Copyright © 1985 by the Genetics Society of America

STATISTICS OF NATURAL POPULATIONS. II. ESTIMATING AN ALLELE PROBABILITY IN FAMILIES DESCENDED FROM CRYPTIC MOTHERS

JONATHAN ARNOLD[1] AND MELVIN L. MORRISON

Department of Genetics, University of Georgia, Athens, Georgia 30602

Manuscript received October 12, 1984
Revised copy accepted December 4, 1984

ABSTRACT

In population studies, adults are frequently difficult or inconvenient to identify for genotype, but a family profile of genotypes can be obtained from an unidentified female crossed with a single unidentified male. The problem is to estimate an allele frequency in the cryptic parental gene pool from the observed family profiles. For example, a worker may wish to estimate inversion frequencies in Drosophila; inversion karyotypes are cryptic in adults but visible in salivary gland squashes from larvae. A simple mixture model, which assumes the Hardy-Weinberg law, Mendelian laws and a single randomly chosen mate per female, provides the vehicle for studying three competing estimators of an allele frequency. A simple, heuristically appealing estimator called the Dobzhansky estimator is compared with the maximum likelihood estimator and a close relative called the grouped profiles estimator. The Dobzhansky estimator is computationally simple, consistent and highly efficient and is recommended in practice over its competitors.

WORKERS (DOBZHANSKY and POWELL 1975) collect adult Drosophila from nature to make inferences about the inversion polymorphism in adult populations. Adult genotypes are *cryptic*, but the genotype becomes visible in larvae. Collected adults are taken into the laboratory, and a fixed number of offspring per adult are identified for inversion genotype. Some uses of this *familial data* are to (i) estimate inversion frequencies, (ii) test population structure (*e.g.*, the Hardy-Weinberg law), (iii) test for selection components (*e.g.*, fertility and viability) and (iv) estimate the frequency of multiple insemination. The complication of cryptic parental genotypes arises not only in the studies of inversions in Drosophila (DOBZHANSKY and EPLING 1944; STALKER 1976; CARSON 1983) but also in the study of human blood groups (FINNEY 1948; CEPPELLINI, SINISCALCO and SMITH 1955), of allozymes in conifer populations (MORRIS and SPIETH 1978), of color morphs in salamanders (HIGHTON 1975) and of allozymes in termites (LUYKX 1981), in Tribolium (SAMOLLOW, DAWSON and RIDDLE 1983) and in insect-pollinated plants such as morning glories (CLEGG and SCHOEN 1984) due to missing parental data. The problem herein to be addressed is the estimation of an allele frequency among cryptic parents from their familial data, as in the examples just given.

Consider a diploid species, such as *Drosophila pseudoobscura* (ANDERSON *et al.*

[1] To whom correspondence should be addressed.

1975), and select a large isolated pocket of the species for study (*e.g.*, Bogotá, Colombia). Adult males and females are collected from this isolate. Preliminary collections, together with a laboratory analysis of offspring, reveal a trait under genetic control of two or more alleles; furthermore, collected females are usually inseminated only once. The collector's interest centers on a particular allele with the other allele(s) treated as one. In deciding to monitor the population, the collector hopes that a combined field and laboratory analysis of the natural population will allow him to estimate the particular allele's frequency.

ARNOLD (1981) describes a model for collecting N cryptic fathers with n offspring per father identified for genotype. The model provides a context in which to estimate a particular allele frequency. Here, we describe a model for the second basic experimental protocol, the collection of N singly inseminated mothers, from each of whom n offspring are identified for genotype. The current model differs from the earlier one (ARNOLD 1981) in two ways: First, there are two unknown parental genotypes in the second protocol (*vs.* one unknown parental genotype in the first protocol), and second, there is an additional assumption about the natural population's mating structure being random. An adult's chromosome in nature carries one allele, hereafter referred to as the allele, with probability θ, or an adult's chromosome in nature carries the other allele with probability $\bar{\theta} = 1 - \theta$. A collected mother carries y copies of the allele. The count y must be 0, 1 or 2, and it can be identified with the cryptic genotype of the collected female. If y is 2, she is homozygous for the allele; if y is 1, she is heterozygous. Assume that each mother mates with one father in nature. The parental genotypes of the mother and father are denoted by y_0 and y_1, respectively, and these parental genotypes $\underline{y} = (y_0, y_1)$ are cryptic to the collector.

The field collector takes the collected mother into the laboratory, obtains progeny and identifies the genotype of each of her offspring. For a given mother, an examination of n of her offspring yields (i) n_2 offspring, which are homozygous for the allele, each with probability w; (ii) n_0 offspring, which are homozygous for the other allele, each with probability u and (iii) n_1 offspring, which are heterozygous, each with probability v. The *family profile* of counts $\underline{n} = (n_0, n_1, n_2)$ is the *familial* genetic *data* on one collected cryptic mother. The number $n = n_0 + n_1 + n_2$ is *fixed* by the experimentor.

When the mother mates only once, then the Mendelian laws of inheritance give values for the probabilities u, v and w, conditional on the cryptic parental genotypes $\underline{y}$ These values are found in Table 1. With a knowledge of the parental genotypes, the Mendelian laws are summarized in the family of conditional densities $F = \{\text{pr}(\underline{n} \mid \underline{y}): \underline{y} = (0, 0), (0, 1), \ldots\}$. A conditional density $\text{pr}(\underline{n} \mid \underline{y})$ for the family profile $\underline{n}$ of one collected mother is trinomial with

$$\text{pr}(\underline{n} \mid \underline{y}) = \binom{n}{\underline{n}} u^{n_0} v^{n_1} w^{n_2}, \tag{1}$$

where the trinomial coefficient $\binom{n}{\underline{n}} = n!/n_0!n_1!n_2!$.

TABLE 1

The model

			Mendelian laws		
$\text{pr}(\underline{y}\mid\theta)$	y_0	y_1	u	v	w
$\bar{\theta}^4$	0	0	1	0	0
$2\bar{\theta}^3\theta$	0	1	½	½	0
$\bar{\theta}^2\theta^2$	0	2	0	1	0
$2\bar{\theta}^3\theta$	1	0	½	½	0
$4\bar{\theta}^2\theta^2$	1	1	¼	½	¼
$2\bar{\theta}\theta^3$	1	2	0	½	½
$\bar{\theta}^2\theta^2$	2	0	0	1	0
$2\bar{\theta}\theta^3$	2	1	0	½	½
θ^4	2	2	0	0	1

