Statistics of Natural Populations. III. Sequential Sampling Plans for the Estimation of Gene Frequencies

Marc J. Sobel; Jonathan Arnold; Milton Sobel


Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198603%2942%3A1%3C45%3ASONPIS%3E2.0.CO%3B2-C

*Biometrics* is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ibs.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Statistics of Natural Populations. III. 
Sequential Sampling Plans for the Estimation 
of Gene Frequencies

Marc J. Sobel,1 Jonathan Arnold,2 and Milton Sobel3

1 Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.; 
2 Department of Genetics, University of Georgia, Athens, Georgia 30602, U.S.A.; 
3 Department of Mathematics, University of California, Santa Barbara, California 93106, U.S.A.

SUMMARY

In previous work several models have been developed for genetic surveys of natural populations. Parents of unknown genotype are collected from a natural population, polymorphic at a single genetic locus. From each of these N cryptic parents a number of offspring are identified for their genotype. Our problem is to select an efficient offspring sampling plan for estimating the frequency of an allele in the cryptic adult population based on the N family profiles of juvenile genotypes. A criterion called the information per unit cost of observation is introduced to evaluate sequential sampling plans, in which the number of offspring per family examined is random. Some simple, practical schemes for stopping the sampling of offspring from a collected parent are introduced; one example is stopping when: (i) the offspring are definitive about the parental genotype(s) for the first time; (ii) a fixed number of one genotype only is seen; or (iii) a fixed maximum feasible number of offspring have been genotyped. This sampling scheme is recommended. For each sampling scheme, the best linear unbiased estimator and the sequential maximum likelihood estimator of the allele frequency are characterized. From the moments of these estimators, it is then possible to tabulate efficient sequential sampling plans, which are better (in the sense of information per unit cost), just as simple, and less costly than corresponding fixed sampling plans in use.

1. Introduction

Population geneticists routinely survey natural populations from a diverse assemblage of organisms to obtain an array of progeny genotypes from collected parents, but find it difficult or impossible to obtain genotypic data on the parents themselves. Statements about the structure of the cryptic parental population are made indirectly through the family profile available on each collected parent. Chromosomal polymorphisms in Drosophila (Dobzhansky and Powell, 1975; Fontdevila et al., 1982; Carson, 1983; Eiges, 1984; and Inoue, Watanabe, and Watanabe, 1984) are one class of relevant examples. Giant gene rearrangements are visible in the salivary chromosomes of Drosophila larvae but are cryptic in adult flies. A system of gene rearrangements along one chromosome is usually inherited in a simple Mendelian fashion like alternate alleles at a single genetic locus. The surveyor wishes to determine the probability, θ, that an adult’s chromosome carries a particular gene arrangement and whether that probability varies spatially or temporally.

Address correspondence to Prof. Jonathan Arnold, Genetics Department, University of Georgia, Athens, Georgia 30602, U.S.A.

Key words: Drosophila pseudoobscura; Familial data; Gene frequencies; Information per unit cost of observation; Levene estimator; Mixture models; Scoring procedure; Sequential design; Sequential estimators.
In designing surveys to estimate an allele probability or allele frequency, the investigator must have some rule specifying the number of offspring to examine per family, the offspring sampling plan. We examine some sampling plans in the context of two models for two basic protocols, one being an allelic assay from collected fathers (Arnold, 1981) and the other being a genotypic assay from collected mothers (Arnold and Morrison, 1985). The notation remains essentially unchanged from earlier papers.

In the male protocol the parameter $\theta$ or $\bar{\theta} = 1 - \theta$ on $[0, 1]$ will denote the probability that a chromosome carries the allele of interest or the other allele(s), respectively. Let $y = 2$, 1, or 0 denote the number of the allele(s) in a collected father, his genotype. If the father’s genotype $y$ is known to be 2, 1, or 0, then the probability that the allele is inherited from the father is $1, \frac{1}{2},$ or 0, respectively. The probability $Pr(y | \theta)$ of drawing a father of genotype $y$ is given by the Hardy–Weinberg law:

$$Pr(y | \theta) = \begin{cases} \theta^2, & y = 2, \\ 2\theta\bar{\theta}, & y = 1, \\ \bar{\theta}^2, & y = 0. \end{cases}$$

(1)

The probability that $x$ offspring inherit the allele from the father and $r - x$, the other allele, from the father will depend on the sampling scheme employed. If $x = (x, r - x)$ denotes a father’s family profile, then the probability of the family profile $Pr(x | \theta)$ will be a $\theta$-mixture:

$$Pr(x | \theta) = \sum_y Pr(x | y)Pr(y | \theta),$$

(2)

where the densities $\{Pr(x | y)\}$ are known and specified by the Mendelian laws and the sampling scheme. For example, if a fixed number of offspring $r$ are examined per father, then the probabilities $\{Pr(x | y)\}$ are binomial with parameters 0, $\frac{1}{2}$, or 1 and $r$. The families of $N$ fathers $x_1, \ldots, x_N$ are a simple random sample from the model specification $Pr(x | \theta)$.

