THE APPLICATION OF THE "METHOD OF MAXIMUM LIKELIHOOD" TO THE ESTIMATION OF LINKAGE

J. B. HUTCHINSON

Empire Cotton Growing Corporation
Cotton Research Station, Trinidad
and
Rothamsted Experiment Station, Harpenden, England

Received October 4, 1928

TABLE OF CONTENTS

INTRODUCTION

The "Method of Maximum Likelihood" is given by FISHER (1928) and is illustrated by a simple case of linkage between two factors. The method consists of multiplying each class frequency by the natural logarithm of its corresponding probability, summing, and determining the value for which the sum is a maximum. The linkage is expressed in terms of p where p is the proportion of gametes carrying both dominants plus the proportion of gametes carrying both recessives. The method is quite general, and may be applied to any type of ratio from which it is desired to estimate a linkage.

A CASE INVOLVING COMPLEMENTARY FACTORS

BRUNSON (1) has studied maize families in which a factor for pale green seedling is linked with one of two complementary factors for aleurone colour, and he gives a formula, derived from EMERSON'S formula, for the estimation of the linkage.

The F_2 is classified as under, where C and R are complementary factors for aleurone colour, and P_{g1} is a factor for pale green seedling:

TABLE 1

Frequencies observed in an F_2 segregating for aleurone colour and pale green seedling (BRUNSON'S *data*).

	CR	$Cr+cR+cr$	SEEDLING TOTAL
P_{g1}	1907	1053	2960
p_{g1}	300	686	986
Aleurone total	2207	1739	n=3946

Then, considering only the factor P_{g1} and the aleurone factor linked with it, the probabilities of the four classes in F_2 will be:

$$\begin{array}{cccc} P_{g1}A & p_{g1}A & P_{g1}a & p_{g1}a \\ \tfrac{1}{4}(2+pp^1): & \tfrac{1}{4}(1-pp^1): & \tfrac{1}{4}(1-pp^1): & \tfrac{1}{4}pp^1 \end{array}$$

where A is either C or R, and p is the proportion of $(P_{g1}A+p_{g1}a)$ gametes in male gametogenesis, and p^1 is the proportion of $(P_{g1}A+p_{g1}a)$ gametes in female gametogenesis. The effect of the linkage is then entirely expressed in the term pp^1, and if crossing over is equal in male and female, the cross-over percentage will be $(1-\sqrt{pp^1})\times 100$ in the coupling phase, and $(\sqrt{pp^1})\times 100$ in the repulsion phase. For simplicity, θ will be written for pp^1, and since the data provide no evidence on the matter, it will be assumed that crossing over is equal in male and female. (See FISHER (1928)).

Then, bringing in the complementary factor for aleurone colour, we get the probabilities in the four classes:

$$\begin{array}{cccc} CRP_{g1} & CRp_{g1} & [Cr+cR+cr]P_{g1} & [Cr+cR+cr]p_{g1} \\ \frac{3}{16}(2+\theta) & \frac{3}{16}(1-\theta) & \frac{3}{16}(2-\theta) & \frac{1}{16}(1+3\theta) \end{array}$$

Then the logarithm of the likelihood will be:

$$L=1907\log\frac{3}{16}(2+\theta)+300\log\frac{3}{16}(1-\theta)+1053\log\frac{3}{16}(2-\theta)+686\log\frac{1}{16}(1+3\theta). \tag{1}$$

And the maximum likelihood value of θ will be that for which the first differential of equation 1, with respect to θ, is zero:

$$=\frac{1907}{2+\theta}-\frac{300}{1-\theta}-\frac{1053}{2-\theta}+\frac{2058}{1+3\theta}=0. \tag{2}$$

Which becomes on multiplying out:

$$11838\theta^3-12802\theta^2-11376\theta+8740=0.$$

An equation of the third degree, which may be solved by HORNER'S method (BURNSIDE and PANTON 1886), or by the method of successive trials developed on page 532.

In the present case the solution is

$$\theta=0.5902.$$

Then, assuming equal crossing over in male and female

$$p=\sqrt{\theta}=0.7682$$

or 23.18 percent crossing over, with coupling.

The variance of this estimate of θ may be obtained by differentiating equation 2 again with respect to θ, substituting the expected values for the class frequencies, and equating to $-1/V(\theta)$

$$-\frac{3n}{16}\cdot\frac{2+\theta}{(2+\theta)^2}-\frac{3n}{16}\cdot\frac{1-\theta}{(1-\theta)^2}-\frac{3n}{16}\cdot\frac{2-\theta}{(2-\theta)^2}-\frac{9n}{16}\cdot\frac{1+3\theta}{(1+3\theta)^2}=-\frac{1}{V(\theta)}$$

or

$$\frac{1}{V(\theta)}=\frac{3n}{16}\left(\frac{1}{2+\theta}+\frac{1}{1+\theta}+\frac{1}{2-\theta}+\frac{3}{1+3\theta}\right) \tag{3}$$

$$V(\theta)=\frac{4}{3n}\cdot\frac{(2+\theta)(1-\theta)(2-\theta)(1+3\theta)}{5+2\theta-4\theta^2}\,. \tag{4}$$

