpha : \beta = -\nu : \mu,$$

when it becomes

$$i(\nu) = \frac{2s}{\gamma(s+3)} \quad . \quad . \quad . \quad . \quad (36)$$

In practice, a provisional value must be taken and scoring based on

$$\alpha : \beta = -\nu_0 : (1-\nu_0),$$

so leading in the usual manner to a revised estimate.

The third possibility, that the recorded parent is NN, is very similar to the first. The probabilities and numbers of N genes are :—

Other parent	No. of children		P	No. of N genes in children (z)
	MN	NN		
MM . . .	s	0	μ^2	s
MN . . .	$s-n$	n	$2\mu\nu\binom{s}{n}2^{-s}$	$s+n$
NN . . .	0	s	ν^2	$2s$

Hence

$$\mathrm{E}(z) = s+s\nu, \quad . \quad . \quad . \quad . \quad . \quad (37)$$
$$\mathrm{V}(z) = \tfrac{1}{2}\gamma s(s+1), \quad . \quad . \quad . \quad . \quad (38)$$

and again the information on ν is :

$$i(\nu) = \frac{2s}{\gamma(s+1)} \quad . \quad . \quad . \quad . \quad (39)$$

As was to be expected, the information when the recorded parent is MN is less than for the other two cases, but the difference is never very great ; the two are in the ratio 1 : 2 when $s = 1$, and nearer equality for all larger s. In a random-mating population, an average value may be used ; there is indeed little advantage in working with the average, since the separate expressions are so simple, but

for the sake of consistency with the scoring methods to be recommended for dominant factors, where use of the average saves some tabulation, it is also recommended here. From equations (33), (36), and (39), the mean information is

$$i(\nu) = \frac{2s}{\gamma(s+1)}(\mu^2+\nu^2)+\frac{2s}{\gamma(s+3)}(2\mu\nu)$$
$$= \frac{2s}{\gamma(s+1)}\left\{1-\frac{4\gamma}{(s+3)}\right\} \quad . \quad . \quad . \quad . \quad (40)$$

If now a score, y, is defined by :—

$$\left.\begin{array}{ll} \text{Recorded parent MM} : & y = 2(s-m) \\ \text{,,} \quad \text{,,} \quad \text{MN} : & y = 4n+2(s-2m-2n)\nu_0 \\ \text{,,} \quad \text{,,} \quad \text{NN} : & y = 2n \end{array}\right\} \quad . \quad (41)$$

where m, n are the numbers of MM, NN children respectively and ν_0 is a provisional estimate of ν, then in all cases, as may be seen from equations (31), (34) and (37),

$$\mathrm{E}(y) = 2s\nu \quad . \quad . \quad . \quad . \quad (42)$$

It may be noted that when ν_0 is small the score for a family with an MN recorded parent may exceed $2s$; this is not an inconsistency in the scoring, but merely a reflection of the strong evidence from the family that the provisional estimate is too small. A total of $2s$ genes are scored for the children, and, from equation (40), the weight per gene is

$$w = \frac{1}{s+1} - \frac{4\gamma}{(s+1)(s+3)} \quad . \quad . \quad . \quad (43)$$

(Fisher used w for the total weight of the family, here denoted by W). This function is tabulated in table 3 ; for a series of families, a weighted estimate based on equations (41), (42),

$$\nu = \frac{\Sigma wy}{\Sigma 2ws} \quad . \quad . \quad . \quad . \quad (44)$$

would obtain the information represented by equation (40) from each family. This scoring has not taken account of the information available from the recorded parents, and these may be scored, like the pairs of parents in (i), as additional unrelated individuals.

The total weight derived from the family in this scoring is :—

$$\mathrm{W}_{1,\,\mathrm{M}}(s) = 2+2ws$$
$$= \frac{2}{s+1}\left\{(2s+1)-\frac{4s\gamma}{s+3}\right\} \quad . \quad . \quad (45)$$

When $s = 1$ this reduces to equation (20), so that for parent-child

pairs scoring by equations (41) is fully efficient. When $s = 2$, comparison with equation (22) shows that

$$\text{Efficiency} = \frac{4(2+\gamma)(25-8\gamma)}{15(14-\gamma+2\gamma^2)} \quad . \quad . \quad . \quad (46)$$

