Heredity (1984), **53**(3), 495-507

THE POPULATION GENETICS OF THE SELF-INCOMPATIBILITY POLYMORPHISM IN *PAPAVER RHOEAS*. IV. THE ESTIMATION OF THE NUMBER OF ALLELES IN A POPULATION

S. O'DONNELL AND M. J. LAWRENCE

Department of Genetics, University of Birmingham, Birmingham B15 2TT

Received 6.i.84

SUMMARY

Three methods of estimating the number of *S*-alleles in a population have been proposed in the literature (Bateman, 1947; Whitehouse, 1949; Paxman, 1963). These methods, all of which assume that genotype frequencies in the population are equal, are described briefly and are used to estimate the number of alleles in populations of *Trifolium pratense* (Williams and Williams, 1947), *Oenothera organensis* (Emerson, 1939) and *Papaver rhoeas* (Campbell and Lawrence, 1981*b*; Lawrence and O'Donnell, 1981). The estimates yielded by Bateman's and Whitehouse's methods are similar to those given by Paxman's maximum likelihood method with the Trifolium and Oenothera data where there is little reason to suppose that the allele frequencies are other than equal. Bateman's method, however, breaks down when used on the Papaver data in which the *S*-allele frequencies are known to be unequal; and Whitehouse's and the maximum likelihood methods yield estimates which are biased downwards when used on these data.

An attempt has been made, therefore, to devise two new estimators of the number of *S*-alleles in a population which do not assume that their frequencies are equal. The properties of these estimators has been investigated with data from eight populations generated on the computer in which the numbers and frequencies of alleles are known. One of these new estimators (E_2) yields estimates which are less biased downwards than those given by Paxman's method when allele frequencies are unequal, but gives estimates which are biased upwards when these frequencies are equal. The other estimator (E_1) is generally less satisfactory than the first, particularly when the number of alleles in the population is large. Though neither of these new estimators are wholly satisfactory, there is some justification for using E_2 when allele frequencies are known to be unequal. Estimates given by E_2 when used on the Papaver data range from 34 to 42 alleles which, bearing in mind that these estimates are still likely to be biased downwards, suggests that the number of alleles in natural populations of this species is likely to be between 40 and 45.

A new procedure for calculating confidence intervals for maximum likelihood estimates, assuming equal allele frequencies, is also described and applied to the Oenothera and Papaver data.

1. INTRODUCTION

The problems with which this paper is concerned is: given that, for a species with a one gene, multi-allelic, gametophytic system of self-incompatibility, *n* different *S*-alleles have been found in a sample of *m* plants, what is the best way to estimate the number of alleles, *N*, in the population from which the sample has been drawn?

The first to consider this problem was Bateman (1947) who argued that the probability of a random pair of plants having an allele in common, assuming that all genotypes in the population are equally frequent, is approximately $4/N$. An empirical estimate of this probability can be obtained from the ratio $2n_p/m(m-1)$, where n_p is the number of pairs of plants in the sample that have at least one allele in common. An estimate of N can then be obtained by equating this ratio to $4/N$; that is

$$\frac{4}{N}=\frac{2n_p}{m(m-1)}. \tag{1}$$

Using this estimator Bateman estimated that there were 171 different S-alleles in one variety and 308 in a second variety of *Trifolium pratense* (data of Williams and Williams, 1947). In fact, if the genotypes in the population are equally frequent, the probability of a pair of plants having an allele in common is $4/N-2/N(N-1)$, rather than $4/N$. However, for the clover data, the second term in this expression can be justifiably ignored. It should also be pointed out that the expected value of the ratio $2n_p/m(m-1)$ is only approximately $4/N$.