Then, since the variance of θ is the mean square deviation of all θ's from the mean of θ, and the variance of p is the mean square deviation of the p's from the mean of p, and since θ equals p^2, we can calculate the variance of p from the variance of θ.

$$V(\theta)=(2p)^2\cdot V(p)$$

and

$$V(p)=\frac{(2+p^2)(1-p^2)(2-p^2)(1+3p^2)}{3np^2(5+2p^2-4p^4)}\,. \tag{5}$$

Substituting for p we get

$$V(p)=0.0001240$$
$$\sigma(p)=0.011$$

BRUNSON'S formula may be reduced to

$$p_B{}^2=\frac{16}{18n}\cdot(a-b-c+3d)$$

where a, b, c, and d are the observed frequencies in the four classes, CRP_{o1}, CRp_{o1}, $[Cr+cR+cr]P_{o1}$, and $[Cr+cR+cr]p_{o1}$ respectively.

BRUNSON finds

$$p_B=0.767.$$

The variance of this estimate may be found from the general equation given by FISHER (1928).

$$\frac{1}{n}V(T)=S\left\{p\left(\frac{dT}{da}\right)^2\right\}-\left(\frac{dT}{dn}\right)^2 \tag{6}$$

where T is any function of the frequencies, and p is the probability of the corresponding class a, b, c, d. For convenience T_B may be used in place of p_B^2, and the variance of p_B found from that of T_B as above.

Then

$$T_B=\frac{16}{18n}(a-b-c+3d).$$

Taking the components of

$$S\left\{p\left(\frac{dT}{da}\right)^2\right\}$$

$$\frac{dT}{da}=\frac{16}{18n}\quad;\qquad\left(\frac{dT}{da}\right)^2=\frac{64}{81n^2}$$

$$\frac{dT}{db}=-\frac{16}{18n};\qquad\left(\frac{dT}{db}\right)^2=\frac{64}{81n^2}$$

$$\frac{dT}{dc}=-\frac{16}{18n};\qquad\left(\frac{dT}{dc}\right)^2=\frac{64}{81n^2}$$

$$\frac{dT}{dd}=\frac{48}{18n}\quad;\qquad\left(\frac{dT}{dd}\right)^2=\frac{64}{9n^2}$$

and

$$S\left\{p\left(\frac{dT}{da}\right)^2\right\}=(pa+pb+pc+9pd)\frac{64}{81n^2}$$

$$=\frac{24(1+\theta)}{16}\times\frac{64}{81n^2}=\frac{32(1+\theta)}{27n^2}.$$

And the component

$$\left(\frac{dT}{dn}\right)^2=\left(-\frac{16(a-b-c+3d)}{18n^2}\right)^2$$

$$=\left(\frac{-18n\theta}{18n^2}\right)^2=\frac{\theta^2}{n^2}$$

Then

$$\frac{1}{n}V(T_B)=\frac{32+32\theta-27\theta^2}{27n^2}$$

$$V(T_B)=\frac{32+32\theta-27\theta^2}{27n}$$

Then since

$$T_B = p_B{}^2\ ; \quad V(p_B) = \frac{V(T_B)}{4/22}$$

as before, and

$$V(p_B) = \frac{32+32p^2-27p^4}{108np^2}\ .$$

Substituting the most likely value of p we get

$$V(p_B) = 0.000165$$

$$\sigma(p_B) = 0.013.$$

The efficiency of BRUNSON'S formula may be estimated by dividing the variance of the Maximum Likelihood equation by the variance of BRUNSON'S formula.

$$E = \frac{(2+p^2)(1-p^2)(2-p^2)(1+3p^2)}{3np^2(5+2p^2-4p^4)} \div \frac{32+32p^2-27p^4}{108np^2}$$

$$= \frac{36(2+b^2)(1-p^2)(2-p^2)(1+3p^2)}{(5+2p^2-4p^4)(32+32p^2-27p^4)}\ .$$

For the distribution of E for values of p from 0 to 1, see figure 1. The formula is about 90 percent efficiency throughout the repulsion phase, but compares badly with the Maximum Likelihood equation for high coupling. Nowhere does it give complete efficiency.

The expected frequencies may be found by substituting the most likely value of θ in the probabilities and multiplying by n =3496.

TABLE 2

Comparison of observation with expectation, obtained by two different solutions (BRUNSON'S *data*).

		CRP_{θ_1}	CRp_{θ_1}	$(Cr+cR+cr)\,P_{\theta_1}$	$(Cr+cR+cr)\,p_{\theta_1}$	n
Observed		1907	300	1053	686	3946
Expected	*M. L.*	*1916.42*	*303.20*	*1043.08*	*683.30*	*3946.00*
	BR.	*1915*	*305*	*1044*	*682*	*3946*

χ^2 from Maximum Likelihood expectation =0.185.

χ^2 from Brunson's expectation =0.2165.

In each case there are 2 degrees of freedom from which to determine P.[1]

[1] FISHER (1922, 1923 and 1924) has shown that when a population, with which a sample is to be compared, has itself been reconstructed from the sample, the distribution of χ^2 is not known simply from the number of frequency classes. When using ELDERTON'S tables, the table must be entered with n^1 equal to *one more than the number of degrees of freedom* in which the sample may

Any number of formulae for the estimation of p may be invented. See Fisher (1928, in press).