In the female protocol (Arnold and Morrison, 1985), the collected mother has mated at random with a single unknown father in nature. [In some populations the female may mate with multiple fathers (Akin, Levene, Levine, and Rockwell, 1984), and the model below would need modification.] Neither parent’s genotype is known. The genotypes of the mother and father are denoted by $y_0$ and $y_1$. If the collector knew the parents to be $y = (y_0, y_1)$, then he could compute the probability that an offspring is homozygous for the allele, heterozygous, or homozygous for the other allele from the Mendelian laws. With both parents heterozygous $[y = (1, 1)]$, these probabilities are $\frac{1}{4}, \frac{1}{2}$, and $\frac{1}{4}$; with one parent being homozygous for the allele and the other, heterozygous, these probabilities are $\frac{3}{4}, \frac{1}{4},$ and 0. A full table of these $9 \times 3$ probabilities is given in Arnold and Morrison (1985, Table 1). In fact, the genotypes of the parents are cryptic, but then we can compute the probability of a mating pair $y$ from the Hardy–Weinberg law (1) and from the assumption of random mating with a single father:

$$Pr(y | \theta) = Pr(y_0 | \theta)Pr(y_1 | \theta).$$

(3)

The probability that among the offspring of one mother, $n_0$ offspring are homozygous for the other allele, $n_1$ offspring are heterozygous, and $n_2$ offspring are homozygous for the allele can be computed from the Mendelian laws and knowing the sampling scheme. If $n = (n_0, n_1, n_2)$ denotes a mother’s family profile, then the probability of the family profile $Pr(n | \theta)$ will be a $\theta$-mixture:

$$Pr(n | \theta) = \sum_y Pr(n | y)Pr(y | \theta).$$

(4)

The densities $\{Pr(n | y)\}$ are known and specified by the Mendelian laws and sampling scheme. For example, if $r$ offspring are examined from each mother, then the probabilities
\[ \text{Pr}(n | y) \] are trinomial. As an example,

\[ \text{Pr}(n | (1, 1)) = \binom{r}{n} \left( \frac{1}{4} \right)^{n_0} \left( \frac{1}{2} \right)^{n_1} \left( \frac{1}{4} \right)^{n_2}, \quad \text{with} \quad \binom{r}{n} = \frac{r!}{n_0!n_1!n_2!}. \]

The families of \( N \) mothers \( n_1, \ldots, n_N \) are a simple random sample from \( \text{Pr}(n | \theta) \).

One sampling scheme already mentioned is to examine a fixed number of offspring from each collected parent. Alternatives, which result in considerable reduction in cost (see Table 3) and gain in efficiency, are sequential sampling schemes. Such schemes allow the family profile (\( x \) or \( n \)), after genotyping the \( i \)th offspring, to determine whether to continue sampling. In sequential sampling schemes the family size varies randomly with each collected parent.

Sequential sampling plans have been considered in only three population genetics contexts. Haldane (1945) and Edwards (1960) suggested their use to estimate the frequency of rare alleles causing genetic disorders in human populations under a negative binomial sampling scheme. This problem takes on a new importance with the introduction of rare DNA markers for human genetic diseases such as Huntington’s chorea (Gusella et al., 1983). Clegg (unpublished Ph.D. dissertation, University of California, Davis, 1972) also suggested sequential sampling plans in the estimation of parental genotypic frequencies from “family profiles” on parental plants with unknown genotype. Clegg’s conclusion was that sequential sampling schemes would be cost-effective provided the bookkeeping problems could be surmounted. Cannings and Thompson (1977) have advocated the use of sequential sampling of pedigrees to infer the underlying mode of inheritance of an allele.

Sequential sampling plans are already used in population genetics. First, they receive clandestine use in the estimation of allele probabilities in cryptic parents. Some family profiles are definitive about the numbers of each allele in the cryptic parents. When we observe both alleles in a father’s family (\( x > 0, n - x > 0 \)) for the first time, we know the father is heterozygous (\( y = 1 \)). There is no further information in this family about the allele probability. The investigator stops examining offspring from the current father and moves to another. An estimator based on such a stopping rule is sequential. Second, Slatkin (1985) has recently shown how estimates of rare allele probabilities allow the estimation of gene flow in natural populations. Haldane (1945) pointed out that an inverse sampling scheme yields a more precise estimate of a rare allele probability than a fixed sampling scheme. The sequential estimators below are based on an inverse sampling scheme and thus could play an important role in obtaining estimates of gene flow in natural populations.

It is well known that sequential procedures offer improved efficiency in estimation and testing over fixed sampling plans and can lead to substantial costs savings in comparison to fixed sampling plans, as documented in Table 3. We now introduce four offspring sampling schemes describing when to stop sampling offspring from a collected parent.

\( \text{(S0):} \) The simplest sampling scheme is that of a fixed sampling plan, namely stop collecting familial data on a collected parent when a fixed number of offspring, \( r \), have been examined. The investigator will have a maximum feasible number of offspring which he can examine per family. The number \( n \) will be referred to as the maximum feasible family size, thus setting the range of \( r \) from 1 to \( n \). For the male protocol, sampling stops when a fixed number of alleles from the father have been examined. Although such a rule is simple, any procedure which stops sampling from a collected parent, when a score is definitive, will be more “efficient.” Once a score is definitive on a collected parent, there is no further information in a family profile about the allele probability. For the male protocol, a score is definitive the first time that both alleles from the father appear in the offspring (\( x \in A_1 \) in Table 1); for the female protocol, a score is definitive the first time both
homozygotes appear in the offspring (n ∈ Λ_{ij} in Table 1). All sequential schemes (S1)–(S3) below involve stopping the first time a score is definitive, and by so doing are uniformly more “efficient” than any fixed sampling scheme (S0).

