THE MAXIMUM LIKELIHOOD ESTIMATION OF THE NUMBER OF SELF-STERILITY ALLELES IN A POPULATION

G. J. PAXMAN

John Innes Institute, Hertford, Herts. England

Received March 11, 1963

PREVIOUS estimates of the number of alleles, n, at a self-sterility locus have been based on a method of DOBZHANSKY and WRIGHT (1941). These authors showed that the probability of two things being of the same kind, when taken at random from a population composed of n different kinds, is $(1/n) + n\sigma_p^2$, where p is the proportion of any one kind. This was adapted to the case of the incompatibility locus by BATEMAN (1947), taking into account the working of the system, whereby the two alleles of every diploid individual must be different. In practice, estimates have been made from data which gave no evidence that the alleles differed in frequency. This is the situation expected in a population at equilibrium provided that there are no selective effects correlated with particular alleles. The value of p may then be assumed equal to $1/n$ for all alleles, making the variance of p (σ_p^2) zero, and the last term of WRIGHT's formula disappears. The estimate then depends on the proportion of pairs of individuals in the sample that carry identical alleles.

In common with BATEMAN (1947) and RAPER, KRONGELB and BAXTER (1958), it is assumed that the alleles going to make up an individual are associated at random, i.e. that the probability of any pair is $\frac{1}{n(n-1)}$. Then the distribution of single alleles, as considered here, is as informative about the total number of alleles as the distribution of pairs. Thus the data from a fully analyzed sample may be reduced to the form a_1 alleles occurred once; a_2 alleles occurred twice; a_r alleles occurred r times; where $\sum_r a_r = a_t$, the total number of different alleles in the sample, and $\sum_r r\, a_r = 2m$, the number of alleles tested, two in each of the m individuals. As the above method does not make full use of the frequency data, it seems possible that it might not be fully efficient in the special case of all alleles being assumed equally frequent.

Consider the probability, P(1), of any given allele occurring once only in a sample of m individuals. It can occur as either of the two alleles in an individual which can be chosen in ${}^mC_1$ ways. The probability of the other allele being different is unity, by the nature of the incompatibility system. Similarly, in each of the remaining $(m-1)$ individuals, the probability that one allele differs from

the given allele is $(1-\frac{1}{n})$ and for the second allele $(1-\frac{1}{n-1})$. Thus:

$$P(1) = {}^{m}C_1\, 2\, (\frac{1}{n})\ (1-\frac{1}{n})^{m-1}\,(1-\frac{1}{n-1})^{m-1}$$

similarly $$P(2) = {}^{m}C_2\, 2^2(\frac{1}{n})^2\,(1-\frac{1}{n})^{m-2}\,(1-\frac{1}{n-1})^{m-2}$$

and $$P(r) = {}^{m}C_r\, 2^r(\frac{1}{n})^r\,(1-\frac{1}{n})^{m-r}\,(1-\frac{1}{n-1})^{m-r}.$$

Collecting together terms of the same power,

$$P(r) = {}^{m}C_r\,(\frac{2}{n})^r\,(1-\frac{2}{n})^{m-r}$$

which is the general term of the binomial $[\frac{2}{n}+(1-\frac{2}{n})]^m$.

The observed frequency distribution of alleles in the sample is a truncated binomial since the number of alleles in the population not occurring in the sample (corresponding to the zero term of the series, $P(0)=(1-\frac{2}{n})^m$) cannot be directly observed. The size of sample, m, is fixed by the experimenter, but a_t may vary from sample to sample between the limits 2 and $2m$. It has been shown that the maximum likelihood estimation of p in the general case $(p+q)^m$ is obtained by equating the mean of the sample to its expectation (HALDANE 1932; FISHER 1934):

$$E\ (\text{mean}) = \frac{mp}{(1-q^m)} = \frac{\sum_{r=1}^{m} r\, a_r}{\sum_{r=1}^{m} a_r}$$

which becomes $$\frac{2m}{n[1-(1-\frac{2}{n})^m]} = \frac{2m}{a_t}$$

that is $$n[1-(1-\frac{2}{n})^m] = a_t.$$

The asymptotic variance of $\hat{p} = 2/\hat{n}$ is given by: $V_{\hat{p}} = \dfrac{pq(1-q^m)^2}{a_t m(1-q^m-mpq^{m-1})}$.