Although the parental genotypes are cryptic, the Hardy-Weinberg law provides a probability density over the parental genotypes. Each adult chromosome carries the allele with probability θ, where the parameter θ lies in the interval [0, 1]. The Hardy-Weinberg law implies that the probability density of parental genotypes denoted by $\text{pr}(y_0\mid\theta)$ for a mother (and $\text{pr}(y_1\mid\theta)$ for a father) is binomial with parameters θ and 2. This probability density is explicitly written out in equation 1.2 of ARNOLD (1981). Furthermore, the Hardy-Weinberg law also assumes random mating and that none of the factors of evolution are operating (*e.g.*, selection), so the probability of drawing a mother of genotype y_0 and a father of genotype y_1 is $\text{pr}(\underline{y}\mid\theta) = \text{pr}(y_0\mid\theta)\text{pr}(y_1\mid\theta)$. The probabilities $\text{pr}(\underline{y}\mid\theta)$ of parental combinations are listed in Table 1. Together, the Mendelian laws and the Hardy-Weinberg law in Table 1 provide the model specification:

$$\text{pr}(\underline{n}\mid\theta) = \Sigma_{\underline{y}}\,\text{pr}(\underline{n}\mid\underline{y})\text{pr}(\underline{y}\mid\theta). \tag{2}$$

The family profiles $\Lambda = \{\underline{n}\colon n_0 \geq 0,\ n_1 \geq 0,\ n_2 \geq 0,\ n_0 + n_1 + n_2 = n\}$ are defined in (and the density (2) lies on) a sample space Λ, which is the interior and sides of an equilateral triangle with height n. The density $\text{pr}(\underline{n}\mid\theta)$ is called a finite *θ-mixture* or simply finite *mixture* of the Mendelian family F. The family F is the *kernel* of the mixture, and the density $\text{pr}(\underline{y}\mid\theta)$ is a *product-mixing density*. A collection of N inseminated mothers is then viewed as a simple random sample drawn from this θ-mixture model specification (2). The statistical analysis of finite mixture models has been reviewed by EVERITT and HAND (1981). The problem is to select one model specification $\text{pr}(\underline{n}\mid\theta)$ out of the family of possible model specifications $M = \{\text{pr}(\underline{n}\mid\theta)\colon\theta \in [0, 1]\}$ to describe the isolate collected, *i.e.*, to estimate θ, the allele probability.

INFERENCE

As reviewed by EVERITT and HAND (1981) and considered by JAMES (1978), HALL (1981) and GANESALINGAM and MCLACHLAN (1981) recently, one prob-

TABLE 2

The score statistic ($\underline{N}$) *and the Dobzhansky score for collected mothers*

Event	Probability	Count ($\underline{N}$)	Weight	Score
$\Lambda_{00} = \{\underline{n}:n_0 = n\}$	$K_{00} = (\bar{\theta}^2 + a\theta\bar{\theta})^2$	N_{00}	0	0 + 0
$\Lambda_{02} = \{\underline{n}:n_1 = n\}$	$K_{02} = 2\theta\bar{\theta}(a + \bar{a}\theta\bar{\theta})$	N_{02}	½	0 + 2
$\Lambda_{22} = \{\underline{n}:n_2 = n\}$	$K_{22} = (\theta^2 + a\theta\bar{\theta})^2$	N_{22}	1	2 + 2
$\Lambda_{01} = \{\underline{n}:n_0 + n_1 = n,\ n_0 > 0,\ n_1 > 0\}$	$K_{01} = 4\theta\bar{\theta}^2(\bar{a}\bar{\theta} + b\theta)$	N_{01}	¼	0 + 1
$\Lambda_{12} = \{\underline{n}:n_1 + n_2 = n,\ n_1 > 0,\ n_2 > 0\}$	$K_{12} = 4\theta^2\bar{\theta}(\bar{a}\theta + b\bar{\theta})$	N_{12}	¾	1 + 2
$\Lambda_{11} = \{\underline{n}:n_0 + n_1 + n_2 = n,\ n_0 > 0,\ n_1 \geq 0,\ n_2 > 0\}$	$K_{11} = 4\theta^2\bar{\theta}^2 c$	N_{11}	½	1 + 1

$\bar{\theta} \equiv 1 - \theta$; $a \equiv (\frac{1}{2})^{n-1}$; $\bar{a} \equiv 1 - a$; $b \equiv (\frac{3}{4})^n - (\frac{1}{2})^n - (\frac{1}{4})^n$; $c \equiv 1 - 2(\frac{3}{4})^n + (\frac{1}{2})^n$.

lem considered by a number of authors is the estimation of the mixing density of a θ-mixture with known kernel. That is, one problem is to infer the allele probability θ. The estimation problem is well defined if, and only if, no two distinct allele probabilities θ and θ^* yield the same model specification $\text{pr}(\underline{n}\,|\,\theta)$. More precisely, the family M is *identifiable* if, and only if, for all $\text{pr}(\underline{n}\,|\,\theta)$ in M, the relationship $\text{pr}(\underline{n}\,|\,\theta) = {}_{\underline{n}}\ \text{pr}(\underline{n}\,|\,\theta^*)$ holding for all family profiles $\underline{n}$ implies $\theta = \theta^*$. A general discussion of identifiability can be found by EVERITT and HAND (1981, pp. 5–7).

Result 1: The family of model specifications, M, is identifiable. The proof proceeds similarly to that by ARNOLD (1981) by using a partition of the sample space generated by the counts of family profiles based on the presence or absence of one or more genotypes that will be described. If a collection of N mothers is taken from nature, and if n offspring are examined in the family of each mother, the collection can be thought of as a simple random sample of size N from the model specification (2). Six mutually exclusive events Λ_{00}, ..., Λ_{11} are singled out for special consideration (Table 2). In this table, $\bar{\theta} \equiv 1 - \theta$, $a \equiv (\frac{1}{2})^{n-1}$; $\bar{a} \equiv 1 - a$; $b \equiv (\frac{3}{4})^n - (\frac{1}{2})^n - (\frac{1}{4})^n$; and $c \equiv 1 - 2(\frac{3}{4})^n + (\frac{1}{2})^n$. The constants a, b and c all lie on [0, 1]. The events Λ_{00}, ..., Λ_{11} constitute a partition of the different possible family profiles $\underline{n}$ in the sample space Λ. This partition identifies types of family profiles $\underline{n}$ by the presence or absence of one or more genotypes in a family. The partition can be described by the corners, two sides, an interior, plus one side of the sample space Λ in Figure 1. The probabilities K_{00}, ..., K_{11}, of each of these six events can be computed from the model specification (2) (Table 2). In a collection of N mothers, the numbers N_{00}, ..., N_{11} count types of family profiles based on the presence or absence of one or more genotypes. For example, the event Λ_{00} in the sample space Λ is a family profile with all offspring homozygous for the other allele. The number N_{00} is a count of those family profiles with only events Λ_{00} in a collection of size N. The joint density of the list $\underline{N} = (N_{00}, \ldots, N_{11})$ is multinomial and is given in result 2.

