Estimation of the Allele Number in a Natural Population by the Method of Isofemale Lines

P. Capy and J. Rouault

Laboratoire de Biologie et Génétique Evolutive, Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette, France

Manuscript received October 17, 1986

Revised copy accepted August 24, 1987

ABSTRACT

The number of alleles present in a natural population of unknown structure is estimated using a sequential sampling procedure applied to isofemale lines. Two questions are raised: how many individuals per isofemale line must be assayed and how many isofemale lines must be sampled to get an adequate sample to estimate the number of alleles, at a given risk, of the natural population? On the one hand, we show that when wild females are inseminated once, only two individuals per line are required. On the other hand, the number of isofemale lines that must be sampled depends on the risk chosen of losing an allele, on the number of alleles present in the population and on their drawing probabilities. When the population structure is known, an accurate answer can be provided. For an unknown population structure, one general sequential sampling previously described by J. Rouault and P. Capy is proposed to estimate the number of alleles in the population from data on isofemale lines.

POLYMORPHISM of natural populations of *Drosophila* species cannot always be analyzed from wild individuals. For example, analysis of an inversion polymorphism in *Drosophila* requires third instar larvae. Often, in such cases, the isofemale lines protocol has been used. The usefulness of the isofemale lines method has been demonstrated in addressing several problems (PARSONS 1980). For analyzing an allozyme polymorphism, this technique avoids the bottleneck effect brought about when few individuals are used to found a laboratory strain. We are interested in the use of such data to estimate allelic diversity within a natural population, as measured by the number of alleles. To address this problem, there are two questions to be answered: how many isofemale lines are required and how many individuals must be assayed per line in order that the overall risk, fixed *a priori*, of missing any allele is small.

The relationship between sample size and the number or the frequencies of variants detected have been described by FISCHER, CORBET and WILLIAMS (1943), COTTERMAN (1954), PEET (1974), and the decrease in the amount of information proportional to the sample size by CANNING and EDWARDS (1968), and GREGORIUS (1980). Otherwise, FINNEY (1948), EWENS (1972), WATTERSON (1974), FUERST, CHAKRABORTY and NEI (1977) and more recently ARNOLD (1981), ARNOLD and MORRISON (1985), CRESSIE and SEHEULT (1985) and SOBEL, ARNOLD and SOBEL (1986) have proposed different methods to estimate a variant probability in a population of unknown structure. Some of these methods are based on sequential sampling plans. Such schemes have already been used or advocated by HALDANE (1945), EDWARDS (1960),

CHOW and YU (1981), CANNING and THOMPSON (1977) and SOBEL, ARNOLD and SOBEL (1986) for different purposes.

However, these methods cannot be used to answer the above questions. Indeed, some of them require the knowledge of the population structure and/or the prior distribution of the alleles number and the others cannot be applied to the isofemale lines protocol. The "structure" of a population will be defined as the number of alleles at a given loci and their respective frequencies.

Our purpose is to describe a sequential sampling method to estimate the number of alleles in a natural population of unknown structure from isofemale line data using the sampling procedure defined by ROUAULT and CAPY (1986).

SAMPLING IN A POPULATION OF KNOWN STRUCTURE

All symbols used in this paper, with their respective meanings, are given in Table 1. Let us consider a natural population containing na alleles with the drawing probabilities (or allele probabilities) $p_1, p_2, \dots, p_{na}$. This population can be thought of as a gametic urn. The probability of drawing each allele a_i times ($0 \leq a_i \leq nt$, where nt is the total number of draws) follows a multinomial distribution. The sum $Pt(na)^1$ of all probabilities related to the events where each $a_i \neq 0$ corresponds to the probability of drawing each of the na alleles, present in the urn at least once. The sampling error (which can be compared to a type I

¹ The computer program used to calculate $Pt(na)$ (written in PASCAL UCSD on an APPLE II micro-computer) will be available upon request to P. CAPY.