TABLE 3

Weight per gene in most efficient linear scoring

ν (provisional)	Number of sibs (s)							
	1	2	3	4	5	6	7	8
	(ii) One parent and s sibs recorded							
0·00, 1·00 . .	0·5000	0·3333	0·2500	0·2000	0·1667	0·1429	0·1250	0·1111
0·05, 0·95 . .	0·4762	0·3207	0·2421	0·1946	0·1627	0·1398	0·1226	0·1092
0·10, 0·90 . .	0·4550	0·3093	0·2350	0·1897	0·1592	0·1371	0·1205	0·1075
0·15, 0·85 . .	0·4362	0·2993	0·2287	0·1854	0·1560	0·1348	0·1186	0·1060
0·20, 0·80 . .	0·4200	0·2907	0·2233	0·1817	0·1533	0·1327	0·1170	0·1046
0·25, 0·75 . .	0·4062	0·2833	0·2187	0·1786	0·1510	0·1310	0·1156	0·1035
0·30, 0·70 . .	0·3950	0·2773	0·2150	0·1760	0·1492	0·1295	0·1145	0·1026
0·35, 0·65 . .	0·3862	0·2727	0·2121	0·1740	0·1477	0·1284	0·1136	0·1019
0·40, 0·60 . .	0·3800	0·2693	0·2100	0·1726	0·1467	0·1276	0·1130	0·1014
0·45, 0·55 . .	0·3762	0·2673	0·2087	0·1717	0·1460	0·1271	0·1126	0·1011
0·50 . .	0·3750	0·2667	0·2083	0·1714	0·1458	0·1270	0·1125	0·1010
	(iii) s sibs recorded							
All values . .	1·0000	0·6667	0·5000	0·4000	0·3333	0·2857	0·2500	0·2222

ν (provisional)	Number of sibs (s)							
	9	10	11	12	13	14	15	16
	(ii) One parent and s sibs recorded							
0·00, 1·00 . .	0·1000	0·0909	0·0833	0·0769	0·0714	0·0667	0·0625	0·0588
0·05, 0·95 . .	0·0984	0·0896	0·0822	0·0759	0·0706	0·0659	0·0618	0·0582
0·10, 0·90 . .	0·0970	0·0884	0·0812	0·0751	0·0698	0·0653	0·0612	0·0577
0·15, 0·85 . .	0·0958	0·0873	0·0803	0·0743	0·0692	0·0647	0·0607	0·0572
0·20, 0·80 . .	0·0947	0·0864	0·0795	0·0736	0·0686	0·0642	0·0603	0·0568
0·25, 0·75 . .	0·0938	0·0857	0·0789	0·0731	0·0681	0·0637	0·0599	0·0565
0·30, 0·70 . .	0·0930	0·0850	0·0783	0·0726	0·0677	0·0634	0·0596	0·0562
0·35, 0·65 . .	0·0924	0·0845	0·0779	0·0723	0·0674	0·0631	0·0593	0·0560
0·40, 0·60 . .	0·0920	0·0842	0·0776	0·0720	0·0671	0·0629	0·0592	0·0559
0·45, 0·55 . .	0·0918	0·0840	0·0774	0·0718	0·0670	0·0628	0·0591	0·0558
0·50 . .	0·0917	0·0839	0·0774	0·0718	0·0670	0·0627	0·0590	0·0557
	(iii) s sibs recorded							
All values . .	0·2000	0·1818	0·1667	0·1539	0·1429	0·1333	0·1250	0·1176

a quantity always greater than Cotterman's efficiency, equation (30), and one which in fact never falls below 0·95. Values of the efficiency are shown in fig. 1. As $s \to \infty$, $W \to 4$, so that the scoring approaches full efficiency for large families; this indeed is obvious because by its very nature the scoring system must always be at least as efficient as Cotterman's.

The method of construction of these scores makes clear the procedure that must be adopted with half-sibs. If sibships of s_1 and s_2 have a common recorded parent (their unrecorded parents being unrelated), they should be scored separately in the manner just described; the parent, of course, must be scored once only. The less common situation in which the common parent is unrecorded but the other parents of the half-sibs are recorded should be scored as though the record were (s_1+s_2) full sibs with one parent, except that there are two recorded parents to be scored as unrelated individuals: both sibships bear the same relationships to the unrecorded parent, and so may be combined for the information they provide. A third possibility is that both parents of a sibship of s_1 are recorded, together with a sibship of s_2 having one parent common with the first and the other unrecorded. Clearly the set of s_1 sibs then provides no information additional to that from their parents and should be discarded; the second set of sibs should be scored with y exactly as described.