The amount of information per plant obtained from the F_2 and its corresponding backcross may be compared with the amount of information obtainable from a simple backcross, that is, one in which complete classification is possible.

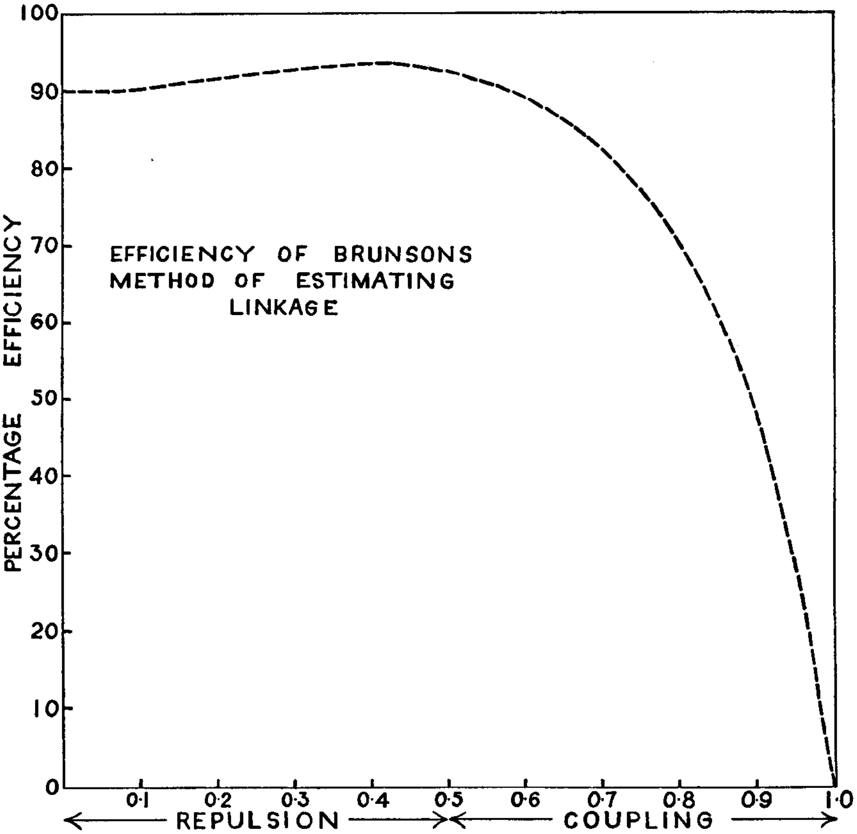

Figure 1.—Efficiency of Brunson's method of estimating linkage in an F_2 involving a factor linked to one of two complementary factors.

For a simple backcross:

$$V(p) = \frac{p(1-p)}{n}.$$

differ from the reconstructed population. Doctor Fisher (1928) gives a table of χ^2 which is more convenient for use than that given by Elderton, and in using Fisher's table, the table is entered with *n equal to the number of degrees of freedom* in which the sample may differ from the reconstructed population.

Therefore the amount of information concerning p available per plant is

$$\frac{I(p)}{n}=\frac{1}{p(1-p)}\,.$$

In a backcross between an F_1 of the type under consideration and its triple recessive, the probabilities of the four classes will be:

$$\begin{array}{cccc} CRP_{o_1} & CRp_{o_1} & (Cr+cR+cr)P_{o_1} & (Cr+cR+cr)p_{o_1} \\ \frac{p}{4} & \frac{1-p}{4} & \frac{2-p}{4} & \frac{1+p}{4} \end{array}$$

And the maximum likelihood value of p is that for which

$$\frac{a}{p}-\frac{b}{1-p}-\frac{c}{2-p}+\frac{d}{1+p}=0$$

and the variance of p

$$\frac{1}{V(p)}=\frac{n}{4}\left(\frac{1}{p}+\frac{1}{1-p}+\frac{1}{2-p}+\frac{1}{1+p}\right)$$

$$V(p)=\frac{2p(1-p)(2-p)(1+p)}{n(1+2p-2p^2)}\,.$$

Then the amount of information concerning p available per plant is the reciprocal of the variance, divided by n

$$\frac{I(p)}{n}=\frac{1+2p-2p^2}{2p(1-p)(2-p)(1+p)}\,.$$

The amount of information available per plant relative to that obtainable from a simple backcross is

$$\frac{1+2p-2p^2}{2p(1-p)(2-p)(1+p)}\div\frac{1}{p(1-p)}=\frac{1+2p-2p^2}{2(2-p)(1+p)}\,.$$

A similar procedure may be applied to the Maximum Likelihood solution for the F_2 and to BRUNSON's formula. The amount of information concerning p made available per F_2 plant by the Maximum Likelihood equation is

$$\frac{I(p)}{n}=\frac{3p^2(5+2p^2-4p^4)}{(2+p^2)(1+p)(1-p)(2-p^2)(1+3p^2)}\,.$$

Dividing by $1/p(1-p)$ we get the amount of information per F_2 plant relative to that supplied by a simple backcross

$$\frac{3p^3(5+2p^2-4p^4)}{(2+p^2)(1+p)(2-p^2)(1+3p^2)}\,.$$

The amount of information concerning p made available per F_2 plant by BRUNSON'S formula is

$$\frac{I(p_B)}{n}=\frac{108p^2}{32+32p^2-27p^4}.$$

Dividing by $1/p(1-p)$, we get the amount of information per F_2 plant relative to that supplied by a simple backcross

$$\frac{108p^3(1-p)}{32+32p^2-27p^4}.$$

These amounts of information are plotted as percentages for values of p from 0 to 1 in figure 2. It will be seen that for values of p from 0.7 to 1.0—