TABLE 1
Meaning of the notation used

a_i	= number of draws of the i th allele
nt	= total number of draws
na	= number of alleles in the base population
na'	= number of alleles detected (in a sample)
p_i	= drawing probability of the i th allele
f_i	= observed frequency of the i th allele
$Pt(na)$	= <i>a priori</i> probability (before sampling) to draw, at least, one time each of the na alleles
rt	= first kind risk associated with $Pt(na)$
$Pe(na)$	= <i>a posteriori</i> probability (after sampling) to draw, at least, one time each of the na alleles
re	= sampling error associated with $Pe(na)$
f_x	= proportion of females inseminated by x males

error) associated to this probability will be,

$$rt = 1 - Pt(na).$$

Some thought must be given to compute $Pt(na)$ when nt and/or na becomes large, so an equivalent expression, depending only on nt , na and p_i , is given by,

$$Pt(na) = \sum_{z=0}^{z=na-1} (-1)^z S_{na,z} \quad (1)$$

where $S_{na,z}$ is the sum of all combinations of sums of $na-z$ elementary probabilities p_i among na , raised to the nt th power. Therefore,

$$S_{na,z} = \sum_{k=1}^{k=C_{na-z}^{na}} T_{na,z,k}^{nt} \quad (2)$$

where $T_{na,z,k}$ is the k th sum of combinations of $na-z$ elementary probabilities p_i among na [see ROUAULT and CAPY (1986) for more details]. When necessary to stress what the probability $Pt(na)$ is a function of, we will write,

$$Pt(na) = E(nt, na, p_1, \dots, p_{na}). \quad (3)$$

For example, with two alleles and the drawing probabilities p_1 and p_2 , the previous expression becomes,

$$Pt(2) = 1 - p_1^{nt} - p_2^{nt}. \quad (4)$$

and for three alleles with the drawing probabilities p_1 , p_2 and p_3 ,

$$Pt(3) = 1 - (p_1 + p_2)^{nt} - (p_1 + p_3)^{nt} - (p_2 + p_3)^{nt} + p_1^{nt} + p_2^{nt} + p_3^{nt}. \quad (5)$$

SAMPLING IN A POPULATION OF UNKNOWN STRUCTURE

Sampling procedure when the population structure is partly unknown: First, let's suppose that the biologist obtains a preliminary estimate of the structure of the natural population. He selects the risk rt , which he is willing to tolerate. Then, the real structure of the natural population will be estimated using the

preliminary estimate and the following sampling procedure:

1. The sample size nt is selected from the expression E using the number of alleles na , their respective drawing probabilities $p_1, \dots, p_{na}$ and the *a priori* probability $Pt(na) = 1 - rt$. This procedure can be summarized as,

$$\left| \begin{array}{l} na \\ p_1, \dots, p_{na} \end{array} \right. \rightarrow E(nt, na, p_1, \dots, p_{na}) \rightarrow nt. \\ Pt(na)$$

2. The sample of nt alleles taken at random in the gametic urn provides a first estimate. Suppose that $na' > na$ alleles have been detected in frequencies $f_1, f_2, \dots, f_{na'}$.
3. The *a posteriori* probability $Pe(na')$, of this sample, is given by

$$\left| \begin{array}{l} nt \\ na' \end{array} \right. \rightarrow E(nt, na', f_1, \dots, f_{na'}) \rightarrow Pe(na) \\ f_1, \dots, f_{na'}$$

and the sampling error by $re = 1 - Pe(na)$.

4. The comparison between re and rt determines whether or not the sampling protocol must be continued or stopped. The sampling protocol ceases when $re \leq rt$.