Whitehouse (1949) gave a method of estimating the number of alleles in a population from the number occurring in a random sample of fruitbodies of a heterothallic fungus; his method can also be used to estimate the number of S-alleles in a population of flowering plants. Treating m and n as continuous variables, Whitehouse showed that, on average, assuming all genotypes in the population are equally frequent

$$\frac{dn}{dm}=\frac{2(N-n)}{N}$$

Then

$$m=\int\frac{N}{2(N-n)}\,dn$$

and

$$\frac{n}{2m}=\frac{n}{N}\frac{1}{\log_e\left(\frac{1}{1-n/N}\right)} \tag{2}$$

from which an estimate of N can be obtained by iteration. One aspect of Whitehouse's solution to this problem, which is particularly attractive, is that dn/dm is the slope of the graph obtained by plotting the number of alleles found against the number of plants examined during the course of an experiment (fig. 1); the gradient of this curve becomes zero, of course, when $n=N$.

Though Paxman (1963) gave the first account of the maximum likelihood solution to this problem, a comment by Fisher (1947) on Bateman's (1947) note leaves no doubt that this solution had been discovered fifteen years earlier; however, characteristically, Fisher gave no details of the method. Paxman showed that the expected number of alleles in a sample of size m, again assuming all genotypes in the population have the same frequency, is:

$$\mathrm{E}n=N\{1-(1-2/N)^m\}. \tag{3}$$

The maximum likelihood estimate of N can be obtained from this expression by equating En to n.

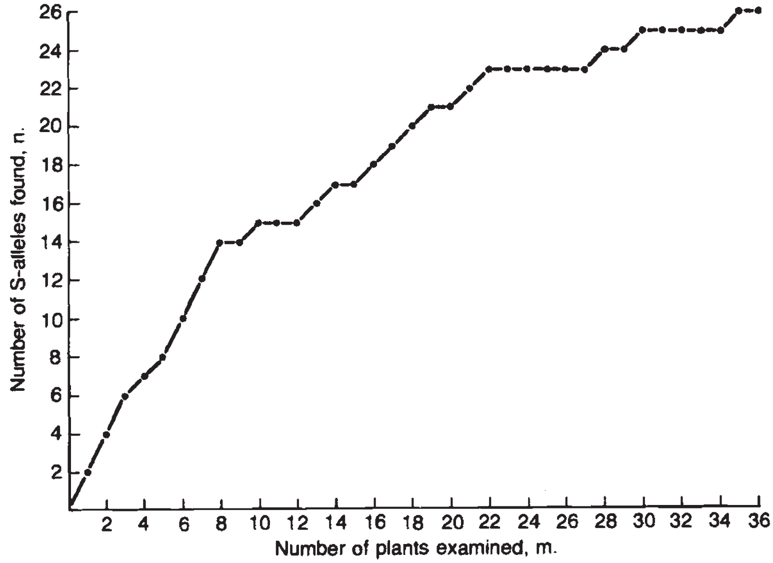

FIG. 1. The graph obtained by plotting the number of alleles found, *n*, against the number of plants examined, *m*, for the R104 sample of *Papaver rhoeas*. Data from table 1 of Lawrence and O'Donnell (1981).

We are indebted to a colleague, Dr Paul Davies, for pointing out that the problem of estimating the number of *S*-alleles in a population is a special case of the more general problem of estimating the size of a closed population by the capture-recapture method. For example, with the Schnabel census (Schnabel, 1938; see Seber, 1982), a series of *m* samples are taken from the population, each sample (except the first) examined for marked individuals and the sample returned to the population after each individual has been re-marked (or marked). A sample of *m* plants from a self-incompatible population of infinite size is equivalent to a series of *m* samples, each of size 2, from a finite population where sampling is by replacement. Chapman (1952) and Darroch (1958) have shown that for the general Schnabel census, the maximum likelihood estimate of *N* can be obtained by solving the polynomial

$$\left(1-\frac{r}{N}\right)=\prod_{i=1}^{m}\left(1-\frac{x_i}{N}\right) \tag{4}$$

where r is the number of different individuals found in m samples and x_i is the size of the *i*th sample. When $r=n$ and all $x_i=2$, this equation is identical to (3).