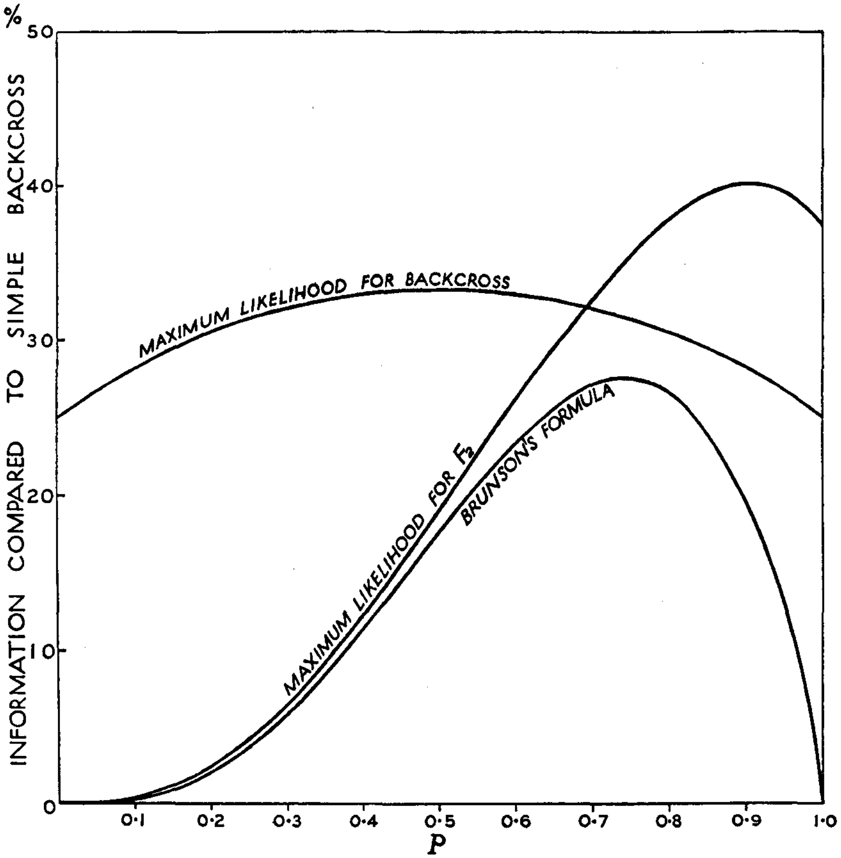

FIGURE 2.—A factor linked to one of two complementary factors: Amount of information concerning linkage supplied per plant by a backcross to a triple recessive, and by an F_2 using (a) Maximum Likelihood solution, (b) Brunson's solution. Amount of information per plant expressed as a percentage of that supplied by a simple backcross.

that is, 30 percent or less crossing over with coupling—an F_2 gives a better estimate of linkage than a backcross to the triple recessive. A simple backcross can, of course, be obtained by backcrossing to a double recessive, which is homozygous dominant for the independent complementary factor. In the present instance, backcrossing is impossible, because one of the linked factors is lethal in the recessive phase. In any case, it is not known which of the complementary factors is independent.

The amount of information supplied by the backcross is never more than 1/3 of what would be available if the classification could be completed. F_2s give very little information concerning linkage in the repulsion phase, and at best, in the coupling phase, give 40 percent of the information that would be supplied by a simple backcross of the same size.

A CASE INVOLVING DUPLICATE FACTORS

WOODWORTH (1921) has studied a case in soy beans in which one of two duplicate factors, I and D, for cotyledon colour is linked with a factor V for seed coat colour, and gives the figures given in table 3 for his F_2, after correcting his data to 15:1 for cotyledon colour. This correction is made necessary because seeds borne on the F_1 plants were classified for cotyledon colour, while seeds borne on F_2 plants were classified for seed coat colour.

TABLE 3
F_2 distribution for cotyledon colour (corrected to 15:1 nearest whole number) and seed coat colour (WOODWORTH'S *data*).

	$ID+Id+iD$	id	COAT COLOUR
V	150	14	164
v	64	0	64
Cotyledon colour	214	14	n=228

The probabilities of the four classes are given in table 4, where θ expresses the effect of the linkage, and, as before, is equal to p^2 if crossing over is the same in male and female.

TABLE 4
Probabilities of the four classes where θ expresses the linkage between either I or D, and V.

	$ID+Id+iD$	id	COAT COLOUR
V	$\frac{11+\theta}{16}$	$\frac{1-\theta}{16}$	$\frac{3}{4}$
v	$\frac{4-\theta}{16}$	$\frac{\theta}{16}$	$\frac{1}{4}$
Cotyledon colour	$\frac{15}{16}$	$\frac{1}{16}$	$n=1$

Then the most likely value of θ is that for which the first differential of the likelihood equation is 0,

$$\frac{150}{11+\theta}-\frac{64}{4-\theta}-\frac{14}{1-\theta}+\frac{0}{\theta}=0.$$

Then, clearly, the most likely value of θ is 0,—that is, complete linkage with repulsion.