If $re > rt$, the sample should not be considered a satisfactory sample of the gametic urn. Therefore, assuming that the structure observed after the first sampling corresponds to the real structure of the gametic urn, i.e.,

$$na = na'$$

and

$$p_1, \dots, p_{na} = f_1, \dots, f_{na},$$

a new sample size nt' ($nt' > nt$) can be defined (step 2 of the previous protocol). The allelic frequencies observed *a posteriori* become the *a priori* drawing probabilities used to estimate the risk rt . Insofar as each sample can be considered as independent from the previous ones (hypothesis of infinite size of the gametic urn), the quantity

$$dt = nt' - nt \quad (6)$$

represents the size of the complementary sample. Indeed, instead of drawing nt' gametes, it is equivalent to complete the first sample with a drawing of dt gametes. Data from the first sample together with data from the complementary sample are pooled in one sample from which new estimations can be inferred. A new comparison between re and rt will be done, and so on until re becomes less than or equal to rt .

The convergence of this algorithm proceeds from the law of the large numbers (ROUAULT and CAPY 1986), i.e., in this protocol from the progressive in-

crease of the size of the final sample (sample containing the initial sample and all complementary samples). This algorithm allows one to get an accurate sample at the risk rt . However, if the drawing probability of an allele is less than rt or $1/na$, it is likely that it will not be in the final sample which will then be biased.

Sampling procedure when the population structure is unknown: When no data about the number of alleles and their respective frequencies are available, it is necessary to make some assumptions to initialize the sampling algorithm. One way is to determine an arbitrary structure of the gametic urn. For example, it could be considered that the urn contains na alleles each of them having a drawing probability $1/na$. In this case any arbitrary structure of the population could be used. However, the uniform distribution (na alleles each of them having the drawing probability $1/na$ allow to minimize the number nt . Nevertheless, in almost all studies from natural population the allele frequency spectrum is not linear (LEWONTIN 1974). So, the arbitrary allele frequencies can be chosen according to this observation. Finally, another solution is to draw a sample of size nt , without making a hypothesis about the population structure. But, the size of the first sample should not be too small. For example, if the population structure is $na = 2$ with $p_1 = 0.9$ and $p_2 = 0.1$, it is more likely that a single allele will be drawn for a small value of nt . In such cases the convergence of the algorithm is direct.

Example: Figure 1 gives an example of the sampling procedure. In this case, the first composition of the population, arbitrarily chosen, is $na = 2$ with $p_1 = 0.5$ and $p_2 = 0.5$. After three successive drawings the total number of alleles sampled is $nt = 21$ and four distinct alleles in frequencies 0.57, 0.19, 0.14 and 0.10, are detected (Figure 2). The true population structure was $na = 4$ alleles with the drawing probabilities: 0.5, 0.3, 0.1 and 0.1. Thus, all alleles present in the simulated population have been detected.

In this example, the formula 1 of the multinomial have been used in two different ways. First, this expression allows us to determine the number of drawings required to get a representative sample with the risk rt for a given population structure (arbitrarily defined or accurately known). Second, this expression allows us to compute the sampling error or the probability of each sample knowing the sample size and assuming that the population structure estimated is right.

APPLICATION TO ISOFEMALE LINES

The analysis of natural populations by the isofemale lines method involves two steps: (1) the sampling of females in the base population and (2) the sampling of offspring in each line. The expression E allows the calculation of the number of draws nt_1 that must be made for a given structure of the base population and

FIGURE 1.—Demonstration run of the sampling procedure. The sampling probability (P_t) is in ordinate and the number of drawing (nt) in abscissa. The structure of the theoretical population is $na = 4$ and the respective drawing probabilities are: 0.5, 0.3, 0.1 and 0.1. For the first drawing, the structure of the population, arbitrarily chosen, is $na = 2$ in proportion 1/1. In these conditions, it is required to draw $nt = 6$ alleles. After the first drawing 3 different alleles are detected in proportion 4/1/1 but the probability of the sample is 0.417. If this first estimation correspond to the true composition of the population, it is necessary to draw a total of 21 alleles to get a probability of the sample superior to 0.95. After the second drawing, 4 distinct alleles are detected in proportion 17/7/5/3 and the probability of the sample is 0.831. According to this new estimation, it is required to draw 32 alleles to get a probability $P_t = 0.951$. The drawing of 11 new alleles allows to get a new composition of the population: $na = 4$ and the respective frequencies are 0.57, 0.19, 0.14 and 0.10. Now, the probability of the sample is 0.953. re being superior to rt , the sampling procedure is stopped. After three successive drawings the sampling error is inferior to the risk $rt = 0.05$.