(iii) Neither Parent Recorded

For a sibship of s with neither parent recorded, no score which is a linear function of observed genotype frequencies can be made to give more information by sub-classification of sibships. Cotterman's score, the total number of N genes, may be written in the notation of this paper as

$$y = s-m+n; \qquad . \quad . \quad . \quad . \quad (47)$$

he shows that

$$\mathrm{E}(y) = 2sv \qquad . \quad . \quad . \quad . \quad (48)$$

and that the weight per gene, which may be used for combining scores just as in equation (44), is

$$w = \frac{2}{s+1} \qquad . \quad . \quad . \quad . \quad (49)$$

Values of this weight are also shown in table 3. Since

$$W_{0,\,C}(s) = \frac{4s}{s+1},$$

comparison with equation (24) shows that for $s = 2$ the efficiency of y is

$$\text{Efficiency}\,(y) = \frac{8(1+\gamma)(2+\gamma)}{3(6+5\gamma+4\gamma^2)}, \qquad . \quad . \quad (50)$$

an expression which Cotterman has graphed and which ranges from 0·89 for $\nu = 0$ or 1 to 1·00 for $\nu = 0{\cdot}5$. Though for larger s the efficiency may decrease a little (as was found for the case of one recorded parent) since $W_{0,\,L}(s)$ can never exceed 4, the efficiency of y must always be at least $s/(s+1)$. For very large s the score approaches full efficiency.

For records of one parent and one child or of two children alone, there is no reason why the maximum likelihood scores and weights, described in section 3 and tabulated in tables 1 and 2, should not be used; these may quite legitimately be added to non-efficient scores for higher values of s. Most investigators, however, will prefer for simplicity to use consistently the instructions of this section which give full efficiency for the first case and only a small loss for the second. Whatever procedure is followed, if the estimate ν differs much from the provisional value ν_0, the scoring may be repeated using ν as the provisional value. For extensive data, sufficient to determine ν with a standard error of much less than 0·05, tables 1-3 may be insufficiently detailed. The formulæ of sections 3 and 4 may then be used to calculate scores and weights for intermediate values of ν_0, or interpolation between adjacent entries in the tables may be used, or the data may be scored in full for the two tabular ν-values nearest to ν_0 and interpolations made in the numerator and denominator of equations (14) before evaluating the revised estimate.

Cotterman has given rules for generalising his system of scoring to pairs of relatives of any degree; his method is probably still of high efficiency. No doubt similar methods could be applied to other types of family, but examination of more complex families will not be undertaken here. The scores discussed in this paper are not applicable to factors which show dominance; for them, Cotterman's scoring can again be modified, to rather greater advantage, in a manner to be described in a later paper.

5. NUMERICAL EXAMPLE

Boyd and Boyd (1941), in a report of blood-group studies of various Near Eastern populations, record blood types for 58 families of Armenians from Ghazir and Beyrouth (Syria). Of these families, 5 have both parents classified, 14 have one parent, and 39 consist of sibs only. On the assumption that the records relate to a population mating at random in respect of blood type, the frequency of the N gene may be estimated by the methods of section 4. The first stage of the calculations is shown in table 4.

Part (*a*) of table 4 shows the recorded parental genotypes for families of types (i) or (ii) (sections 3 and 4). In part (*b*), the genotypes of the children are classified, first for families with one recorded parent and secondly for families with no recorded parent. The totals at the bottom of the table show 131 N genes out of a total of 298,

a proportion of about 0·44. A provisional estimate $\nu_0 = 0{\cdot}45$ was therefore used, in conjunction with equations (41) and (47) and