The variance of this estimate is obtained, as before, by differentiating again, substituting the probabilities for the observed frequencies, and equating to $-1/V(\theta)$

$$\frac{1}{V(\theta)}=\frac{n}{16}\left\{\frac{1}{11+\theta}+\frac{1}{4-\theta}+\frac{1}{1-\theta}+\frac{1}{\theta}\right\}$$

$$V(\theta)=\frac{4(11+\theta)(4-\theta)(1-\theta)\theta}{n(11+2\theta-4\theta^2)}\cdot$$

Then, since $\theta=p^2$ we can derive the variance of p as before.

$$V(p)=\frac{(11+p^2)(4-p^2)(1-p^2)}{n(11+2p^2-4p^4)}\cdot$$

Substituting $\theta=0$ we get

$$V(p)=\frac{44}{11\times 228}=0.01754$$

$$\sigma(p)=0.132.$$

The F_2 evidence indicates, therefore, 0 percent crossing over, with a standard deviation of 13.2 percent. WOODWORTH (1921 and 1923) gives equations[2] for the estimation of linkage, which may be reduced to:

$$p_w{}^2=\theta_w=\frac{a+b-15c+d}{n}\cdot$$

On applying this to the corrected data, WOODWORTH gets 13.25 percent crossing over. This, however, he shows to be due to the fact that the correction is to the nearest whole number, only, and when the exact corrected figures are taken, there is no evidence of crossing over.

The variance of θ_w may be obtained by applying equation 6 as before.

[2] WOODWORTH (1921) gives: $r=.25\sqrt{a+b+d+15c}$.
This is clearly an error for $r=.25\sqrt{a+b+d-15c}$.

$$S\left\{p\left(\frac{d\theta w}{da}\right)^2\right\}=\frac{11+\theta}{16}\times\frac{1}{n^2}+\frac{4-\theta}{16}\times\frac{1}{n^2}+\frac{1-\theta}{16}\times\frac{225}{n^2}+\frac{\theta}{16}\times\frac{1}{n^2}$$

$$=\frac{15-14\theta}{n^2}$$

$$\left(\frac{d\theta}{dn}\right)^2=\left(\frac{-a-b+15c-d}{n^2}\right)^2$$

$$=\left(\frac{-16n\theta}{16n^2}\right)^2$$

$$=\frac{\theta^2}{n^2}\,.$$

Then

$$\frac{1}{n}V(\theta_w)=\frac{15-14\theta-\theta^2}{n^2}$$

Then, since $\theta_w=p_w^2$, as before, the variance of p_w will be

$$V(p_w)=\frac{15-14p^2-p^4}{4np^2}\,.$$

Then, since $p=0$, in this case the variance of WOODWORTH'S estimate is infinitely large.

The expected numbers are obtained by substituting $\theta=0$ in the probabilities, and multiplying by $n=228$. Having corrected the ratio for cotyledon colour, and fitted an estimate of p, only 1 degree of freedom remains, and P =about 0.28. (See footnote to page 523.)

TABLE 5

Observed and expected frequencies for cotyledon colour and seed coat colour in soy beans. $\chi^2=1.1548$.

	$[ID+Id+iD]V$	$[ID+Id+iD]v$	idV	idv	n
Observed	150	64	14	0	228
Expected	*156.75*	*57.00*	*14.25*	*0.00*	*228*
Deviations	−6.75	+7.0	−0.25	0	

The probabilities of the classes in a backcross involving duplicate genes, one of which is linked to a third gene, will be

$$[ID+Id+iD]V:[ID+Id+iD]v:idV:idv$$

$$\frac{1+p}{4}\qquad\frac{2-p}{4}\qquad\frac{1-p}{4}\quad\frac{p}{4}$$

and the variance of p will be

$$V(p)=\frac{2p(1-p)(2-p)(1+p)}{n(1+2p-2p^2)}$$

the same as for complementary factors.

The amounts of information concerning p available per plant, compared to a simple backcross are calculated as before and give

Maximum Likelihood equation for F_2

$$\frac{p(11+2p^2-4p^4)}{4(11+p^2)(4-p^2)(1+p)}$$

WOODWORTH's formula for F_2

$$\frac{4p^3(1-p)}{15-14p^2-p^4}.$$

Maximum Likelihood equation for backcross

$$\frac{(1+2p-2p^2)}{2(2-p)(1+p)}.$$

These amounts of information are plotted as percentages of that obtainable from a simple backcross in figure 3. It is quite clear that very little information concerning linkage can be obtained from an F_2 involving duplicate factors under any circumstances, and such F_2s should always be carried on to F_3. A backcross to the triple recessive gives from 1/4 to 1/3 of the information per plant that would be obtained from a simple backcross.

TABLE 6

Soy bean F_2 segregating for seed coat colour and cotyledon colour: yellow cotyledon plants only included and these analysed according to F_3 behaviour (WOODWORTH's *data*).