a given type I error (rt_1). Assuming that a wild female is inseminated by a single male, drawing from an isofemale line is equivalent to drawing four alleles. Then, the number nl of isofemale lines that must be sampled is,

$$nl = nt_1/4. \quad (7)$$

In the same way, the number ni of individuals that must be drawn in the total progeny of the isofemale lines, for a given allelic composition among the progeny and for a given type I error (rt_2) is,

$$ni = nt_2/2. \quad (8)$$

If it is assumed that the number of alleles and their probabilities are identical in the base population and in the descendants of the isofemale lines, then $nt_1 = nt_2$. Therefore, the total number of individuals sampled in the progeny of overall isofemale lines is,

$$\frac{ni}{nl} \cdot nl = \frac{nt_2}{2} \cdot \frac{4}{nt_1} \cdot nl = 2 \cdot nl. \quad (9)$$

The apportionment of these total numbers of individuals leads to draw an average of two individuals per isofemale line.

Thus, when the two type I errors rt_1 and rt_2 are identical, it is required to draw two times more indi-

FIGURE 2.—Number of alleles detected according to the number of individuals assayed per line from a sample of 10 (a), 20 (b), 30 (c) or 40 (d) isofemale lines. Four base populations are simulated: 1) $na = 2$; $p_1 = p_2 = 0.5$; 2) $na = 2$; $p_1 = 1 - p_2 = 0.9$; 3) $na = 5$; $p_1 = p_2 = 0.4$; $p_3 = 0.1$; $p_4 = p_5 = 0.05$; 4) $na = 5$; $p_1 = 0.890$; $p_2 = 0.049$; $p_3 = 0.030$; $p_4 = 0.016$; $p_5 = 0.015$.

TABLE 2

Examples of relationship between the first parameters of the sampling algorithm when the isofemale lines method is used and $na = 3$, $p_1 = 0.6$, $p_2 = 0.3$ and $p_3 = 0.1$

First kind error rt	Associated probability $Pt(na)$	Intermediate probability $Pt_i(na)$	Associated error rt_i	No. of draw nt	No. of lines nl
0.25	0.750	0.8660	0.1340	20	5
0.05	0.950	0.9747	0.0253	36	9
0.01	0.990	0.9950	0.0050	52	13
0.001	0.999	0.9995	0.0005	76	19

viduals, among the progeny of isofemale lines, than isofemale lines.

Assuming that the two steps involved in this protocol are independent, and choosing two identical type I errors ($rt_1 = rt_2$, involving $Pt_1(na) = Pt_2(na)$) for each step, the total type I error rt , corresponding to an *a priori* probability $Pt(na)$, is,

$$rt = 1 - \sqrt{Pt_1(na) \cdot Pt_2(na)}. \quad (10)$$

Table 2 gives some numerical examples of the relationships between the different parameters for a given structure.

SIMULATIONS OF SAMPLING IN NATURAL POPULATION

In order to test the algorithm, some simulations have been made using the RND (random) function of

an Apple IIe computer with the DOS 3.3 operating system.

This algorithm can be initialized in many different ways (Figure 3). The path C of the algorithm where nt is defined arbitrarily, has been chosen for all the simulations.

Table 3 gives the number of alleles detected and their frequencies, in four independent repetitions of the same process. The base population contains 5 alleles with the drawing probabilities 0.4, 0.4, 0.1, 0.05 and 0.05. The number nt of isofemale lines of the first sample is always 10.