Result 2: The probability density for the list of counts of family profile types,

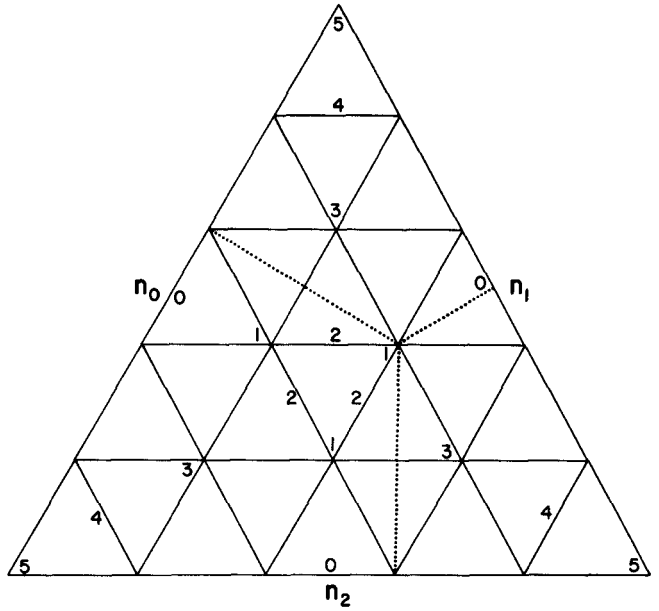

FIGURE 1.—The sample space Λ of possible family profiles $\underline{n}$ for $n = 5$. The dotted perpendicular distance from the side labeled n_2 is the number of offspring in a family of size 5 with genotype $y = 2$. The other two perpendicular distances represent n_1 and n_0.

$\underline{N} = (N_{00}, \ldots, N_{11})$, is multinomial:

$$\text{pr}(\underline{N} \mid \theta) = \binom{N}{\underline{N}} K_{00}^{N_{00}} K_{02}^{N_{02}} K_{22}^{N_{22}} K_{01}^{N_{01}} K_{12}^{N_{12}} K_{11}^{N_{11}}. \tag{3}$$

The multinomial coefficient $\binom{N}{\underline{N}} = N!/N_{00}!N_{02}!N_{22}!N_{01}!N_{12}!N_{11}!$.

If we accept the model, *nearly all* of the necessary information about the allele probability in a collection is contained in the list of counts $\underline{N}$, as will be shown later in this section. Thus, if a collector thought the model was appropriate and wished to know what data to record on each collected mother, we would recommend that the collector record the counts of different types of family profiles and the number of offspring per mother.

For 50 yr, Dobzhansky and coworkers have used certain heuristic procedures to estimate allele probabilities without formal justification of those procedures. ARNOLD (1981) justifies the use of an allele probability estimator based on familial data from collected fathers. A similar procedure has been used throughout the study of the inversion polymorphism in *D. pseudoobscura* for familial data on collected mothers. The "Dobzhansky estimator" for collected mothers was first fully described by DOBZHANSKY *et al.* (1963). It is defined as a weighted average of the list of counts $\underline{N}$:

$$\hat{\theta}_D = \left(N_{22} + \frac{3}{4}N_{12} + \frac{1}{2}N_{11} + \frac{1}{2}N_{02} + \frac{1}{4}N_{01}\right) \Big/ N. \tag{4}$$

From the well-known properties of the multinomial distribution in (3), result 3 follows.

Result 3: The Dobzhansky estimator $\hat{\theta}_D$ is unbiased, consistent and has the following variance

$$\text{var}(\hat{\theta}_D) = \frac{\theta\bar{\theta}}{4N}[1 + a + 2\theta\bar{\theta}(2\bar{a} - d)]. \tag{5}$$

As before, the constant $a \equiv 2^{-(n-1)}$ and $d \equiv 2 - (\frac{3}{4})^n - (\frac{1}{2})^n - 3(\frac{1}{4})^n$. The heuristic argument for the estimator is based on a "scoring procedure." A scoring procedure is a rule for inferring the number of each allele in the parents in nature and observed in offspring in the laboratory. A "score" is the inference made from the scoring procedure. The Dobzhansky scoring procedure is found in Table 2. Given a family profile $\underline{n}$, his procedure is equivalent to selecting parental genotypes $\underline{y}$ so that $\text{pr}(\underline{n} \mid \underline{y})$ is maximum, and the score equals $y_0 + y_1$. For example, suppose that the family profile obtained was $\underline{n} = (0, 0, n)$; all offspring are homozygous for the allele. Then, (1) both parents could be homozygous for the allele ($\underline{y} = (2, 2)$, $\text{pr}(\underline{n} \mid \underline{y}) = 1$); (2) one parent could be homozygous for the allele and the other heterozygous ($\underline{y} = (2, 1)$ or $(1, 2)$, $\text{pr}(\underline{n} \mid \underline{y}) = a/2$) or (3) both parents could be heterozygous ($\underline{y} = (1, 1)$, $\text{pr}(\underline{n} \mid \underline{y}) = a^2/4$). Given the parental karyotypes, the odds for these three possibilities are $1{:}a/2{:}a^2/4$. The probability $\text{pr}(\underline{n} \mid \underline{y})$ is maximum when $\underline{y} = (2, 2)$, yielding a score of 2 + 2. The scoring procedure is applied to the family profile of each mother. Summing the scores and normalizing by $4N$ (the number of alleles in all parents) yields the Dobzhansky estimator (4).

The variance $\text{var}(\hat{\theta}_D)$ is approximately $\theta\bar{\theta}[1 + a]/4N$ as could be predicted from result 3 by ARNOLD (1981). The extra term $\theta\bar{\theta}[2\theta\bar{\theta}(2\bar{a} - d)]/4N$ is exactly zero for $n = 1, 2$ and is maximal for $n = 5$. This extra term adds no more than 6% to the variance $\text{var}(\hat{\theta}_D)$ in (5). The approximate variance is intuitively appealing. The variance, $\text{var}(\hat{\theta}_D) \simeq \theta\bar{\theta}[1 + a]/4N$, takes the usual sample variance used by geneticists, $\theta\bar{\theta}/4N$, and enlarges it by $(1 + a)$ because of the uncertainty about cryptic parental genotypes. This result 3 here and result 3 by ARNOLD (1981) provide a simple adjustment to the collection size N reported in current geographic surveys, thus allowing one to compute a standard error on allele probability estimates. From result 3 in this paper, the relevant adjusted collection size N_A should be $N_A = N/(1 + a)$.