(S1): For the female protocol, a simple sequential sampling scheme is one which leads to termination when a run of r homozygotes only is observed. For the male protocol, sampling ceases when a run r of one allele only is observed. This rule is one in which an examination of further offspring from a collected parent ceases when: (i) a score is definitive for the first time; (ii) there is r of one homozygote (allele) only; or (iii) the maximum feasible number of offspring n have been examined. If a run of heterozygotes is observed, sampling continues until the nth offspring has been examined.

(S2): There may be diminishing returns in the amount of “information” from a run of heterozygotes. A simpler sampling scheme than (S1) is to stop whenever a run r of one genotype only is seen. For the male protocol, sampling scheme (S2) does not arise. From sampling scheme (S2), sampling offspring stops when: (i) a score is definitive for the first time; (ii) there is a run r of one genotype only; or (iii) the maximum feasible number of offspring n have been examined. This rule can lead to early termination with a run of r heterozygotes.

(S3): The last sampling scheme represents a compromise between a scheme which stops based on runs (S1) of length r and a scheme which allows gaps between the r homozygotes. This scheme does not arise under the male protocol. In fact, one might suggest a scheme which involves stopping the first time there are r of either homozygote, but the gain in “efficiency” by allowing gaps can be more simply addressed by stopping rule (S3). Sampling scheme (S3) is one in which an examination of further offspring ceases when: (i) the score is definitive for the first time; (ii) the first time there are r homozygous offspring in the first r + 2 offspring examined; or (iii) the maximum feasible number of offspring n have been examined. Under sampling scheme (S3), there could be up to 2 heterozygotes interspersed with the r individuals homozygous for one allele [e.g., n = (r, 2, 0) or (0, 2, r)] when sampling ceases. Thus, this scheme can lead to early termination of sampling without a run of one homozygote. By examining this scheme we will be able to settle whether there is much more “information” in a family profile beyond the runs observed. As a note, consider a scheme which involves stopping the first time there are r homozygous offspring in the first r + k offspring. Calculations under (S3) in Tables 1 and 2 for k = 2 are typical of and are the simplest among the r + k schemes, while similar calculations for k = 1 are not typical.

Once one of these sampling schemes is chosen, almost all the information about the allele probability is summarized by a score statistic, which counts families according to the presence or absence of alleles (or genotypes) seen in a family profile x (or n). The score statistic for the male protocol is a list of counts N = (N_{00}, N_{01}, N_{2}) of N families with only the other allele (N_{0}), at least one of each allele (N_{1}), or only the allele (N_{2}). The score statistic for the female protocol is a list of 6 counts N = (N_{00}, \ldots, N_{11}) of N families with only offspring homozygous for the other allele (N_{00}), \ldots, or at least one offspring homozygous for each allele (N_{11}). The subscripts denote an inference on the number of the allele, which the parent(s) carry.

The allele probability estimates advocated for the male and female protocols are

\[ \hat{\theta} = (N_{2} + \frac{1}{2}N_{1})/N \]

and

\[ \hat{\theta} = (N_{22} + \frac{3}{4}N_{12} + \frac{1}{2}N_{11} + \frac{1}{2}N_{02} + \frac{1}{4}N_{01})/N. \] (5)
Either estimator is known as the *Dobzhansky estimator* under the fixed offspring sampling scheme (S0) and will be termed the *Levene estimator* under the sequential sampling scheme (S2) [(S1)] under the female (male) protocol in honor of Howard Levene's long-standing contribution to the population genetics of *Drosophila* (Dobzhansky and Levene, 1948). We recommend the use of the Levene estimator.

In addition to being more efficient than fixed sampling plans, the sequential plans above are just as simple as a fixed plan in (S0), provided there is no problem in genotyping offspring one at a time (Armitage, Stratton, and Worthington, 1985). As with a fixed sampling scheme, there is only one number to remember as offspring are genotyped, that is, where one is relative to the *stopping constant r*. The sequential estimates are standard [e.g. (5)], and tables and a program are available for selecting an efficient offspring sampling plan, as specified by the stopping constant r.

2. Inference

When these two models are adequate, as in Dobzhansky and Levene (1948), Anderson and Mcguire (1978), and Anderson et al. (1979), almost all the information is contained in the score statistic N for the male or female protocol. For the male protocol, the statistic $N = (N_0, N_1, N_2)$ counts the number of each type of score. This statistic is sufficient, as has been shown for sampling scheme (S0) (Arnold, 1981, Result 2). For the female protocol, the score statistic $N = (N_{00}, \ldots, N_{11})$ is nearly sufficient (Arnold and Morrison, 1985, p. 794), where, for example, $N_{00}$ is the number of family profiles inferred to have resulted from a cross of two parents homozygous for the other allele. Both score statistics have a multinomial distribution. For the male protocol the score statistic is trinomial:

$$\Pr(N | \theta) = \left(\frac{N!}{N_0!N_1!N_2!}\right)K_0^{N_0}K_1^{N_1}K_2^{N_2},$$

(6.M)

where $$(N) = N!/(N_0!N_1!N_2!)$$. For the female protocol, the score statistic is multinomial:

$$\Pr(N | \theta) = \left(\frac{N!}{N_{00}!N_{01}!N_{02}!N_{10}!N_{11}!N_{12}!}\right)K_{00}^{N_{00}}K_{01}^{N_{01}}K_{02}^{N_{02}}K_{10}^{N_{10}}K_{11}^{N_{11}}K_{12}^{N_{12}},$$

(6.F)

where $$(N)$$ is a multinomial coefficient. The multinomial cell probabilities $$K$$ are quadratic (quartic) polynomials in $$\theta$$ in the male (female) protocol with constants determined by the sampling schemes used to terminate sampling from a family (Table 1). The constants $$g$$, $$\ldots$$, $$p$$, determining the score probabilities $$K$$, vary with the sampling scheme (S1) and depend only on the stopping constant $$r$$ and the maximum feasible family size $$n$$.