	TRUE BREEDING YELLOWS	SPLITTING YELLOWS		COAT COLOUR
		Double Het.	Single Het.	
V	57	45	45	147
v	55	3	5	63
Cotyledon colour	112	48	50	n=210

WOODWORTH is able to classify the 210 plants which came from embryos having yellow cotyledons into: (1) plants homozygous dominant for I or D or both; (2) plants segregating for both I and D; and (3) plants homozygous recessive for either I or D and segregating for either D or I. (See table 6.) Then the probabilities of the classes will be as given in table 7.

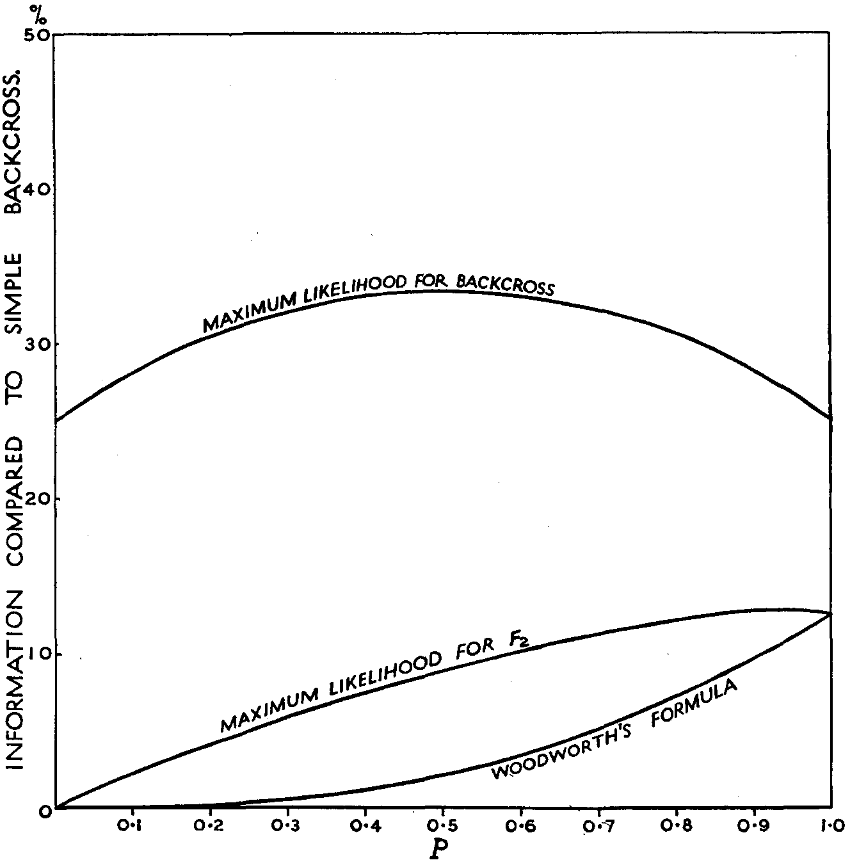

FIGURE 3.—A factor linked to one of two duplicate factors: Amount of information concerning linkage supplied per plant by a backcross to a triple recessive, and by an F_2, using (a) Maximum Likelihood solution, (b) Woodworth's solution. Amount of information per plant expressed as a percentage of that supplied by a simple backcross.

TABLE 7

Probabilities of the classes involved in the soy bean segregation tabulated in table 6.

	TRUE BREEDING YELLOWS	SPLITTING YELLOWS		COAT COLOUR
		Double Het.	Single Het.	
V	$\frac{3}{15}(1+2p-p^2)$	$\frac{4}{15}(1-p+p^2)$	$\frac{2}{15}(2-p)$	$\frac{11+p^2}{15}$
v	$\frac{1}{15}4-6p+3p^2)$	$\frac{4}{15}p(1-p)$	$\frac{2p}{15}$	$\frac{4-p^2}{15}$
Cotyledon colour	$\frac{7}{15}$	$\frac{4}{15}$	$\frac{4}{15}$	$n=1$

Then the garitholm of the likelihood will be

$$L=57\log\frac{3}{15}(1+2p-p^2)+45\log\frac{4}{15}(1-p+p^2)+45\log\frac{2}{15}(2-p)$$

$$+55\log\frac{4-6p+3p^2}{15}+3\log\frac{4p}{15}(1-p)+5\log\frac{2p}{15}\cdot$$

And the most likely value of p will be that for which

$$\frac{114(1-p)}{1+2p-p^2}-\frac{45(1-2p)}{1-p+p^2}-\frac{45}{2-p}-\frac{330(1-p)}{4-6p+3p^2}+\frac{8}{p}-\frac{3}{1-p}=x=0.$$

This equation may be multiplied up and solved by HORNER'S method, but an easier method is to obtain the value of p by successive trials. This method has the advantage that it avoids all algebraic manipulation, and provides not only an estimate of the linkage, but also an estimate of its standard error. A good guess for the first trial can usually be made after inspecting the data. In this example the F_2 indicated no crossing over, while the data in table 6 show that repulsion is not complete. A fair first trial is therefore $p=0.1$. Since the amount of information $[=1/V(p)]$ is the rate of change of x relative to p, the two final approximations, if worked out to sufficient accuracy, will give an estimate of the variance of p. The two final approximations used were $p=0.144$ and $p=0.115$.