The counts $\underline{N}$ of family profile types allow us to establish that the model is identifiable and to characterize the Dobzhansky estimator. A natural question is whether or not the list of counts $\underline{N}$ contains *all* of the relevant information for estimating the allele probability θ. There remains a minute amount of information relevant for estimating θ beyond that in the list of counts $\underline{N}$. When there are two kinds of juvenile genotypes in a family, one kind being homozygous and the other kind being heterozygous (*i.e.*, $\underline{n} \in \Lambda_{01}$ *or* Λ_{12}), the *number* of heterozygotes in the offspring provides a small amount of additional information about θ. So, we need to further subdivide family profiles in Λ_{01} or Λ_{12} according to their number of heterozygotes. We introduce new events $\Lambda_{01}(j)$ and $\Lambda_{12}(j)$ in Λ_{01} and Λ_{12}, respectively. These new events partition some family

profiles by the number j of heterozygotes observed in a family profile $\underline{n}$. Define the new events:

$$\Lambda_{01}(j) = \{\underline{n}: n_0 = n - j,\ n_1 = j\},$$

$$\Lambda_{12}(j) = \{\underline{n}: n_2 = n - j,\ n_1 = j\}.$$

The event $\Lambda_{01}(j)$ is a family profile that contains j heterozygous offspring and $n - j$ offspring homozygous for the other allele. These new events are points on the sides $n_2 = 0$ and $n_0 = 0$, respectively, of Figure 1. The probabilities of these new events contained in Λ_{01} or Λ_{12} can be calculated from (2):

$$K_{01}(j) = \text{pr}(\underline{n} = (n - j, j, 0) \mid \theta) = 4\theta\bar{\theta}^2(\bar{a}_j\bar{\theta} + b_j\theta),$$

$$K_{12}(j) = \text{pr}(\underline{n} = (0, j, n - j) \mid \theta) = 4\theta^2\bar{\theta}(\bar{a}_j\theta + b_j\bar{\theta}),$$

where the constants $\bar{a}_j = \binom{n}{j}(1/2)^n$ and $b_j = \binom{n}{j}(1/4)^{n-j}(1/2)^j$. When these constants are summed ($j = 1, \ldots, n - 1$) in the probabilities $K_{01}(j)$ or $K_{12}(j)$ to yield $\bar{a} = \bar{a}_+$ and $b = b_+$, then we recover the probabilities K_{01} or K_{12} of a family profile containing one homozygous type and the heterozygous type in Table 2. The joint density of the list of counts $\underline{N}' = (N_{00}, N_{02}, N_{22}, N_{01}(1), \ldots, N_{01}(n - 1), N_{12}(1), \ldots N_{12}(n - 1), N_{11})$ is also multinomial like the density of $\underline{N}$ and is given in result 2A.

Result 2A: The minimal sufficient statistic is $\underline{N}'$, the list of counts of family profile types, where $\underline{N}' = (N_{00}, N_{02}, N_{22}, N_{01}(1), \ldots, N_{01}(n - 1), N_{12}(1), \ldots N_{12}(n - 1), N_{11})$. The probability density of this sufficient statistic is also multinomial:

$$\text{pr}(\underline{N}' \mid \theta) = \binom{N}{\underline{N}'} K_{00}^{N_{00}} K_{02}^{N_{02}} K_{22}^{N_{22}} \prod_{j=1}^{n-1} K_{01}(j)^{N_{01}(j)} \prod_{j=1}^{n-1} k_{12}(j)^{N_{12}(j)} K_{11}^{N_{11}}. \quad (6)$$

The multinomial coefficient $\binom{N}{N'}$ is defined analogously to $\binom{N}{\underline{N}}$ in result 2.

When the multinomial probability given in (6) is considered as a function of θ and fixed at some list of realized counts $\underline{N}'$, the resulting likelihood also yields the maximum likelihood estimator $\hat{\theta}_{\text{ML}}$ advocated by several authors. These authors describe three different methods of computing the maximum likelihood estimator for models such as (2): (i) Fisher's maximum likelihood scoring (FINNEY 1948); (ii) gene counting (CEPPELLINI, SINISCALCO and SMITH 1955; HABERMAN 1977) and (iii) iteratively reweighted least squares (THOMPSON and BAKER 1981). It is desirable then to explore the relative properties of the Dobzhansky estimator $\hat{\theta}_{\text{D}}$ and the maximum likelihood estimator $\hat{\theta}_{\text{ML}}$. Computing the fully efficient maximum likelihood estimator by iteratively reweighted least squares (BURN 1982) is an attractive *alternative* to the Dobzhansky estimator because it can easily be generalized for use in other inversion studies; it provides a variance; it provides goodness of fit, and it can be implemented by use of existing statistical packages. The relative merits of the other two methods are discussed by HABERMAN (1977).

There is a third estimator analogous to the maximum likelihood estimator, which is based on the list of counts of family profile types $\underline{N}$. We will call this third alternative, $\hat{\theta}_{GP}$, the *grouped profiles estimator*. When the multinomial probability given in (3) is considered as a function of θ and fixed at some list of realized counts $\underline{N}$, the grouped profiles estimator $\hat{\theta}_{GP}$ will be that function of the counts $\underline{N}$ maximizing $\text{pr}(\underline{N} \mid \theta)$ with respect to θ. The major difference between this grouped profiles estimator $\hat{\theta}_{GP}$ and the maximum likelihood estimator $\hat{\theta}_{ML}$ is that the former sacrifices the small amount of information in the number j of heterozygotes in families with one homozygous type and the heterozygous type (*i.e.*, $\underline{n} \in \Lambda_{01}$ or Λ_{12}). As a result, the grouped profiles estimator is simpler to compute than $\hat{\theta}_{ML}$ but not as simple as the Dobzhansky estimator.

Result 4: The log likelihood function $\ln \text{pr}(\underline{N}' \mid \theta)$ and the function $\ln \text{pr}(\underline{N} \mid \theta)$ are concave over the interval (0, 1).

Result 4 implies that the maximum likelihood and grouped profiles estimators are unique when they exist. By taking the derivatives of $\ln \text{pr}(\underline{N}' \mid \theta)$ and $\ln \text{pr}(\underline{N} \mid \theta)$ with respect to θ, polynomial equations can be found, for which the estimators $\hat{\theta}_{ML}$ and $\hat{\theta}_{GP}$ are roots. Denote the likelihood equation by $\text{LE}(\hat{\theta}_{ML}) = 0$ and the equation for $\hat{\theta}_{GP}$, by $\text{L}(\hat{\theta}_{GP}) = 0$. These equations establish the existence of these two estimators.