Linear unbiased estimators of the allele probability $$\theta$$ have been advocated by Haldane (1945), Arnold (1981), and Arnold and Morrison (1985) for their simplicity, efficiency (for $$\theta \in [0.01, 0.99]$$), and heuristic appeal. Cotterman (1941, 1947, 1954) has recommended similar estimators. The sequential estimators below share all these properties. For each protocol the estimator advocated is a linear combination of the counts in the score statistic N:

$$\hat{\theta} = (N_2 + w_1N_1)/N$$

(7.M)

or

$$\hat{\theta} = (N_{22} + w_{12}N_{12} + w_{11}N_{11} + w_{02}N_{02} + w_{01}N_{01})/N.$$  

(7.F)

The weights $$w$$ in (7) were particularly simple in the fixed sampling plans as given in (5). When the familial data on each parent are collected sequentially, then the weights must be changed in order to keep the sequential estimator $$\hat{\theta}$$ in (7), unbiased. The class of all linear, sequential unbiased estimators based on (7) can be fully characterized in Result 1 for both protocols.
### Table 1

Probabilities of scores under varied sampling schemes

<table>
<thead>
<tr>
<th>Event</th>
<th>Probability</th>
<th>Constant</th>
<th>(S0)</th>
<th>(S1)</th>
<th>(S2)</th>
<th>(S3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male protocol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A_0 = {x: x = 0})</td>
<td>(K_0 = 3\bar{\theta} + g\bar{\theta})</td>
<td>(g = 1 - \bar{g})</td>
<td>((\frac{1}{3})^{n-1} = a)</td>
<td>((\frac{1}{3})^{n-1})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A_1 = {x: 0 &lt; x &lt; n})</td>
<td>(K_1 = 2\bar{\theta})</td>
<td>(g = )</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>(A_2 = {x: x = r})</td>
<td>(K_2 = \theta^2 + g\bar{\theta})</td>
<td>(g = )</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>Female protocol</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(A_{00} = {n: n = (n_0, 0, 0)})</td>
<td>(K_{00} = K_0)</td>
<td>(g = 1 - \bar{g})</td>
<td>((\frac{1}{3})^{n-1} = a)</td>
<td>((\frac{1}{3})^{n-1})</td>
<td>((\frac{1}{3})^{n-1})</td>
<td>()</td>
</tr>
<tr>
<td>(A_{01} = {n: n = (0, n_1, 0)})</td>
<td>(K_{01} = 2\bar{\theta}(p + \theta\bar{\theta}))</td>
<td>(p = )</td>
<td>((\frac{1}{3})^{n-1})</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>(A_{02} = {n: n = (0, 0, n_2)})</td>
<td>(K_{02} = K_0)</td>
<td>(g = )</td>
<td>((\frac{1}{3})^{n-1})</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>(A_{01} = {n: n = (n_0, n_1, 0)})</td>
<td>(K_{01} = 4\bar{\theta}^2(b\bar{\theta} + h\bar{\theta}))</td>
<td>(b = )</td>
<td>(1 - (\frac{1}{3})^{n-1} = a)</td>
<td>(1 - (\frac{1}{3})^{n-1})</td>
<td>(1 - (\frac{1}{3})^{n-1})</td>
<td>(1 - (\frac{1}{3})^{n-1})</td>
</tr>
<tr>
<td>(A_{02} = {n: n = (0, n_1, 0)})</td>
<td>(K_{02} = K_0)</td>
<td>(h = )</td>
<td>()</td>
<td>()</td>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>(A_{01} = {n: n = (n_0, n_1, 0)})</td>
<td>(K_{01} = 4\bar{\theta}^2(b\bar{\theta} + h\bar{\theta}))</td>
<td>(f = )</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
</tr>
<tr>
<td>(A_{02} = {n: n = (0, n_1, 0)})</td>
<td></td>
<td></td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
</tr>
<tr>
<td>(A_{01} = {n: n = (n_0, n_1, 0)})</td>
<td></td>
<td></td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
<td>(1 - 2(\frac{1}{3})^n)</td>
</tr>
</tbody>
</table>

\* The constants \(a, \bar{a}, b, c\) in Results 2 of Arnold (1981) and Arnold and Morrison (1985, Table 2) are listed in boldface type under (S0).