When

$$p=0.114 \qquad x=+0.26904$$

$$p=0.115 \qquad x=-0.54503.$$

Therefore p is approximately 0.1143, or 11.43 percent crossing over. And for a change of 0.001 in p, x changes 0.81407, and the rate of change of $p=1/V(p)=814.07$

$$V(p)=0.001228$$

$$\sigma(p)=0.03504.$$

This may be checked by differentiating the likelihood equation a second time, equating to $-1/V(p)$ and substituting $p=0.11433$

$$\frac{1}{V(p)}=\frac{114(3-2p+p^2)}{(1+2p-p^2)^2}-\frac{45(1+2p-2p^2)}{(1-p+p^2)^2}+\frac{45}{(2-p)^2}+\frac{330(2-6p+3p^2)}{(4-6p+3p^2)^2}$$

$$+\frac{8}{p^2}+\frac{3}{(1-p)^2}=816.05$$

$$V(p)=0.001225$$

$$\sigma(p)=0.03500.$$

The two estimates of the variance differ by less than 0.5 percent.

We have, therefore, 11.43 percent crossing over with a standard deviation of 3.5 percent.

CONSIDERATION OF THEORY WHEN COMPLEMENTARY OR DUPLICATE FACTORS ARE THEMSELVES LINKED

These cases are rare, but may be considered for the sake of completeness.

(a) *Complementary Factors*

The probabilities of the two classes will be

$$\frac{2+\theta}{4}:\frac{2-\theta}{4}$$

where, as before, $\theta=p^2$ if crossing over is the same in male and female.

The most likely value of θ will be that for which

$$\frac{a}{2+\theta}-\frac{b}{2-\theta}=0$$

or

$$p^2=\theta=\frac{2(a-b)}{n}\cdot$$

And

$$\frac{1}{V(\theta)}=\frac{n}{4}\left(\frac{1}{2+\theta}+\frac{1}{2-\theta}\right)$$

$$V(\theta)=\frac{(2+\theta)(2-\theta)}{n}\cdot$$

And since $\theta=p^2$, we calculate the variance of p as before

$$V(p)=\frac{(2+p^2)(2-p^2)}{4p^2n}\cdot$$

And the amount of information per plant compared to a (theoretical) simple backcross will be the reciprocal of this divided by $n/p(1-p)$

$$\frac{4p^3(1-p)}{4-p^4}$$

(b) *Duplicate Factors*

The probabilities of the two classes will be

$$\frac{4-\theta}{4}:\frac{\theta}{4}\cdot$$

The most likely value of θ will be that for which

$$\frac{b}{\theta}-\frac{a}{4-\theta}=0\ .$$

or

$$p^2=\theta=\frac{4b}{n}\cdot$$

And differentiating again, substituting probabilities for a and b and equating to $-1/V(\theta)$ we get

$$\frac{1}{V(\theta)}=\frac{n}{4}\left\{\frac{1}{\theta}+\frac{1}{4-\theta}\right\}$$

$$V(\theta)=\frac{\theta(4-\theta)}{n}$$

$$V(p)=\frac{4-p^2}{4n}\cdot$$

And the amount of information per plant compared with a simple backcross will be

$$\frac{4p(1-p)}{4-p^2}\cdot$$

The probabilities of the classes in the two backcrosses will be

$$\text{Complementary}\quad \frac{p}{2}[AB]:\frac{2-p}{2}[Ab+aB+ab]$$

$$\text{Duplicate}\quad \frac{2-p}{2}[AB+Ab+aB]:\frac{p}{2}ab.$$

And the most likely values of p will be of those for which

$$\frac{a}{p}-\frac{b}{2-p}=0 \text{ for Complementary Factors.}$$

and

$$\frac{b}{p}-\frac{a}{2-p}=0 \text{ for Duplicate Factors.}$$

The variance of p is the same in each case:

$$\frac{1}{V(p)}=\frac{n}{2}\left(\frac{1}{p}+\frac{1}{2-p}\right)$$

$$=\frac{n}{p(2-p)}$$

$$V(p)=\frac{p(2-p)}{n}$$

Then the amount of information per plant relative to a simple backcross will be the reciprocal of the variance divided by $n/p(1-p)$

$$=\frac{1-p}{2-p}.$$

These amounts of information are plotted as percentages of the amount of information supplied by a simple backcross for values of p from 0 to 1. See figure 4.