Result 5: The likelihood equation $\text{LE}(\hat{\theta}_{ML}) = 0$ is a polynomial equation of degree at least 7 and has a unique root $\hat{\theta}_{ML} \in [0, 1]$. The equation $\text{L}(\hat{\theta}_{GP}) = 0$ is a polynomial equation of degree 7 and has a unique root $\hat{\theta}_{GP} \in [0, 1]$. The equation for $\hat{\theta}_{GP}$ is

$$\alpha\hat{\theta}_{GP}^7 + \beta\hat{\theta}_{GP}^6 + \gamma\hat{\theta}_{GP}^5 + \delta\hat{\theta}_{GP}^4 + \xi\hat{\theta}_{GP}^3 + \zeta\hat{\theta}_{GP}^2 + \iota\hat{\theta}_{GP} + \kappa = 0, \tag{7}$$

where $\alpha = \bar{a}^3(\bar{a} - b)^2 4N$, $\kappa = N_2 a^2 \bar{a} b$, $N_2 = 2N_{22} + 2N_{12} + 2N_{11} + N_{20} + N_{01}$ and the remaining coefficients are given in Table 3. A unique root always exists within the interval [0, 1].

TABLE 3

Equation for grouped profiles estimator

	ι
N_{00}	$(\bar{a} + 1)(-2a^2\bar{a}b)$
N_{01}	$(b - 2\bar{a})a^2 b$
N_{02}	$a\bar{a}^2 b$
N_{12}	$a^2\bar{a}^2$
N_{22}	$2a\bar{a}b$
N_2	$[-5a\bar{a}b + a\bar{a}^2 + \bar{a}b + ab^2 + \bar{a}^3 b]a$

	ζ
N_{00}	$[5a\bar{a}^2 b - a\bar{a}^3 - 2\bar{a}^2 b - a\bar{a}b^2 - \bar{a}^3 b + 3a\bar{a}b - a\bar{a}^2 - \bar{a}b - ab^2]2a$
N_{01}	$[-4ab^2 + 7a\bar{a}b - 2a\bar{a}^2 + b^2 - 2\bar{a}b + \bar{a}^2 b^2 - 2\bar{a}^3 b]a$
N_{02}	$[-5a\bar{a}b + \bar{a}^2 a + ab^2 + \bar{a}^3 b]\bar{a}$
N_{12}	$[-2a\bar{a} + \bar{a} + ab + \bar{a}^3]a\bar{a}$
N_{22}	$[-4a\bar{a}^2 b + 2a\bar{a}^3 + 2a\bar{a}b^2 + 2\bar{a}^4 b - 6a^2\bar{a}b + 2a^2\bar{a}^2 + 2a\bar{a}b + 2a^2 b^2 - 2a^2\bar{a}^2 b]$
N_2	$[10a^2\bar{a}b - 4a^2\bar{a}^2 - 4a\bar{a}b - 4a^2 b^2 + 5a^2\bar{a}^2 b - 7a\bar{a}^2 b + a\bar{a}^2 + ab^2 + a\bar{a}^4 + \bar{a}^2 b + a\bar{a}^2 b^2 - a\bar{a}^3 b]$

TABLE 3—continued

	ξ
N_{00}	$[-10a^2\bar{a}^2b + 4a^2\bar{a}^3 + 8a\bar{a}^2b + 4a^2\bar{a}b^2 + 6a\bar{a}^3b - 2a\bar{a}^3 - 2a\bar{a}b^2 - a\bar{a}^4 - \bar{a}^3b - a\bar{a}^2b^2 - 4a^2\bar{a}b + 2a^2\bar{a}^2 + 2a\bar{a}b + 2a^2b^2 - a\bar{a}^2 - ab^2 - \bar{a}^2b]2$
N_{01}	$[5a^2b^2 - 9a^2\bar{a}b + 4a^2\bar{a}^2 - 3ab^2 + 5a\bar{a}b + 4a^2\bar{a}b^2 - 7a^2\bar{a}^2b - 6a\bar{a}b^2 + 11a\bar{a}^2b - 2a\bar{a}^2 - 2a\bar{a}^4 + \bar{a}b^2 - 2\bar{a}^2b - a\bar{a}^2b^2 + 2a\bar{a}^3b]$
N_{02}	$[10a\bar{a}b - 4\bar{a}^2a - 4ab^2 - 6\bar{a}^3b + \bar{a}^4 + \bar{a}^2b^2]\bar{a}$
N_{12}	$[ab + 2a^2\bar{a} - 5a\bar{a} + a\bar{a}b + \bar{a} - a\bar{a}^2]\bar{a}^2$
N_{22}	$[-a\bar{a}^2b - a\bar{a}^3 - a\bar{a}b^2 + 2a\bar{a}^3b - 4\bar{a}^3b + \bar{a}^5 + \bar{a}^3b^2 - \bar{a}^4b + 4a^2\bar{a}b - 2a^2\bar{a}^2 - 2a\bar{a}b - 2a^2b^2 + 3a^2\bar{a}^2b + a\bar{a}^2 + ab^2 - a^2\bar{a}^3 - a^2\bar{a}b^2 + \bar{a}^2b]2$
N_2	$[-10a^2\bar{a}b + 5a^2\bar{a}^2 + 6a\bar{a}b + 5a^2b^2 - 10a^2\bar{a}^2b + 19a\bar{a}^2b - 3a\bar{a}^2 - 3ab^2 + 4a^2\bar{a}^3 + 4a^2\bar{a}b^2 - 6a\bar{a}^3 - 5\bar{a}^2b - 6a\bar{a}b^2 + 6a\bar{a}^3b + \bar{a}^3 + \bar{a}b^2 - a\bar{a}^4 - \bar{a}^3b - a\bar{a}^2b^2]$
	δ
N_{00}	$[10a^2\bar{a}^2b - 5a^2\bar{a}^3 - 15a\bar{a}^2b - 5a^2\bar{a}b^2 - 15a\bar{a}^3b + 7a\bar{a}^3 + 7a\bar{a}b^2 + 5a\bar{a}^4 + 5\bar{a}^3b + 5a\bar{a}^2b^2 - \bar{a}^4 - \bar{a}^2b^2 + 3\bar{a}^2b - \bar{a}^3 - \bar{a}b^2]2$
N_{01}	$[-5a^2\bar{a}b^2 + 9a^2\bar{a}^2b + 14a\bar{a}b^2 - 25a\bar{a}^2b - 4a^2\bar{a}^3 + 10a\bar{a}^3 - 4\bar{a}b^2 + 7\bar{a}^2b + 5a\bar{a}^2b^2 - 9a\bar{a}^3b - 2\bar{a}^3 + 2a\bar{a}^4 - \bar{a}^2b^2 + 2\bar{a}^3b]$
N_{02}	$[-10a\bar{a}b + 5\bar{a}^2a + 5ab^2 + 10\bar{a}^3b - 5\bar{a}^4 - 5\bar{a}^2b^2 + 5\bar{a}^3b]\bar{a}$
N_{12}	$[-a^2\bar{a} + 4a\bar{a} + a^2b - 2\bar{a} - 3ab + 3a\bar{a}^2 + b - \bar{a}^2 - a\bar{a}b]\bar{a}^2$
N_{22}	$[18a\bar{a}^2b - 8a\bar{a}^3 - 8a\bar{a}b^2 + 12\bar{a}^3b + 2a\bar{a}^4 + 2a\bar{a}^2b^2 - 6\bar{a}^4 - 6\bar{a}^2b^2 + 8\bar{a}^4b - 2\bar{a}^5 - 2\bar{a}^3b^2 - 8a^2\bar{a}^2b + 4a^2\bar{a}^3 + 4a^2\bar{a}b^2 - 6\bar{a}^2b + 2\bar{a}^3 + 2\bar{a}b^2]$
N_2	$[-2a^2\bar{a}^2 - 2a^2b^2 + 10a^2\bar{a}^2b - 30a\bar{a}^2b + 2a\bar{a}^2 + 2ab^2 - 5a^2\bar{a}^3 - 5a^2\bar{a}b^2 + 12a\bar{a}^3 + 10\bar{a}^2b + 12a\bar{a}b^2 - 15a\bar{a}^3b - 4\bar{a}^3 - 4\bar{a}b^2 + 5a\bar{a}^4 + 5\bar{a}^3b + 5a\bar{a}^2b^2 - \bar{a}^4 - \bar{a}^2b^2]$
	γ
N_{00}	$[-4a^2\bar{a}b + 2a^2\bar{a}^2 + 10a\bar{a}b + 2a^2b^2 + 20a\bar{a}^2b - 5a\bar{a}^2 - 5ab^2 - 9a\bar{a}^3 - 10\bar{a}^2b - 9a\bar{a}b^2 + 4\bar{a}^3 + 4\bar{a}b^2 - 4\bar{a}b + 2\bar{a}^2 + 2b^2]2\bar{a}$
N_{01}	$[2a^2b^2 - 9ab^2 - 4a^2\bar{a}b + 17a\bar{a}b + 5b^2 - 9a\bar{a}b^2 - 9\bar{a}b + 16a\bar{a}^2b + 4\bar{a}b^2 - 7\bar{a}^2b + 2a^2\bar{a}^2 - 8a\bar{a}^2 + 4\bar{a}^2 - 6a\bar{a}^3 + 2\bar{a}^3]\bar{a}$
N_{02}	$[4a\bar{a}b - 2\bar{a}^2a - 2ab^2 - 20\bar{a}^3b + 9\bar{a}^4 + 9\bar{a}^2b^2]\bar{a}$
N_{12}	$[3\bar{a} - 2b]\bar{a}^4$
N_{22}	$[-10a\bar{a}^2b - 6\bar{a}^2b + 4a\bar{a}^3 + 4a\bar{a}b^2 - 14\bar{a}^3b + 6\bar{a}^4 + 6\bar{a}^2b^2 + 4a^2\bar{a}b - 16a\bar{a}b - 2a^2\bar{a}^2 - 2a^2b^2 + 4a\bar{a}^2 + 8\bar{a}b + 4ab^2]\bar{a}$
N_2	$[-4a^2\bar{a}b + 18a\bar{a}b + 2a^2\bar{a}^2 + 2a^2b^2 - 9a\bar{a}^2 - 10\bar{a}b - 9ab^2 + 20a\bar{a}^2b + 5\bar{a}^2 + 5b^2 - 9a\bar{a}^3 - 10\bar{a}^2b - 9a\bar{a}b^2 + 4\bar{a}^3 + 4\bar{a}b^2]\bar{a}$
	β
N_{00}	$[7a - 6](\bar{a} - b)^2 2\bar{a}^2$
N_{01}	$[7b - 6\bar{a}](\bar{a} - b)\bar{a}^3$
N_{02}	$-7(\bar{a} - b)^2\bar{a}^3$
N_{12}	$(b - \bar{a})\bar{a}^4$
N_{22}	$(\bar{a} - b)^2(-2\bar{a}^2)$
N_2	$(\bar{a} - b)^2(-7\bar{a}^3)$