\* For the (S3) rule,

\[
h = \left(\frac{3}{4}\right)^n - \left(\frac{1}{2}\right)^n - \left(\frac{1}{2}\right)\left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1} - \left(\frac{1}{4}\right)^{n-1}.
\]
<table>
<thead>
<tr>
<th>Sampling scheme</th>
<th>$\eta_{1n}$</th>
<th>Var($\hat{\theta}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Male protocol</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(S0) $r$</td>
<td>$\frac{\theta \bar{\theta}}{2N} (1 + g)$</td>
<td></td>
</tr>
<tr>
<td>(S1) $r(\theta^2 + \bar{\theta}^2) + 2\theta \bar{\theta} (1 + 2g)$</td>
<td>$\frac{\theta \bar{\theta}}{2N} (1 + g)$</td>
<td>(Levene)</td>
</tr>
<tr>
<td><strong>Female protocol</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(S0) $r$</td>
<td>$\frac{\theta \bar{\theta}}{4N} \left[ 1 + g - 2\theta \bar{\theta} [\hat{\theta} + 2f + 3(h - b)] \right]$</td>
<td></td>
</tr>
<tr>
<td>(S1) $r(\theta^4 + \bar{\theta}^4) + 2\theta \bar{\theta} (\theta^2 + \bar{\theta}^2) [2n + (r - n)g] + (\theta \bar{\theta})^2 [24 + 2n + 4a - 32n(\hat{\theta})^n + (\hat{\theta})^r - (\hat{\theta})^2(n + r)]$</td>
<td>$\frac{\theta \bar{\theta}}{4N} \left{ \frac{\hat{\theta}^2}{b} + 2a - 2\theta \bar{\theta} [\hat{\theta} + w_{11} (1 - w_{11}) + f + 16\theta \bar{\theta} (h - b) w_{01} (1 - w_{01})] \right}$</td>
<td>(Levene)</td>
</tr>
<tr>
<td>(S2) $r(\theta^4 + \bar{\theta}^4) + 4\theta \bar{\theta} (\theta^2 + \bar{\theta}^2) [6n + rg]$</td>
<td></td>
<td>Same as expression (S0)</td>
</tr>
<tr>
<td></td>
<td>$+ (\theta \bar{\theta})^2 [24 + 2r + 4g - 32n(h + (\hat{\theta})^r + \hat{\theta}^r)]$</td>
<td>(Levene)</td>
</tr>
<tr>
<td>(S3)* $r(\theta^4 + \bar{\theta}^4) + 2\theta \bar{\theta} (\theta^2 + \bar{\theta}^2) [an + rg + r(r + 1)(\hat{\theta})^r + \hat{\theta} + (r + 1)(\hat{\theta})^r + 2n(\hat{\theta} - 3r(\hat{\theta})^r - r(r + 1)(\hat{\theta})^r]$</td>
<td></td>
<td>Same as expression (S1)</td>
</tr>
<tr>
<td></td>
<td>$+ (\theta \bar{\theta})^2 [24 + 2n + 4a - 4r(r + 5)(\hat{\theta})^r]$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$- 32n(h + (\hat{\theta})^r + (\hat{\theta})^r - (\hat{\theta})^r - r(r + 1)(\hat{\theta})^r/8)]$</td>
<td></td>
</tr>
</tbody>
</table>

*The constants $g, \ldots, p$ vary with each sampling scheme. For the expectation $\eta_{1n}^{(2)}$ to hold, one must have $n \geq r + 2$. 

51
Result 1. For (7.M) to be unbiased under the male protocol and \( r > 1 \), we must have \( w_1 = \frac{1}{2} \). For (7.F) to be unbiased under the female protocol and \( r > 1 \), we must have:

\[
\begin{align*}
    w_{12} &= (3 - 2g - 2pw_{02})/(4b); \\
    w_{11} &= (3 - g^2)b + 2w_{02}(2ph - (1 + p)b) - 2h(2 - g))/(4fb); \\
    w_{01} &= (1 - 2pw_{02})/(4b)
\end{align*}
\]

with the constants \( g, \ldots, p \) given in Table 1.

In the male protocol there is one and only one linear, unbiased sequential estimator with weight \( w_1 = \frac{1}{2} \) as in fixed samples with the Dobzhansky estimator \((N_1 + \frac{1}{2}N_0)/N\) under sampling scheme (S0). If \( r = 1 \), then the weight \( w_1 \) is irrelevant. In the female protocol there is a whole class of linear sequential unbiased estimators indexed by the weight \( w_{02} \). We restrict the weight \( w_{02} \in [0, 1] \) so that \( \hat{\theta} \) in (7) is in [0, 1]. Under the condition \( w_{02} \in [0, 1] \), the remaining weights lie in [0, 1]. If \( r = 1 \), then the weights \( w_{12}, w_{11}, \) and \( w_{01} \) are irrelevant.

Having characterized the class of linear sequential unbiased estimators, it is then possible to pull out the best such estimator. An estimator (7) minimizing the mean square error MSE(\( \hat{\theta} \)) for a given sampling scheme among unbiased procedures based on a score statistic N can be found and is termed the BLUS estimator.

There is a more general criterion that will allow us to compare BLUS estimators from different sampling schemes (S1) with each other and with other estimators, such as the sequential maximum likelihood estimators \( \hat{\theta}_{ML} \) from each sampling scheme. The criterion is called the information per unit cost of observation \( i_\theta(\theta) \) and is defined asymptotically (large N) as

\[
i_\theta(\theta) = \lim_{N \to \infty} [\text{MSE}(\hat{\theta})E(m)N]^{-1},
\]

where \( E(m) \) denotes the expected family size under the sampling scheme in question. This quantity is motivated in the next section. The asymptotic efficiency of an estimator \( \hat{\theta} \) is measured by the ratio of the information per unit cost of observation of the estimator \( \hat{\theta} \) to that of our sequential maximum likelihood estimator \( \hat{\theta}_{ML} \) in large samples, namely

\[
i_\theta(\theta)/i_{\theta ML}(\theta).
\]

When two estimators within the same sampling scheme are being compared, this measure reduces to the usual notion of asymptotic efficiency. One can think of the BLUS estimator as the one which maximizes the information per unit cost of observation in the class of linear unbiased estimators (7), since the expected family size \( E(m) \) is independent of the weights \( \{w\} \).