The estimates obtained by using Bateman's (B), Whitehouse's (W) and the maximum likelihood (ML) method on the red clover data, on Emerson's (1939) data from the *Oenothera organensis* population and on our own data from three natural populations of *Papaver rhoeas* (Campbell and Lawrence, 1981*b*; Lawrence and O'Donnell, 1981) are shown in table 1. There is little to choose between the estimators when used on the clover and Oenothera data, though on general grounds, the maximum likelihood method must be regarded as the best. However, Bateman's method breaks down completely,

TABLE 1

Estimates of the number of S-alleles, $\hat{\mathrm{N}}$, *in populations of the indicated species by Bateman's* (B), *Whitehouse's* (W) *and the maximum likelihood* (ML) *methods. Values of* N *shown in the table are given to the nearest integer. The repeatability,* R = (2m − n)/(2m − 3) *gives a measure of the thoroughness of an analysis* (*Campbell and Lawrence, 1981a*)

Species	Source of data	*m*	*n*	*R*	$\hat{N}$			99% confidence interval
					B	W	ML	
Trifolium pratense: 1	Bateman (1947)*	25	43	0·15	171	162	**156†**	—
Trifolium pratense: 2	Bateman (1947)*	22	41	0·07	308	308	**294†**	—
Oenothera organensis	Emerson (1939)	37	28	0·65	32	31	**30**	**28–37**
Papaver rhoeas R102	Lawrence and O'Donnell (1981)	36	30	0·61	(25)	34	**34**	**30–43**
Papaver rhoeas R104	Lawrence and O'Donnell (1981)	36	26	0·67	(22)	28	**28**	**26–35**
Papaver rhoeas R106	Campbell and Lawrence (1981*b*)	51	31	0·72	(23)	32	**32**	**31–37**

* Data from Williams and Williams (1947) interpreted, perhaps questionably, by Bateman.

† The ML estimates of *N* for the clover populations shown above are the correct values; Paxman's (1963) values for these populations (192 and 215 respectively) are incorrect.

as Fisher (1947) predicted it would, when used on data where the allele frequencies are known to be significantly unequal, as is the case with the three poppy samples. On the other hand, while neither Whitehouse's nor the maximum likelihood method give nonsensical estimates when used on our own data, both must be biased downwards (see later). Neither method, therefore, can be regarded as very satisfactory when used in these circumstances. Before, however, we turn to consider the problem of the estimation of N in populations where the genotype frequencies are not assumed equal, it is convenient to deal first with the question of confidence intervals for maximum likelihood estimates.

2. Confidence intervals for $\hat{N}$

In many circumstances, the variance of an estimator can be used to calculate a confidence interval for the parameter, by means of the familiar "±2 standard error" type of argument. There are two closely related reasons why this is a not very satisfactory procedure in the present case. First, the distribution of the estimate may not be symmetrical, which is particularly to be expected where m is much greater than N. Second, irrespective of whether the distribution of the estimate is symmetrical, N cannot be less than n. Hence it follows that the possible values of a reasonable estimate must be truncated at n, thus introducing a further potential source of asymmetry. In these circumstances, the attachment of a symmetrical standard error to the estimate is not a sensible procedure. What is clearly required in order to indicate the precision of $\hat{N}$ is a procedure for calculating a confidence interval which takes both sources of asymmetry into account. The procedure given below achieves this objective, for it is independent of whatever method of point estimation is used, so that any asymmetry in the distribution of the estimate will not affect it.

Consider an infinite population of self-compatible plants in which N S-alleles occur and where all $N(N-1)/2$ genotypes are equally frequent. We then ask; what is the probability, $P(x;\, m, N)$ that a sample of m plants from this population will contain x alleles? A solution to this problem can be obtained from the following argument. When the $(m-1)$th plant was sampled either x, $x-1$ or $x-2$ alleles had been found; similarly for the $(m+1)$th plant, either x or $x+1$ or $x+2$ alleles will have occurred. Given that x alleles have been found in m plants, the probability that no new alleles will be found in the $(m+1)$th plant is:

$$\frac{x(x-1)}{N(N-1)}$$

the probability that one new allele will be found is

$$\frac{2x(N-x)}{N(N-1)}$$

and the probability that two new alleles will be found is

$$\frac{(N-x)(N-x-1)}{N(n-1)}.$$

Using "transitional" probabilities like these, the probability $P(x; N, m)$ can be expressed in terms of $P(x; N, m-1)$, $P(x-1; N, m-1)$ and $P(x-2; N, m-1)$ in the following way:

$$\begin{aligned} P(x; N, m) = {} & \frac{x(x-1)}{N(N-1)} \,.\, P(x; N, m-1) \\ & + \frac{2(x-1)(N-x+1)}{N(N-1)} \,.\, P(x-1; N, m-1) \\ & + \frac{(N-x+2)(N-x+1)}{N(N-1)} \,.\, P(x-2; N, m-1). \end{aligned} \tag{5}$$

From the properties of the self-incompatibility system,

$$P(2; N, 1) = 1$$

and

$$P(x; N, 1) = 0 \quad \text{for } x \neq 2;$$

all other values of $P(x; N, m)$ can be calculated from (5) on the computer.

Expression (5) defines the probability distribution of the random variable x for fixed N and m. Suppose that in a sample of m plants, n distinct S-alleles have been found and suppose, further, that the true value of N is N_0. The sum Θ can then be defined as:

$$\Theta = \sum_{n} P(x; N_0, m)$$

where the x's fulfill the condition

$$P(x\colon N_0, m) \leqq P(n; N_0, m).$$

The set of all possible values of N_0 can be divided into two subsets; those for which Θ is less than some significance level α and those for which Θ is greater than α. The latter set defines a $(1-\alpha) \times 100$ per cent confidence interval for N. This procedure is equivalent to carrying out a significance test on each of the possible values of N_0.

The results obtained when this method is applied to the *Oe. organensis* and *P. rhoeas* data are shown in the last column of table 1 (see, also, table 5 of Lawrence and O'Donnell, 1981). There are two points worth making about these 99 per cent confidence intervals for the maximum likelihood estimates of N. First, the width of each interval is inversely correlated with the repeatability (R) of the experiment in question. This is clearly a desirable property of any limits that might be attached to an estimate, for the repeatability is a measure of the thoroughness of an investigation (Campbell and Lawrence, 1981*a*). Second, in a thorough investigation, most of the alleles present in the population will have occurred in the sample. In consequence, $\hat{N}$ should be close to n; and, because $\hat{N}$ cannot be less than n, the limit should be asymmetrically located about $\hat{N}$, as is, indeed, the case. The procedure we have used, therefore, appears to give sensible limits to $\hat{N}$.

Before we turn to consider the problem of estimating N where genotype frequencies are not assumed equal, it is worth pointing out that equation (5) can be used to derive the maximum likelihood estimate. Thus if n distinct

S-alleles have occurred in m plants, the value of N which maximises $P(n; N, m)$ gives the maximum likelihood estimate of this parameter. Consideration of the algebraic form of equation (5) for low values of m make it clear that the terms N, n and m in $P(n; N, m)$ follow a simple pattern. From this an explicit likelihood function can be obtained which is

$$1(N; n, m) = k.\frac{N!}{N^m(N-1)^m(N-n)!}$$

where k is not a function of N. The maximum of this function can be found by solving

$$N^{m-1}(N-n)-(N-2)^m=0 \tag{6}$$

which is a polynomial with only one real root greater than n; it can be shown that this equation yields the same estimates of N as equations (3) and (4).

3. Unequal genotype frequencies

Each of the estimators we have considered so far assume that the incompatability genotypes in the population are equally frequent. We know that this assumption is most unlikely to be true for *P. rhoeas*, because in each of the three samples we have examined, the allele frequencies were significantly unequal (Campbell and Lawrence, 1981*b*; Lawrence and O'Donnell, 1981). It is possible that this assumption does not hold for the Oenothera data either, for although there is no evidence that alleles are unequally frequent in Emerson's (1939) initial sample, the fact that he ultimately found 45 different S-alleles (Emerson, 1940), eight more than the number at the upper bound of the 99 per cent confidence interval (assuming equal frequencies), suggests that the allele frequencies may be unequal in this population also. It is obvious that in these circumstances the use of the maximum likelihood expression (3) will lead to an estimate of N which is biased downwards because the more frequent S-alleles are more likely to occur in a sample than the less frequent ones. There is considerable justification, therefore, for considering the problem of estimating N in populations where the genotype frequencies are unequal.