The eight coefficients $\alpha, \ldots, \kappa$ of the equation for $\hat{\theta}_{GP}$ are linear in the score statistic $\underline{N}$ and are obtained by multiplying each row by its row label (*e.g.*, $N_{00}, \ldots, N_2$) and summing all entries under a given coefficient. The constants a, $\bar{a}$ and b are defined in Table 2. The coefficients α and κ are in result 5.

Equation (7) for $\hat{\theta}_{GP}$ was verified by computing $\hat{\theta}_{GP}$ via two distinct numerical methods, Fisher's scoring method and the method of iteratively reweighted least squares (BURN 1982), and comparing the result with the root of (7). In the case in which family size is 2, the equation for $\hat{\theta}_{GP}$ is the likelihood equation, *i.e.*, $\hat{\theta}_{ML} = \hat{\theta}_{GP}$. The easiest way to solve (7) is to graph this polynomial over the interval [0, 1]. Note that (7) expresses the grouped profiles estimator $\hat{\theta}_{GP}$ in terms of the statistic $\underline{N}$. If $n = 1$, $n = \infty$ or $N = 1$ in result 5, the Dobzhansky, grouped profiles and maximum likelihood estimators are the same.

The three estimators can be compared in large samples by their relative efficiency; for this purpose, the variances of the three estimators are needed. Let CRLB denote the Cramer-Rao lower bound (RAO 1973) and $I(\theta) = (CRLB)^{-1}$ denote the Fisher information about θ in a sample of size 1 described by (6). Similarly, we can define the Cramer-Rao lower bound $CRLB_{GP}$ and Fisher information $I_{GP}(\theta)$ about θ in a sample of size 1 described by (3). As in RAO (1973, p. 368), we can establish result 6.

Result 6: If the allele probability $\theta \in [0, 1]$, then

$$I_{GP}(\theta) = \frac{1}{K_{00}}\left[\frac{dK_{00}}{d\theta}\right]^2 + \ldots + \frac{1}{K_{11}}\left[\frac{dK_{11}}{d\theta}\right]^2$$

and (8)

$$I(\theta) = \frac{1}{K_{00}}\left[\frac{dK_{00}}{d\theta}\right]^2 + \frac{1}{K_{02}}\left[\frac{dK_{02}}{d\theta}\right]^2 + \frac{1}{K_{22}}\left[\frac{dK_{22}}{d\theta}\right]^2 + \sum_{j=1}^{n-1}\frac{1}{K_{01}(j)}\left[\frac{dK_{01}(j)}{d\theta}\right]^2 + \sum_{j=1}^{n-1}\frac{1}{K_{12}(j)}\left[\frac{dK_{12}(j)}{d\theta}\right]^2 + \frac{1}{K_{11}}\left[\frac{dK_{11}}{d\theta}\right]^2$$

The Fisher information $I(\theta)$ and $I_{GP}(\theta)$ are symmetric about $\theta = \frac{1}{2}$, are convex, have a minimum value at $\theta = \frac{1}{2}$ and increase uniformly with family size n.