FINNEY (1949) and LEJEAUNE (1958), have tabulated functions allowing rapid solution of the above equation of estimation. However, both were concerned with the application of the method to the segregation of recessive traits in human families (HALDANE 1932). Neither author considers samples greater than 20 nor values of $p = 2/n$ below 0.01. Their tables are therefore unlikely to be of use with incompatibility data. However, it is no very laborious task to solve the equation for n by iteration. Fiducial limits can be calculated for the estimate of $2/n$ from its variance, and hence for the estimate of n.

This estimate may be regarded as the likely minimum number of alleles in the population, since any inadequacy of the assumption of equal frequency of alleles results in an underestimate of n. This led FISHER (1947) to suggest that

TABLE 1

Comparison of methods

Sample		Size (m)	Published estimate $\hat{n}$	5% fiducial limits		Maximum likelihood estimate $\hat{n}$	5% fiducial limits	
BATEMAN 1947	1st population	25	171	83	429	192	149	268
	2nd population	22	308	105	1490	215	110	3690
RAPER *et al.* 1958	*A* factor	57	336	217	562	295	208	504
	B factor	57	64	53	79	69	56	90

upper fiducial limits should not be ascribed to $\hat{n}$. He pointed out, using data from *Trifolium pratense* (WILLIAMS and WILLIAMS 1947), that a moderate degree of variation in p between alleles could account for the observed repetitions in the sample, even if the number of alleles, n, were infinite. Lower fiducial limits are not open to this criticism.

The maximum likelihood estimate of n may be compared with BATEMAN's method in two sets of published data. BATEMAN (1947) used the data of WILLIAMS and WILLIAMS (1947) from two populations of *Trifolium pratense*, and RAPER *et al.* (1960) estimated n for the *A* and *B* mating type factors in *Schizophyllum commune*. Although little weight can be attached to upper fiducial limits, these have been calculated for comparison with the published figures and are shown in Table 1. The two methods give somewhat different values of $\hat{n}$ in any one sample, but neither gives a biased estimate. The upper fiducial interval may be larger in the maximum likelihood case, but the important lower interval is always smaller for a given value of $\hat{n}$. Where the sample size is large, i.e. of the same order as the number of alleles in the population (*cf.* line 4 of Table 1) both methods are efficient. In such a case, the lower fiducial limit may be fixed in practice by the number of different alleles observed in the sample, which clearly cannot exceed the actual number in the population. It may be noted that RAPER *et al.* observed 56 different *B* factors in their sample, yet give a lower five percent fiducial limit of $\hat{n}$ as 53. The advantage of the maximum likelihood method increases with higher ratios of $n:m$. Thus, in the second Trifolium population, in which the total number of alleles was about ten times the number of plants investigated, the five percent fiducial interval of the maximum likelihood estimate was only one half of the corresponding published figure.

SUMMARY

A maximum likelihood expression is derived for estimating the number of self-sterility alleles in a population, when the frequency of all alleles can be assumed equal. This method has the advantage of giving a smaller lower fiducial interval than the method previously used, particularly when the number of alleles is large compared with the size of the sample.

ACKNOWLEDGMENT

I am grateful to MR. N. GILBERT for much helpful discussion.

LITERATURE CITED

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

DOBZHANSKY, TH., and S. WRIGHT, 1941 Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of *Drosophila pseudoobscura*. Genetics **26:** 23–51.

FINNEY, D. J., 1949 The truncated binomial distribution. Ann. Eug. **14:** 319–328.

FISHER, R. A., 1934 The effect of methods of ascertainment upon the estimation of frequencies. Ann. Eug. **6:** 13–25.

FISHER, R. A., 1947 Number of self sterility alleles. Nature **160:** 797–798.

HALDANE, J. B. S., 1932 A method of investigating recessive characters in man. J. Genet. **25:** 251–255.

LEJEAUNE, J., 1958 Sur un solution *a priori* de la methode *a posteriori* de Haldane. Biometrics **14:** 153–520.

RAPER, J. R., G. S. KRONGELB, and M. G. BAXTER, 1958 The number and distribution of incompatibility factors in Schizophyllum. Am. Naturalist **92:** 221–232.

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