Result 2. For the male protocol and either sampling scheme, estimator (7.M) with \( w_1 = \frac{1}{2} \) is the BLUS estimator. For the female protocol, estimator (7.F) with the weights in Result 1 and \( w_{02} = \frac{1}{2} \) is the BLUS estimator, provided a sampling scheme leads to (6).

Sampling schemes that do not necessarily lead to score probabilities structured as in Table 1 include asymmetrical sampling schemes in which the stopping constant \( r \) depends on the genotype involved. Such schemes may become important in estimating a rare allele probability of a DNA marker for a genetic disease. In particular, Result 2 implies that the Dobzhansky estimator in the previous papers is the best linear unbiased estimator among fixed sampling plans. Furthermore, the BLUS alternatives to the Dobzhansky estimator based on the constants in columns (S1) through (S3) are uniformly more “efficient” than the Dobzhansky estimator (S0) as measured by the information per unit cost of observation.
For the male protocol, the variance of the BLUS estimator (7.14) is found in Table 2. For the female protocol, the variance of any linear sequential estimator (7.15) with weights selected to make it unbiased is given by equation (A.2) in the Appendix. The variances (A.2) of the BLUS estimators for the female protocol simplify for the four sampling schemes in Table 1 because \( w_{02} = \frac{1}{2} \). The variances of the BLUS estimators are shown in Table 2. For example, the Levene estimators in equation (5) have their variances listed in Table 2 based on the constants under column (S2) or (S1) in Table 1.

The BLUS estimators can be compared with the corresponding [same sampling scheme (S1) and constants] sequential maximum likelihood estimator obtained by maximizing the density in Result 2 (Arnold, 1981) or Result 2A (Arnold and Morrison, 1985), using the new constants in Table 1. Results 1–6 in these previous papers go through for each sampling scheme, but are not repeated for the sake of space. It suffices to say that the BLUS estimator (e.g., \( w_1 = \frac{1}{2} \) or \( w_{02} = \frac{1}{2} \)) is nearly fully efficient (\( \geq 0.90 \)) for the allele probability \( \theta \in [0.01, 0.99] \) relative to the sequential maximum likelihood estimator by comparing the BLUS variance in Table 2 with its Cramer–Rao lower bound. Alternatively, for the female protocol, the sequential grouped profiles estimator (Arnold and Morrison, 1985, p. 792) obtained by maximizing density (6) is fully efficient relative to the maximum likelihood estimator to two decimal places for \( \theta \in [0.01, 0.99] \), but easier to compute than the sequential maximum likelihood estimator. A completely analogous estimator has been advocated by DeGroot and Li (1960).

3. Sequential Design

In designing surveys to estimate an allele probability \( \theta \), the investigator must decide on the number of offspring to identify from each parent of \( N \) such collected parents. Sequential sampling plans, in which the number of offspring per family is random, have been introduced. Schemes (S0)–(S3) describe when to stop sampling a family. It is well known that certain sequential procedures for estimation and testing are asymptotically optimal (and certainly better than fixed sampling plans) for a broad array of statistical problems (Lai, 1981; Chow and Yu, 1981). The sequential probability ratio test minimizes the expected number of observations under fairly general conditions (Lai, 1981). A sequential estimator of the mean due to Anscombe (1952) minimizes the asymptotic mean square error plus a linear cost of observation under fairly general conditions (Chow and Yu, 1981). Sequential procedures are well known to be cost-effective (Wald, 1947, p. 54 or Table 3).

In the protocols in Arnold (1981) and Arnold and Morrison (1985), practical considerations usually rule out the use of a fully sequential procedure, in which both collection size \( N \) and family size \( m \) are allowed to be random. We focus on sequential plans in which family size \( m \) is random, but collection size \( N \) is large and fixed. In this special setting there are a number of additional reasons that sequential schemes (S1)–(S3) are advantageous. An examination of the sample space \( \Lambda = \{x\} \) or \( \{n\} \) under a plan in the sequential schemes (S1)–(S3) reveals that each sequential plan corresponds to a subset of the sample space of a fixed plan in (S0) with the same stopping constant \( r \). This fact implies that the sufficient statistics for each sequential offspring sampling plan are of lower dimensionality than those of the corresponding fixed sampling plan, meaning the geneticist need follow fewer events. The sequential procedures considered (S1)–(S3) are also just as simple as the fixed sampling plans (S0), and the geneticists clandestinely use one aspect of them already.

Several rules to stop sampling a given family have been proposed on grounds of their simplicity and ease of implementation in a nonautomated setting. Each sampling scheme involves waiting until \( r \) of some genotype are seen. In designing surveys to estimate an allele probability \( \theta \), the investigator must select a sampling scheme (S0)–(S3) and then a stopping constant \( r = 1, 2, \ldots, \), or \( m \) in order to fully specify the offspring sampling plan.
A number of authors (Kempthorne, 1969, p. 184; Brown, Weir, and Marshall, 1970; Brown, 1975; and Morris and Spieth, 1978) either explicitly or implicitly use a criterion called the information per unit cost of observation to select an efficient sampling plan \( r \). If an investigator had the resources to make parental collection size \( N \) very large, then he would wish to choose an offspring sampling plan \( r \), that provides the most information at the least cost. We need some measure of the information in an estimator \( \hat{\theta} \), a measure which is larger when the estimator has more precision. One such measure is the inverse of the asymptotic mean square error (large \( N \)) of the estimator \( \hat{\theta} \), namely \( \text{MSE}^{-1}(\hat{\theta}) \). In dividing the information \( \text{MSE}^{-1}(\hat{\theta}) \) by the sampling plan’s expected total cost \( E(m)N \), we obtain the information per unit cost of observation \( \text{MSE}(\hat{\theta})E(m)N^{-1} \). We wish to choose an offspring sampling plan \( r \) with a large information per unit cost of observation. This criterion credits the investigator for increased precision in his estimate, but penalizes him for an increase in the relative cost \( E(m) \) of sampling to achieve the same desired precision. Since the information \( I_m(\theta) = \text{MSE}^{-1}(\hat{\theta}) \) is proportional to the collection size \( N \), the information per unit cost of observation is independent of collection size and thus is an ideal asymptotic criterion for selecting an offspring sampling plan \( r \) (Govindaraju, 1975, p. 321; Federgruen, Hordijk, and Tijms, 1979).