Results 3 and 6 permit one to calculate the *asymptotic* (large N, fixed n) relative efficiency of the Dobzhansky estimator, $\hat{\theta}_D$, and of the grouped profile estimator, $\hat{\theta}_{GP}$. The respective efficiencies are $CRLB/var(\hat{\theta}_D)$ and $CRLB/CRLB_{GP}$ (see Figure 2). When family size $n = 1$, the sufficient statistic $\underline{N}'$ in result 2A is binomial, and these relative efficiencies are 1. If the family size n is greater than 1, the relative efficiency of the Dobzhansky estimator is at least 0.90 on the interval [0.01, 0.99]; it is also concave on [0, 1]; it is symmetric about $\theta = \frac{1}{2}$ and has a maximum value at $\theta = \frac{1}{2}$. The same statements hold true for the relative efficiency of the grouped profiles estimator except that it is fully efficient for $n = 2$. Minimum values of these two relative efficiencies on the interval [0.01, 0.99] for different family sizes are

n	1	2	3	4	5
$\min\{CRLB/CRLB_{GP}\}$ $\theta \in [0.01, 0.99]$	1	1	0.995	0.995	0.995
$\min\{CRLB/var(\hat{\theta}_D)\}$ $\theta \in [0.01, 0.99]$	1	0.90	0.92	0.95	0.97

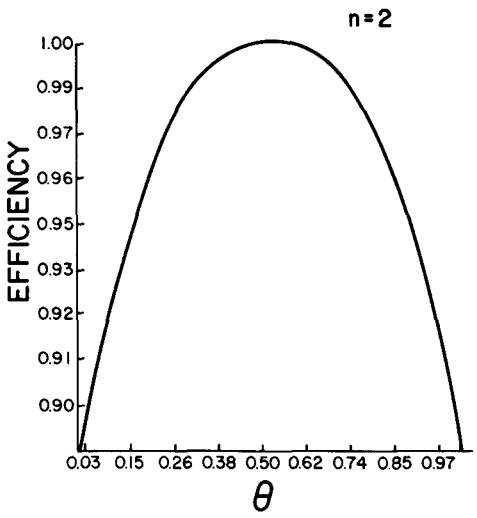

FIGURE 2.—Efficiency of the Dobzhansky estimator as a function of inversion frequency in the least favorable case, $n = 2$.

First, the counts $\underline{N}$ of family profile types contain *nearly all* of the information in a sample relevant for estimating an allele probability θ. Second, the Dobzhansky estimator $\hat{\theta}_D$ is highly efficient.

In conclusion, the Dobzhansky estimator is simple to compute, unbiased, consistent and highly efficient for the two basic experimental protocols in inversion studies (ARNOLD 1981), and its use is recommended.

AN ILLUSTRATION

A large geographic survey of the *D. pseudoobscura* inversion polymorphism is nearing completion. J. R. POWELL and L. B. KLACZKO have kindly provided us with the data in Table 4. The study has included four collections in the Texan Davis Mountains over a 50-yr period. The data in Table 4 are a random subsample of ten families obtained from mothers collected on July 24–25, 1982, in and around the Davis Mountains State Park. Previous data by ANDERSON *et al.* (1975) suggest that one can expect to find at least two inversions in the offspring, denoted Pike's Peak (PP) and Arrowhead (AR). The ten collected mothers were isolated in separate laboratory vials and allowed to lay eggs. Five larvae from each female were identified for genotype (Table 4).

The Dobzhansky estimator can be computed by the scoring procedure in Table 2 or by (4). Counting the numbers of family profiles in which only *PPPP* homozygotes appear (*e.g.*, $N_{22} = 4$, $N_{12} = 5$, etc.), then $\hat{\theta}_D = \left(4 + \frac{3}{4}5 + \frac{1}{2}0 + \frac{1}{2}0 + \frac{1}{4}1\right) \Big/ 10 = {}^{8}/_{10}$. This is the same as counting the number of PP inversions in the scores (*i.e.*, 32) and dividing by the number of parental chromosomes (*i.e.*, 40). The Dobzhansky estimate of the PP probability in the natural population is then 0.8. Substituting the Dobzhansky es-

TABLE 4

Familial data from D. pseudoobscura *mothers collected in the Davis Mountains, Texas (7/24–25/82)*

Game	Outcomes	Score	Event	Weight
1	PPPP, ARPP, ARPP, ARPP, ARPP	1 + 2	Λ_{12}	3/4
2	ARPP, ARPP, PPPP, ARPP, PPPP	1 + 2	Λ_{12}	3/4
3	PPPP, PPPP, ARPP, ARPP, ARPP	1 + 2	Λ_{12}	3/4
4	ARAR, ARPP, ARPP, ARAR, ARAR	0 + 1	Λ_{01}	1/4
5	PPPP, PPPP, PPPP, PPPP, PPPP	2 + 2	Λ_{22}	4/4
6	PPPP, PPPP, PPPP, PPPP, PPPP	2 + 2	Λ_{22}	4/4
7	PPPP, ARPP, PPPP, ARPP, ARPP	2 + 1	Λ_{12}	3/4
8	PPPP, PPPP, PPPP, PPPP, PPPP	2 + 2	Λ_{22}	4/4
9	PPPP, PPPP, PPPP, PPPP, PPPP	2 + 2	Λ_{22}	4/4
10	PPPP, ARPP, ARPP, ARPP, ARPP	1 + 2	Λ_{12}	3/4
				8

$N_{00} = 0$ $N_{02} = 0$ $N_{22} = 4$
$N_{01} = 1$ $N_{12} = 5$ $N_{11} = 0$
$N_{01}(2) = 1$ $N_{12}(3) = 3$
$N_{12}(4) = 2$
$\hat{\theta}_D = 8/10$

timate ($\hat{\theta}_D = 0.8$), the collection size ($N = 10$) and the family size ($n = 5$) into (5) yields a standard error estimate of 0.0666.

For inference purposes the only data needed are the Dobzhansky scores, given that the number of offspring examined from each family is five (Table 4). If family size were variable, it would also be necessary to record family size along with each Dobzhansky score.

For comparison, the maximum likelihood estimator $\hat{\theta}_{ML}$ and its variance estimator $1/\{NI(\hat{\theta}_D)\}$ were computed by maximum likelihood scoring. The Fisher information is provided in result 6. The Dobzhansky estimator ($\hat{\theta}_D = 0.80$, SE = 0.0666) is very close to the maximum likelihood estimator ($\hat{\theta}_{ML} = 0.79$, SE = 0.0720) and grouped profiles estimator ($\hat{\theta}_{GP} = 0.80$, SE = 0.0659). In more than 100 trials the Dobzhansky, grouped profiles and maximum likelihood estimators agreed to two decimal places. It is clear that the extra effort needed to obtain the maximum likelihood or grouped profiles estimates and the estimates of their standard errors is not warranted.