Very little information concerning linkage can be obtained from an F_2 involving linked complementary factors. Where there is a considerable

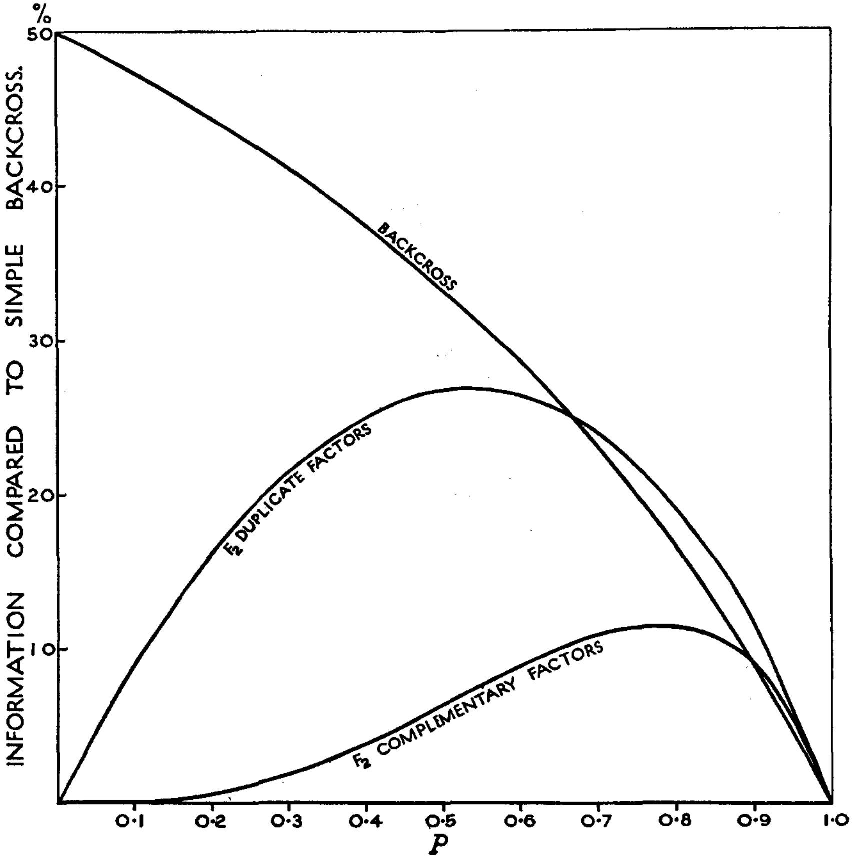

FIGURE 4.—Complementary factors themselves linked, and duplicate factors themselves linked. Amount of information concerning linkage supplied per plant by a backcross to a double recessive, and by F_2 using Maximum Likelihood solution. Amount of information per plant expressed as a percentage of that supplied by a simple backcross.

amount of crossing over—that is, where p is near 0.5—an F_2 involving linked duplicate factors will give about 1/4 of the information that would be given by a simple backcross of the same size. The backcrosses give up to 1/2 of the information obtainable from a simple backcross with high linkage in the repulsion phase, but become less efficient than the corresponding F_2s with high coupling.

SUMMARY

1. The "Method of Maximum Likelihood" developed by DOCTOR R. A. FISHER is applied to the problem of estimating linkage in cases involving complementary and duplicate factors.

2. Variances are calculated for existing formulae, and their efficiencies are determined to show that the "Method of Maximum Likelihood" is in all cases superior to any other method of estimation.

3. The amount of information supplied per plant by Maximum Likelihood formulae for F_2s and backcrosses, and by other formulae for F_2s is calculated and compared with the amount of information supplied per plant by a simple—that is, completely classified—backcross. (See figures 2, 3 and 4.) From these curves it is possible to estimate the size of family necessary to give any required degree of accuracy.

ACKNOWLEDGMENTS

The investigation here presented was carried out at ROTHAMSTED EXPERIMENTAL STATION under the guidance of DOCTOR R. A. FISHER, during a period of study leave from the Genetics Department of the COTTON RESEARCH STATION, Trinidad, B. W. I. The author is indebted to the EMPIRE COTTON GROWING CORPORATION for granting the leave, and to DOCTOR R. A. FISHER for suggesting the problem, and for the care and patience with which he supervised both its general development, and its detailed application.

Since the above was written a paper on the calculation of linkage intensities by F. V. OWEN has come under notice. FISHER'S Method of Maximum Likelihood is considered, but its particular virtue, namely, that, in the theory of large samples, no statistic can have a smaller sampling variance, is overlooked; see FISHER (1928).

LITERATURE CITED

BRUNSON, A. M., 1924 The inheritance of a lethal pale green seedling character in maize. Cornell Univer. Agric. Exp. Sta. Memoir 72: 1924. Cornell Univ. Press, Ithaca, N. Y.

BURNSIDE and PANTON, 1886 Theory of equations. 2nd Edition, pp. 209.

FISHER, R. A., 1922 On the interpretation of χ^2 from contingency tables and the calculation of P Jour. of the Royal Statistical Soc. **85:** Part 1, 88–94.

1923 Statistical tests of agreement between observation and hypothesis. Economica **3:** 139–147.

1924 The conditions under which χ^2 measures the discrepancy between observation and hypothesis. Jour. of the Royal Statistical Soc. **87:** Part 3.

1928 Statistical methods for research workers. 2nd Edition, Oliver and Boyd.

1928 (In Press) On a property connecting the χ^2 measure of discrepancy with the method of maximum likelihood. Bologna: International Congress of Mathematics.

WOODWORTH, C. M., 1921 Inheritance of cotyledon, seed-coat, hilum, and pubescence colors in soy beans. Genetics **6:** 487–553.

1923 The calculation of linkage intensities where duplicate factors are concerned. Genetics **8:** 106–115.

OWEN, F. V., 1928 Calculation of linkage intensities by product moment correlation. Genetics **13:** 80–110.