There are also finite sampling justifications leading to the same criterion. Let us suppose that an investigator has a fixed amount of resources \( T = E(m)N \) available to him and wishes to choose an offspring sampling plan \( r \) with small mean square error \( \text{MSE}(\hat{\theta}) \). Since this criterion is proportional to \( N^{-1} \), we can write it as a product of a quantity independent of \( N \) and of \( N^{-1} \):

\[
\text{MSE}(\hat{\theta})N \times [N^{-1}] = [I_m^{-1}(\theta)N] \times [E(m)/T],
\]

the last quantity being proportional to \( \text{MSE}(\hat{\theta})E(m)N \). To make the quantity \( \text{MSE}(\hat{\theta})E(m)N \) small is tantamount to making the information per unit cost of observation large. In the description below we will couch our discussion in terms of the information per unit cost of observation because of its simplicity and wider applicability. This criterion also yields a Cramer–Rao lower bound for sequential estimators (Govindaraju, 1975, p. 321).

An efficient sampling plan involves selecting a stopping constant \( r^* \) to maximize the information per unit cost of observation in large samples (\( N \) large). The current wisdom (Morris and Spieth, 1978) is that efficient sampling plans to estimate an allele probability require \( r = 1 \), one offspring per collected adult. Generally, the information per unit cost of observation will depend on the unknown parameter \( \theta \), as well as the stopping constant \( r \). In selecting a sampling plan it is necessary to study the criterion \( \text{MSE}(\hat{\theta})E(m)N^{-1} \) over the parameter space as well as for variation in the sampling plan \( r \). By using a preliminary estimate of the allele probability \( \theta \) or by assuming a least favorable state of nature (e.g., \( \theta = \frac{1}{2} \)), we can select an efficient sampling plan.

There is another cost ignored by previous analyses—the set-up cost per collected adult. When a collection yields few parents, the investigator has invested considerable time per parent in collecting adults and setting up cultures for offspring, and the limited information available from small collections is at a premium. He then wishes to examine more offspring per collected parent. We introduce a unit set-up cost \( s \) to be added to the unit cost of observation \( E(m) \). The information per unit cost of observation then becomes

\[
\text{MSE}(\hat{\theta})E[(s + m)N]^{-1}.
\]

For the examples we shall use, the units of \( s \) are the number of offspring per family that could be genotyped in the time taken to make the parental collection and set up offspring cultures in the laboratory.
An efficient sequential sampling plan involves selecting the stopping constant $r^*$, which maximizes the information per unit cost of observation. To find such plans we must: (i) calculate the expected number of offspring per family $E(m)$ for each sampling scheme (Si); (ii) compare sampling schemes via the criterion $[\text{MSE}(\theta)E(m + s)N]^{-1}$; and (iii) select an efficient sampling plan $r^*$ from within a sampling scheme (Si).

The expected number of offspring examined or relative cost of observation under sampling scheme (Si) will be denoted by $n_{r,n}^{0}$. Relative cost increases with the stopping constant $r$. The expected number of offspring per family $n_{r,n}^{0}$ under the various sampling schemes are found in Table 2. Its maximum value is $n$, the maximum feasible family size, and its minimum value, $r$. For a fixed scheme (S0), we have $n_{r,n}^{0} = r$, a constant; for sequential schemes the expected number of offspring $n_{r,n}^{0}$ will depend on the allele probability $\theta$, as summarized in Result 3.

Result 3. Under sampling schemes (Si) above the expected number of offspring examined $n_{r,n}^{0}$ is a concave function of $\theta \in [0, 1]$, symmetric about a maximum ($\leq n$) at $\theta = \frac{1}{2}$, and has a minimum of $r$ at $\theta = 0$ or 1.

The relative cost of sampling for the sequential procedures increases (Figure 1) as the population becomes more heterozygous (i.e., as $2\theta \theta \rightarrow \frac{1}{2}$). The relative cost of sampling has its smallest value $r$ when the population is monomorphic. The symmetry of the expected relative cost is a consequence of the symmetry in the sampling schemes.