The Dobzhansky scoring procedure is simple and appropriate for collected females who are singly inseminated, whether or not their families are small or large. This recommended procedure produces an allele probability estimator that is unbiased, consistent and highly efficient.

CONCLUDING REMARKS

A number of new experimental protocols have been developed in inversion studies of *D. pseudoobscura*. A number of new mixture models need to be developed to handle the new protocols and existing situations, such as when there are variable numbers of offspring. Other situations have been described previously (ARNOLD 1981), and CLEGG and SCHOEN (1984) give a direct extension of the model in this paper for insect-pollinated plants. Simple estimators for these new mixture models need to be developed.

Having fitted various models to existing data, we should now test their goodness of fit and test various hypotheses made in the models. As an illustration, several models for differing experimental protocols in Drosophila inversion studies assume the Hardy-Weinberg law, such as those for cryptic fathers or mothers. Given this assumption, it would be worthwhile using the conditional probability of the sample given the sufficient statistic to construct a significance test (ANSCOMBE 1982, chapter 12). The probability of the sample is the basic test statistic first discussed in studies of Drosophila by LEVENE (1949).

There are several experimental protocols used for making inferences about cryptic parental population data, depending on the sex of the collected parent. Within any one protocol there is the decision of how many offspring, n, per collected parent are to be identified as to genotype and how many adults N are to be collected for a fixed total amount of field/experimental effort. The problem is to choose an optimal n and N to estimate an allele probability. Also, different experimental protocols must be compared for relative sampling efficiency. BROWN (1975) and BROWN, WEIR and MARSHALL (1970) consider similar design problems for estimating genetic parameters in plant populations.

M.L.M. would like to thank A. SOKOLOFF and K. W. COOPER for their continued encouragement. J.A. wishes to thank B. TAYLOR and M. ASMUSSEN for fruitful discussions. We are grateful to A. F. MACRAE, C. J. BROWN, R. BURN, B. WEIR, I. R. SAVAGE, and referees for their comments on this manuscript. The computing in this paper was made possible by Digital Equipment Corporation's generous gift of a PDP-11/34A minicomputer. This material is based upon work supported by the National Science Foundation under grant BSR-8315821.

LITERATURE CITED

ANDERSON, W. W., TH. DOBZHANSKY, O. PAVLOVSKY, J. R. POWELL and D. YARDLEY, 1975 Genetics of natural populations. XLII. Three decades of genetic change in *Drosophila pseudoobscura*. Evolution **29:** 24–36.

ANSCOMBE, F. J., 1982 *Computing in Statistical Science Through APL*. Springer-Verlag, New York.

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.

BROWN, A. H. D., 1974 Efficient experimental designs for the estimation of genetic parameters in plant populations. Biometrics **31:** 145–160.

BROWN, A. H. D., B. S. WEIR and D. R. MARSHALL, 1970 Optimum family size for the estimation of heterozygosity in plant populations. Heredity **25:** 145–160.

BURN, R., 1982 Loglinear models with composite link functions in genetics. pp. 144–154. In: *Proceedings of the International Conference Generalized Linear Models*, Edited by R. GILCHRIST. Springer-Verlag, New York.

CARSON, H. L., 1983 Chromosomal sequences and interisland colonizations in the Hawaiian Drosophila. Genetics **103:** 465–482.

CEPPELLINI, R., M. SINISCALCO and C. A. B. SMITH, 1955 The estimation of gene frequencies in a random-mating population. Ann. Hum. Genet. (Lond.) **20:** 97–115.

CLEGG, M. T. and D. J. SCHOEN, 1984 Estimation of mating system parameters when outcrossing events are correlated. Proc. Natl. Acad. Sci. USA **81:** 5258–5262.

DOBZHANSKY, TH. and C. EPLING, 1944 Contributions to the genetics, taxonomy, and ecology of Drosophila pseudoobscura and its relatives. Carnegic Inst. Wash. Publ. 554.

DOBZHANSKY, TH., A. S. HUNTER, O. PAVLOVSKY, B. SPASSKY and B. WALLACE, 1963 Genetics of natural populations. XXXI. Genetics of an isolated marginal population of *Drosophila pseudoobscura*. Genetics **48:** 91–103.

DOBZHANSKY, TH. and J. R. POWELL, 1975 *Drosophila peseudoobscura* and its American relatives, *Drosophila persimilis* and *Drosophila miranda*. pp. 537–587. In: *Handbook of Genetics*, Vol. 3, Edited by R. C. KING. Plenum Press, New York.

EVERITT, B. S. and D. J. HAND, 1981 *Finite Mixture Distributions*. Chapman and Hall, New York.

FINNEY, D. J., 1948 The estimation of gene frequencies from family records. I. Factors without dominance. Heredity **2:** 199–218.

GANESALINGAM, S. and G. J. MCLACHLAN, 1981 Some efficiency results for the estimation of the mixing proportion in a mixture of two normal distributions. Biometrics **37:** 23–33.

HABERMAN, S., 1977 Product models for frequency tables involving indirect observation. Ann. Statist. **5:** 1124–1147.

HALL, P., 1981 On the non-parametric estimation of mixture proportions. J. R. Statist. Soc. (B) **43:** 147–156.

HIGHTON, R., 1975 Geographic variation in genetic dominance of the color morphs of the red-backed salamander, *Plethodon cinereus*. Genetics **80:** 363–374.

JAMES, I. R., 1978 Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements. Biometrics **34:** 265–275.

LEVENE, H., 1949 On a matching problem arising in genetics. Ann. Math. Statist. **20:** 91–94.

LUYKX, P., 1981 A sex-linked esterase locus and translocation heterozygosity in a termite. Heredity **46:** 315–320.

MORRIS, R. W. and P. T. SPIETH, 1978 Sampling strategies for using female gametophytes to estimate heterozygosity in conifers. Theor. Appl. Genet. **51:** 217–222.

RAO, C. R., 1973 *Linear Statistical Inference and Its Applications*. Jon Wiley and Sons, New York.

SAMOLLOW, P. B., P. S. DAWSON and R. A. RIDDLE, 1983 X-linked and autosomal inheritance patterns of homologous genes in two species of *Tribolium*. Biochem. Genet. **21:** 167–176.

STALKER, H. D., 1976 Chromosome studies in wild populations of *Drosophila melanogaster*. Genetics **82:** 323–347.

THOMPSON, R. and R. J. BAKER, 1981 Composite link functions in generalized linear models. Appl. Statist. **30:** 125–131.

Communicating editor: J. R. POWELL