Except for the second repetition, all alleles have been detected. However, some alleles present in the base population can not be detected (e.g., repetition 2, Table 3). This disadvantage depends on the value of the type I error and/or on the drawing probabilities of alleles. In this case, the type I error was 0.05 and the drawing probability of the allele not detected 0.05. For this repetition the number of lines sampled was particularly small (13 isofemale lines). Nevertheless, it must be stressed that the allelic frequencies observed are very close to the drawing probabilities when all alleles are detected.

The number of alleles detected is always inferior or equal to the exact value. Moreover, this number is generally weak and rarely exceeds ten alleles. One possible method that can be applied in this case, to calculate the confident interval of na , is to use some methods like Bootstrap (EFRON 1984). Nevertheless, the risk rt can be considered as a measure of the uncertainty in the estimation of na .

These four examples demonstrate that the number of isofemale lines drawn at the end of the algorithm can be very different from one repetition to another (37, 13, 68, 19 in the four examples given). In Table 4, some estimations of the number of alleles have been achieved for three situations. In the first two situations, only two alleles exist in the base population in proportions 0.5/0.5 or 0.9/0.1 whereas in the third situation five alleles were simulated in proportions 0.4, 0.4, 0.1, 0.05 and 0.05. The values presented in this table are the means of 100 repetitions. Each simulation consisted of some initial samples of 5, 10 or 20 isofemale lines. If two alleles exist in the proportion 0.5/0.5 the number of lines drawn at the end of the algorithm remains small. Except when the size of the first sample is equal to 5, all alleles are detected when 10 and 20 isofemale lines and two individuals per line are sampled. When the proportion of these alleles is 0.9/0.1, even if 20 isofemale lines are drawn in the first sample, only one allele can be detected (11 cases in 100 repetitions). With regard to the locus with five alleles, the number of cases where at least one allele is not detected, increases. Between the base population and sample, it is necessary to shuttle a number of times, but in all situations the average number remains small and does not exceed 3. The

FIGURE 3.—Algorithm of the sampling protocol when isofemale lines method is used. *Path A* is used when some data about the population structure under study are known and *Path B* and *Path C* when no data are available: *Path B* corresponds to the choice of an arbitrary structure of the population and *Path C* to an arbitrary choice of the first sample size n_t without making a hypothesis about the population structure.

average of allelic frequencies observed when the algorithm is finished are very close to the initial drawing probabilities and the variances (not given) of these estimations range between 10^{-3} and 10^{-4} .

DISCUSSION

The number of isofemale lines required in a sample depends on the initial risk chosen, on the number of alleles and on their respective frequencies in the natural population. When the population structure is accurately known it is easy to compute the number of isofemale lines that must be drawn to get a reliable (as measured by r_t) estimate on the allelic diversity measured (n_a). When no data are available about the population structure we suggest using the sequential procedure described herein. This sampling plan converges to an estimate of the population structure for a given risk.

Assuming that a wild female can be inseminated by a single male, it is advocated to draw an average of two individuals per line. In fact, in *Drosophila* species, it has been shown that wild females can be inseminated by more than one male. However, the rate of multiple matings seems to be different from one species to another and from one population to another (ZOUROS and KRIMBAS 1970; MILKMAN and ZEITLER 1974; PETIT, BOURGERON and MERCOT 1980). In the following expression it is assumed that a wild female can be inseminated by a maximum of m males and f_x represents the proportion of females inseminated by x males. In these conditions, the average number of alleles present in a single line which come from the different male parents is,

$$A1 = 2 \left(1 + \sum_{x=1}^{x=m} f_x \right). \quad (11)$$

TABLE 3

Examples of four repetitions of the estimation of the allele number in the same theoretical population