TABLE 4

Estimation of frequency of N gene from records of 58 Armenian families

(*a*) Parents

Frequencies of			Sum of values of	
M	MN	N	W	x
8	12	4	48	20

(*b*) Children (scored for $\nu_0 = 0{\cdot}45$)

s	Recorded parent	No. of families	Numbers of children			Sum of values of		
			M	MN	N	w	y	$2s$
2	MM	4	7	1	0	·267	2·0	16
	MN	6	4	8	0	·267	3·6	24
	NN	1	0	0	2	·267	4·0	4
3	MM	1	0	3	0	·209	6·0	6
	MN	1	0	1	2	·209	7·1	6
4	MN	1	0	1	3	·172	10·2	8
						$\Sigma wy = 7{\cdot}056$		$\Sigma 2ws = 15{\cdot}632$
2	none	26	21	21	10	·267	41·0	104
3	none	11	11	15	7	·500	29·0	66
4	none	2	0	3	5	·400	13·0	16
						$\Sigma wy = 47{\cdot}047$		$\Sigma 2ws = 108{\cdot}768$

table 3, to give the scores and weights in part (*b*) of table 4. For example, the 6 pairs of sibs having a recorded MN parent give a total score

$$\begin{aligned} y &= 4\times 0 + 2\times(12-8)\times 0{\cdot}45 \\ &= 3{\cdot}6 \end{aligned}$$

and the weight per gene for $s = 2$, $\nu_0 = 0{\cdot}45$ in table 3 is 0·267. From the weighted totals of scores, the revised estimate is

$$\begin{aligned} \nu &= \frac{20+7{\cdot}056+47{\cdot}047}{48+15{\cdot}632+108{\cdot}768} \\ &= \frac{74{\cdot}102}{172{\cdot}400} \\ &= 0{\cdot}4298. \end{aligned}$$

Furthermore, the variance of the estimate is

$$V(\nu) = \frac{\nu(1-\nu)}{172{\cdot}4}$$
$$= 0{\cdot}00142.$$

Hence the records lead to an estimate that in the population sampled 43·0 per cent.±3·8 per cent. of the genes are of type N. The difference from the provisional value 45 per cent. is small, so that no re-scoring is necessary; indeed, almost the same result would have been obtained if a provisional value of 40 per cent. had been taken.

6. SUMMARY

The estimation of a population gene frequency from records containing related individuals requires special statistical treatment if the estimate is to be of high efficiency and if an unbiased assessment of its variance is also desired. Scoring of individuals in accordance with rules based on the principle of maximum likelihood fulfils both conditions. In section 3 above, tables of maximum likelihood scores are given for use in the absence of dominance with family records like those discussed by Fisher (1940) in relation to a factor with dominance, namely parent-child pairs and sib-pairs. The complexity of the formulæ is likely to prevent the extension of the method to larger families.

Cotterman (1947) has suggested a scoring system based on a simple count of genes, and has shown this to be of high efficiency. In the present paper, a modification of Cotterman's proposals is shown to lead to a system of " most efficient linear scores " which are always a little more efficient than Cotterman's but very little more trouble to compute. The method is described, and tables to facilitate its use are given, for the estimation of a gene frequency in the absence of dominance from data including sibships with or without parental records. Under these conditions, the differences from Cotterman's method are slight; as will be seen in a subsequent paper, they are more important in the presence of dominance. The method is illustrated by application to data on MN blood types from 58 Armenian families, and an estimate of 43·0 per cent.±3·8 per cent. is obtained for the frequency of the N gene.

REFERENCES

BOYD, W. C., AND BOYD, L. S. 1941.
Data for testing for genetic linkage on 500 pairs of sibs.
Ann. Eugen., Lond. 11, 1-9.

COTTERMAN, C. W. 1947.
A weighting system for the estimation of gene frequencies from family records.
Contributions from the Laboratory of Vertebrate Biology, University of Michigan, No. 33.

FINNEY, D. J. 1940.
The detection of linkage.
Ann. Eugen., Lond. 10, 171-214.

FINNEY, D. J. 1941.
The detection of linkage. III. Incomplete parental testing.
Ann. Eugen., Lond. 11, 115-135.

FINNEY, D. J. 1943.
The detection of linkage. VII. Combination of data from matings of known and unknown phase.
Ann. Eugen., Lond. 12, 31-43.

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

FISHER, R. A. 1935*b*.
The detection of linkage with recessive abnormalities.
Ann. Eugen., Lond. 6, 339-351.

FISHER, R. A. 1940.
The estimation of the proportion of recessives from tests carried out on a sample not wholly unrelated.
Ann. Eugen., Lond. 10, 160-170.

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

STEVENS, W. L. 1938.
Estimation of blood-group gene frequencies.
Ann. Eugen., Lond. 8, 362-375.