The advantage of the equal frequency assumption is that it imposes a constraint which makes inference relatively simple, for in these circumstances a particular sample determines uniquely the population from which it has been drawn; that is, the number of alleles found in the sample, n is a sufficient statistic. If equal S-allele frequencies are not assumed this constraint is removed. Though the sample contains other information, namely, the frequencies of the allele present, it is unlikely that any statistic will be found that can discriminate between all possible populations that might have given the observed sample. Strictly speaking, then, inference in these circumstances is impossible.

Despite this rather unpromising conclusion, it is nevertheless possible to make some progress with this problem. Thus, while in theory an allele can have an infinitesimally low frequency, in practice its frequency cannot be less than $1/2N_p$, where N_p is the number of plants in the population. Furthermore, an allele with a frequency as low as this would be in danger of loss by drift. Hence the only way a large number of alleles could be maintained at these very low frequencies would be if the mutation rate was very high; as is well known, the evidence on this matter suggests that the mutation rate at the *S*-locus is very low (Lewis 1948, 1951). It follows, therefore, that there is little reason to believe that populations of *P. rhoeas* contain large numbers of *S*-alleles at very low frequencies. This is an encouraging conclusion because it is these alleles, if present, which would cause the most serious difficulty for any estimator that might be devised.

Now while equation (3) is not true, if genotype frequencies are unequal, the relation

$$\mathrm{E}n \leqq \{1-(1-2/N)^m\}$$

is true whatever the genotype frequencies may be. Furthermore, if the expectation of the maximum likelihood estimator (assuming equal genotype frequencies) is denoted as ML, then any estimator Θ which fulfills the condition

$$\mathrm{EML} \leqq \mathrm{E}\Theta \leqq N \tag{7}$$

can justifiably be regarded as an improvement on the maximum likelihood estimate, since it is less biased, provided, of course, that it does not have the disadvantage of a substantially greater variance.

One way of improving the maximum likelihood estimator would be to find an expression for E*n* which was true whatever the allele frequencies might be. An expression with this property can be obtained in the following way. The factor $\{1-(1-2/N)^m\}$ in equation (3) is the probability of an allele of frequency $1/N$ appearing in a sample of *m* plants. If the allele has a frequency of x_i, the corresponding expression is $\{1-(1-2x_i)^m\}$ and a quality which can be called "*the average probability of an allele appearing in a sample of size m*" can be defined as:

$$\frac{1}{N}\cdot\sum_{i=1}^{N}\{1-(1-2x_i)^m\}. \tag{8}$$

In an analogous fashion to Paxman's equation (3), the equivalent formula for the case of unequal genotype frequencies might be expected to be

$$\mathrm{E}n=\sum_{i=1}^{N}\{1-(1-2x_i)^m\}. \tag{9}$$

We are, again, indebted to Dr Paul Davies for the following simple proof that (9) is indeed the exact value of *n*.

Let $J_i=1$ if allele *i* is found in the sample; otherwise, $J_i=0$. It follows that:

$$n=\sum_{i=1}^{N}J_i$$

and

$$\mathrm{E}n=\sum_{i=1}^{N}\mathrm{E}J_i.$$

Since EJ_i = the probability that allele i is found in the sample

$$EJ_i = 1-(1-2x_i)^m$$

so that

$$En = \sum_{i=1}^{N} \{1-(1-2x_i)^m\}.$$

An estimator of (8), "the average probability of an allele appearing in a sample of size m", if such an estimator can be found, will also estimate N because

$$En = N \times \frac{1}{N} \sum_{i=1}^{N} \{1-(1-2x_i)^m\}$$

and En can be equated to the value of n from the sample.