Repetitions	Step	No. of lines sampled	No. of complementary lines	Allelic frequencies observed (in brackets the exact frequencies)					$Pe(na')$	re
				A1 (0.400)	A2 0.400	A3 0.100	A4 0.050	A5 0.050)		
1	1	10		0.350	0.375	0.175	0.075	0.025	0.607	0.397
	2	37	27	0.432	0.338	0.135	0.027	0.068	0.983	0.017
2	1	10		0.275	0.550	0.100	0.075		0.941	0.059
	2	13	3	0.269	0.558	0.077	0.096		0.979	0.021
3	1	10		0.425	0.450	0.100	0.025		0.627	0.373
	2	37	27	0.392	0.459	0.067	0.014	0.068	0.866	0.134
	3	68	31	0.415	0.415	0.088	0.011	0.071	0.951	0.049
4	1	10		0.400	0.425	0.075	0.100		0.941	0.059
	2	13	3	0.423	0.385	0.058	0.058	0.076	0.896	0.104
	3	19	6	0.408	0.355	0.105	0.039	0.093	0.952	0.048

$na = 5$; $p_1 = 0.4$; $p_2 = 0.4$; $p_3 = 0.1$; $p_4 = 0.05$; $p_5 = 0.05$; na' = number of alleles detected; $Pe(na')$ = *a posteriori* probability of the sample; re = sampling error.

TABLE 4

Mean and variance of the number of alleles detected and allelic frequencies observed when the sampling algorithm is stopped

Size of first sample	Population structure	Isofemale lines		Steps		Alleles detected		Allele frequencies				
		M	V	M	V	M	V	A1	A2	A3	A4	A5
5	A	5.13	1.70	1.01	0.01	2.00	0.00	0.497	0.503			
	B	9.54	36.20	1.57	0.50	1.80	0.16	0.887	0.113			
	C	30.29	501.00	2.64	1.12	4.48	0.82	0.405	0.388	0.116	0.051	0.040
10	A	10.00	0.00	1.00	0.00	2.00	0.00	0.494	0.506			
	B	15.20	94.20	1.30	0.25	1.95	0.05	0.890	0.110			
	C	30.13	304.00	2.19	0.64	4.70	0.33	0.393	0.388	0.110	0.063	0.046
20	A	20.00	0.00	1.00	0.00	2.00	0.00	0.500	0.500			
	B	20.50	8.60	1.03	0.03	1.99	0.01	0.893	0.107			
	C	38.74	474.00	1.73	0.46	4.89	0.09	0.401	0.394	0.990	0.058	0.048

The same process has been performed 100 times. Three situations were simulated: A) $na = 2$; $p_1 = 0.5$; $p_2 = 0.5$; B) $na = 2$; $p_1 = 0.9$; $p_2 = 0.1$ and C) $na = 5$; $p_1 = 0.4$; $p_2 = 0.4$; $p_3 = 0.1$; $p_4 = 0.05$; $p_5 = 0.05$. For each situation, the size of the first sample was 5, 10 or 20 isofemale lines. M = mean and V = variance.

Therefore, the average number of individuals that must be assayed per line becomes,

$$\frac{ni}{nl} = \frac{A1}{2} \quad (12)$$

As it has been already stressed the expression of multinomial can be used in two different ways: 1, to determine the sample size and 2, to compute the probability of a sample. This last application is useful for the comparison between two or several samples of natural populations. In all statistical tests, the type I error chosen must be superior to the higher sampling error or each sample with a sampling error superior to the type I error chosen, must be eliminated.

In the present work, the choice of the statistic is based on the detection of a number of alleles and not on the confidence interval of their respective frequencies. However, when the sample size is computed according to this statistic, the allelic frequencies observed are close to the drawing probabilities simu-

lated. In all cases, the confidence interval of the allelic frequency observed includes the initial drawing probability. To improve the estimation of the allelic frequencies, it is always possible, after the determination of the number of alleles, to continue the sampling with a new statistic like "the length of confidence interval of each allelic frequency inferior to a given value." However, if a new allele is detected, it is necessary to come back to the first statistic.