Though any number of estimates of "the average probability of an allele appearing" can be devised, two only will be considered here. The most obvious estimator of (8) is:

$$\frac{1}{n} \sum_{i=1}^{n} \{1-(1-f_i/m)^m\}$$

where f_i is the number of times the ith allele occurred in the sample. Clearly, to some extent the quantities f_i/m for the observed alleles only are overestimates of the true value of $2x_i$. Then N can be estimated from

$$\hat{N} = \frac{n^2}{\sum_{i=l}^{n} \{1-(1-f_i/m)^m\}}. \tag{10}$$

This estimator of N will be referred to as E_1.

A slightly more complicated estimator can be derived by the following argument. A sample of m plants drawn from a population containing N alleles will contain n of these alleles and not contain $(N-n)$. Let the sum of the frequencies of the latter be $x/(1+x)$ and let $w=(1+x)$. The population is assumed to be of infinite size so that sampling does not alter the genotype frequencies in the population. If the ith allele has occurred in the sample f_i times, an estimate of its frequency in the population is $f_i/2mw$. Then an estimate of the average gene frequency in the population is:

$$\frac{1}{n} \sum_{i=1}^{n} f_i/2mw = 1/nw.$$

Given w, this estimate should be biased upwards, on average, because the sample will contain more of the high frequency alleles than the low frequency ones. So a downwards biased estimate of the number of alleles in the population, N, is given by nw. n can then be estimated from

$$n = w \sum_{i=1}^{n} \{1-(1-f_i/mw)^m\} \tag{11}$$

by equating En to n in equation (9) with substitution of our estimates of x_i. This expression can be solved for w and N estimated as nw. This estimator will be referred to as E_2.

4. Simulations

Since it is not obvious how these estimators will perform in practice, eight populations, which differed in the number and frequency of alleles they contained, were simulated on the computer. Samples were then drawn at random from these populations and E_1, E_2 and the maximum likelihood estimate calculated for each. The means of these statistics over all samples drawn from the same population gave an empirical estimate of their expected

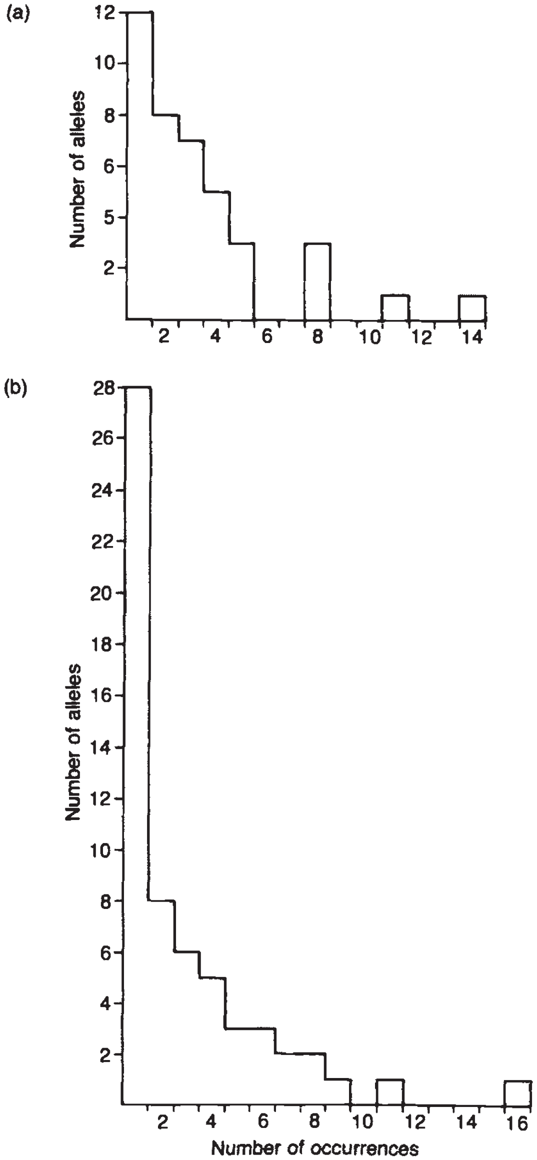

FIG. 2. Frequency distributions for runs 9(a) and 14(b) of the simulation investigation (see table 2).

value. The inclusion of the maximum likelihood estimator in these tests enables us to find out how bad this is when allele frequencies are unequal.

Between 1000 and 3000 samples of size 30 or 40 were drawn from each population; populations were of size 4000 and contained either 10, 40 or 60 different *S*-alleles. In three of these populations, the alleles were, approximately, equally frequent ($f_i \simeq 1/N$); in a further three, these frequencies were determined by the convenient method of equating f_i to $2i/N(N+1)$; and in the remaining pair of populations, the allele frequencies were distributed as in figure 2 in an attempt to simulate the distributions found in practice with the poppy data (Campbell and Lawrence, 1981*b*; Lawrence and O'Donnell, 1981). Further details of these simulations are given in O'Donnell (1983).

There are five points worth making about the results obtained from these simulations (table 2). First, the maximum likelihood (ML) and E_2 estimators appear to be virtually insensitive to the size of sample drawn from the population; for E_1, on the other hand, a marked improvement is obtained by increasing sample size from 30 to 40 when the population contains a realistic number of alleles (runs (10) and (11); and runs (12) and (13)).

TABLE 2

A comparison of the behaviour of the maximum likelihood (ML), E_1 *and* E_2 *estimators by repeated sampling from populations of size 4000 containing either 10, 40 or 60 different* S-*alleles*

Run	N	f_i	m	No. of samples	ML	E_1	E_2
1	10	1/10	30	2000	10·000	10·132	10·164
2			40	1000	10·000	10·033	10·066
3		$i/55$	30	1000	9·586	10·054	10·133
4			40	1000	9·765	10·092	10·145
5	40	1/40	30	2000	40·256	39·522	46·793
6			40	1000	40·146	41·819	45·908
7		$i/820$	30	3000	32·896	34·175	38·765
8			40	1000	33·521	36·154	38·892
9		fig. 2(a)	30	1000	30·244	32·641	37·385
10	60	1/60	30	2000	60·636	51·149	70·874
11			40	1000	60·182	57·180	70·402
12		$i/1830$	30	3000	48·049	44·857	57·030
13			40	2000	48·785	49·543	57·768
14		fig. 2(b)	30	1000	40·126	40·443	49·742

Second, when only 10 alleles are present in the population, there is little to choose between the estimators, irrespective of whether the alleles are equally frequent or not. This is, of course, as expected because when the number of alleles in the population is, relative to the number sampled ($2m$), as low as this, there is little need for statistical inference; however, the results obtained for this extreme situation at least show that the estimators behave as expected. Third, while, as expected the ML estimator gives satisfactory estimates when the frequencies of the alleles in the population are equal, these estimates are biased downwards when allele frequencies are unequal; furthermore, we note that this bias is greatest for the two runs ((9) and (14)) where the frequency distributions of the alleles are similar to those found in practice. Fourth, though the E_1 estimates appear to be less biased than the ML estimates for populations containing either 10 or 40 alleles, the former appear to be as poor, if not worse, than the latter when $N = 60$. It is clear, therefore, that E_1 cannot be regarded as a satisfactory estimator over the range of conditions that have been simulated in these tests. Lastly, though E_2 always gives estimates which are less biased than ML estimates when allele frequencies are unequal, it yields estimates which are biased upwards when these frequencies are equal.

Taking these results as a whole, therefore, it is clear that neither E_1 nor E_2 completely fulfil criterion (7); hence neither can be regarded as a satisfactory replacement for the ML estimator. On the other hand, our simulations suggest that E_2 fulfils this criterion when allele frequencies are unequal. Hence there is some justification for using E_2, rather than the ML estimator, when allele frequencies are known to be unequal; in all other circumstances, however, in the present state of knowledge, it is better to use the latter. Further investigation of estimators of the E_2 type are necessary both to improve their bias properties and to investigate their variance.