We thank S. McEVEY and two anonymous reviewers for their helpful suggestions and comments on this paper.

LITERATURE CITED

- ARNOLD, J., 1981 Statistics of natural populations. I. Estimating an allele probability in cryptic fathers with a fixed number of offspring. *Biometrics* **37**: 495-504.
- ARNOLD, J., and M. L. MORRISON, 1985 Statistics of natural populations. II. Estimating an allele probability in families descended from cryptic mothers. *Genetics* **109**: 785-798.
- CANNINGS, C., and A. W. F. EDWARDS, 1968 Expected genotypic frequencies in small sample: deviation from Hardy-Weinberg

- equilibrium. *Am. J. Hum. Genet.* **21**: 245–247.
- CANNINGS, C., and E. A. THOMPSON, 1977 Ascertainment in the sequential sampling of pedigrees. *Clin. Genet.* **12**: 208–212.
- CHOW, Y. S., and K. F. YU, 1981 The performance of a sequential procedure for the estimation of the mean. *Ann. Stat.* **9**: 184–189.
- COTTERMAN, C. W., 1954 Estimation of gene frequencies in non-experimental populations. In *Statistics and Mathematical In: Biology*, Edited by O. KEMTHORNE, T. A. BANCROFT, J. W. GOWEN and J. L. LUSH. Iowa State College Press, Ames, Iowa.
- CRESSIE, N., and A. SEHEULT, 1985 Empirical Bayes estimation in sampling inspection. *Biometrika* **72**: 451–458.
- EDWARDS, A. W. F., 1960 On a method of estimating frequencies using the negative binomial distribution. *Ann. Hum. Genet.* **24**: 313–318.
- EFRON, B., 1984 Bootstrap confidence interval for a class of parametric problems. *Biometrika* **72**: 45–58.
- EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**: 87–112.
- FINNEY, D. J., 1948 The estimation of the gene frequencies from family records. I. Factors without dominance. *Heredity* **2**: 199–218.
- FISHER, R. A., A. S. CORBET and C. B. WILLIAMS, 1943 The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**: 42–58.
- FUERST, P. A., R. CHAKRABORTY and M. NEI, 1977 Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* **86**: 455–483.
- GREGORIUS, H. R., 1980 The probability of losing an allele when diploid genotypes are sampled. *Biometrics* **36**: 643–652.
- HALDANE, J. B. S., 1945 On a method of estimating frequencies. *Biometrika* **33**: 222–225.
- LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.
- MILKMAN, R., and R. R. ZEITLER, 1974 Concurrent multiple paternity in natural and laboratory populations of *Drosophila melanogaster*. *Genetics* **78**: 1191–1193.
- PARSONS, P. A., 1980 Isofemale strains and evolutionary strategies in natural populations. *Evol. Biol.* **13**: 175–217.
- PEET, R. K., 1974 The measurement of species diversity. *Annu. Rev. Ecol. Syst.* 295–306.
- PETIT, C., P. BOURGERON and H. MERCOT, 1980 Multiple matings, effective population size and sexual selection in *Drosophila melanogaster*. *Heredity* **45**: 281–292.
- ROUAULT, J., and P. CAPY, 1986 Comment évaluer un nombre de catégories par échantillonnage. *Ann. Econ. Stat.* 111–124.
- SOBEL, M. J., J. ARNOLD and M. SOBEL, 1986 Statistics of natural populations. III. Sequential sampling plans for the estimation of gene frequencies. In press.
- WATTERSON, G. A., 1974 The sampling theory of selectively neutral alleles. *Adv. Appl. Prob.* **6**: 463–488.
- ZOUROS, E., and C. B. KRIMBAS, 1970 Frequency of female digamy in natural population of the olive fruitfly *Dacus oleae* as found by using enzyme polymorphism. *Entomol. Exp.* **13**: 1–9.