5. The number of alleles in populations of *P. rhoeas*

Estimates, $\hat{N}$, of the number of S-alleles in populations of *P. rhoeas* given by the E_2 estimator are shown in table 3. Insofar that these estimates are larger than those yielded by the ML method (table 1), they must be regarded as more satisfactory. On the other hand, as is clear from the simulations (table 2), estimates of N given by the E_2 estimator are also biased downwards, though to a lesser extent, than the ML estimates. It follows, therefore, that the estimates shown in table 3 should be regarded as minimum estimates of the number of alleles in these populations. The most reasonable conclusion that can be drawn from these results, therefore, is that there are at least 40 different S-alleles in each of these populations, though probably not very many more than this number.

TABLE 3

Estimates of the number of S-*alleles in* P. rhoeas *populations obtained from the* E_2 *estimator* (*see text*)

Population	$\hat{N}$
R102	42
R104	34
R106	38

Acknowledgements. We are indebted to Dr Paul Davies of the Department of Statistics for his advice both during the course of the work and the preparation of this paper; and also to Dr J. S. Gale for much useful discussion. The simulations were carried out on the University of Manchester's CDC7600 computer to whom our thanks are due. One of us (S.O.) wishes to acknowledge receipt of an SERC Studentship which enabled the work described in this paper to be carried out.

6. References

BATEMAN, A. J. 1947. Number of *S*-alleles in a population. *Nature*, *160*, 337.

CAMPBELL, J. M. AND LAWRENCE, M. J. 1981*a*. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. I. The number and distribution of *S*-alleles in families from three localities. *Heredity*, *46*, 69–79.

CAMPBELL, J. M. AND LAWRENCE, M. J. 1981*b*. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. II. The number and frequency of *S*-alleles in a natural population (R106). *Heredity*, *46*, 81–90.

CHAPMAN, D. G. 1952. Inverse, multiple and sequential sample census. *Biometrics*, *8*, 286-306.

DARROCH, J. N. 1958. The multiple-recapture census. I. Estimation of a closed population. *Biometrika*, *45*, 343-359.

EMERSON, S. 1939. A preliminary survey of the *Oenothera organensis* population. *Genetics*, *24*, 524-537.

EMERSON, S. 1940. Growth of incompatible pollen tubes in *Oenothera organensis*. *Bot. Gaz.*, *101*, 890-911.

FISHER, R. A. 1947. Number of self-sterility alleles. *Nature*, *160*, 797.

LAWRENCE, M. J. AND O'DONNELL, S. 1981. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. III. The number and frequency of *S*-alleles in two further natural populations (R102 and R104). *Heredity*, *47*, 53-61.

LEWIS, D. 1948. Structure of the incompatibility gene. I. Spontaneous mutation rate. *Heredity*, *2*, 219-236.

LEWIS, D. 1951. Structure of the incompatibility gene. III. Types of spontaneous and induced mutation. *Heredity*, *5*, 399-414.

O'DONNELL, S. 1983. Population genetics of self-incompatibility in *Papaver rhoeas* L. Ph.D. thesis, University of Birmingham.

PAXMAN, G. J. 1963. The maximum likelihood estimation of the number of self-sterility alleles in a population. *Genetics*, *48*, 1029-1032.

SCHNABEL, Z. E. 1938. The estimation of the total fish population of a lake. *Amer. Math. Mon.*, *45*, 348-352.

SEBER, G. A. F. 1982. *The estimation of Animal Abundance and Related Parameters*. Charles Griffin, London (2nd ed.).

WHITEHOUSE, H. L. K. 1949. Multiple allelomorph heterothallism in the Fungi. *New Phytol.*, *48*, 212-244.

WILLIAMS, W. AND WILLIAMS, R. D. 1947. Genetics of red clover (*Trifolium pratense* L.) compatibility. III. The frequency of incompatibility *S* alleles in two non-pedigree populations of red clover. *J. Genet.*, *48*, 69-79.