The savings in numbers of offspring examined per family can be substantial (Table 3). The savings can be computed by comparing an efficient sampling plan $r = r^*$ with the plan that stops only for definitive scores ($r = n$). The efficient plan $r^*$ for a given sampling

![Figure 1. The information per unit cost of observation (---), the expected number of offspring per family (-----), and the efficient sequential plans $r^*$ (------) for stopping rule (S1) are graphed as a function of allele probability. The set-up cost is zero and the maximum family size, 5. The vertical scale is 10 times each of these quantities and the horizontal scale, percent.](image-url)
Table 3

<table>
<thead>
<tr>
<th>n</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>2</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>3</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>4</td>
<td>.8 ( .8)</td>
<td>.6 ( .7)</td>
<td>.6 (0)</td>
<td>.5 (0)</td>
<td>.5 (0)</td>
</tr>
<tr>
<td>5</td>
<td>1.5 (1.6)</td>
<td>1.2 (1.4)</td>
<td>1.1 (0.6)</td>
<td>1.0 (.5)</td>
<td>1.0 (.5)</td>
</tr>
<tr>
<td>6</td>
<td>2.3 (2.5)</td>
<td>5.1 (2.0)</td>
<td>5.2 (1.2)</td>
<td>5.2 (1.0)</td>
<td>5.2 (1.0)</td>
</tr>
<tr>
<td>7</td>
<td>3.0 (3.3)</td>
<td>6.0 (2.7)</td>
<td>6.0 (1.7)</td>
<td>6.0 (1.6)</td>
<td>6.0 (1.5)</td>
</tr>
<tr>
<td>8</td>
<td>7.0 (4.1)</td>
<td>6.9 (3.4)</td>
<td>6.8 (2.3)</td>
<td>6.8 (2.1)</td>
<td>6.7 (2.0)</td>
</tr>
<tr>
<td>9</td>
<td>7.9 (4.9)</td>
<td>7.8 (4.1)</td>
<td>7.6 (2.9)</td>
<td>7.5 (2.6)</td>
<td>7.5 (2.5)</td>
</tr>
<tr>
<td>10</td>
<td>8.9 (5.7)</td>
<td>8.7 (4.8)</td>
<td>8.5 (3.5)</td>
<td>8.3 (3.1)</td>
<td>8.3 (3.0)</td>
</tr>
<tr>
<td>11</td>
<td>9.9 (6.6)</td>
<td>9.6 (5.4)</td>
<td>9.3 (4.1)</td>
<td>9.1 (3.6)</td>
<td>9.0 (3.5)</td>
</tr>
<tr>
<td>12</td>
<td>10.8 (7.4)</td>
<td>10.5 (6.1)</td>
<td>10.1 (4.6)</td>
<td>9.9 (4.2)</td>
<td>9.8 (4.0)</td>
</tr>
</tbody>
</table>

The savings for the male protocol are in parentheses. The set-up cost used is 3.8 larvae. Some of the optimal stopping constants \( r^* \) for each pair \((n, \theta)\) can be looked up in Table 4.

The savings \((r^* \text{ maximizing the information per unit cost of observation, namely } [\eta_{r,n} N \text{ var}(\theta)]^{-1})\) assuming zero set-up cost. We choose to compare the efficient sampling plan \( r = r^* \) with \( r = n \), rather than with the fixed plan, because any rational sampling plan involves stopping for definitive scores. From Table 2 the relative cost under stopping rule \((S2)\) is

\[
\eta_{r,n}^{(2)} = r(\theta^4 + \bar{\theta}^4) + 2\bar{\theta}(\theta^2 + \bar{\theta}^2)(\bar{g}n + rg) + (\bar{\theta})^2[2r + 24 + 4g - 32(h + (\frac{1}{n})' + (\frac{1}{n})')]\]

Setting \( r = n \) in this expression and subtracting the result from \( \eta_{r,n}^{(2)} \) yields the savings in numbers of offspring examined per family:

\[
(r - n)(\theta^4 + \bar{\theta}^4) + 2\bar{\theta}(\theta^2 + \bar{\theta}^2)(\bar{g}n + rg - n) + (\bar{\theta})^2[2(r - n) + 4(g - a) - 32(h + (\frac{1}{n})' + (\frac{1}{n})' - (\frac{1}{n})')]\]

This savings can be substantial, as shown in Table 3.

The second problem is choosing an efficient sampling scheme, and the third, choosing the most efficient plan \( r^* \) among plans \( r = 1, \ldots, n \). Sampling schemes can be cross-compared by their relative efficiency at fixed \( r \), and then the information per unit cost of observation can be used to select the efficient sequential plan \( r^* \). In order to make these comparisons, we summarize in Result 4 some properties of the information per unit cost of observation.

**Result 4.** For each sampling scheme \((S_i)\) and protocol, the information per unit cost of observation is a convex function for \( \theta \in (0, 1) \) and is symmetric about its minimum of \( \theta = \frac{1}{2} \). For a given \( r \) and for each protocol, the Levene estimator is more efficient than the other sampling schemes \((S_i)\) in Table 1 for intermediate values of \( \theta \) indicated below and \( n < 25 \).

Result 4 holds true for the BLUS, sequential maximum likelihood, and sequential grouped profiles estimators. As the heterozygosity of the population \((i.e., 2\bar{\theta})\) increases, there is less information about the allele probability \( \theta \) (Figure 1). As indicated in Figure 1, monomorphic populations provide more information about the allele probability with the fewest number of offspring examined per family on average.
Figure 2. The information per unit cost of observation for stopping rules S1 (-----), S2 (-----), and S3 (· · · · · ·) are plotted as a function of allele probability. The graphs are for three distinct sampling plans: (left) \( s = 0, r = 3, n = 5 \); (middle) \( s = 0, r = 3, n = 10 \); (right) \( s = 0, r = 3, n = 35 \). The vertical scale is 10 times the information and the horizontal scale, percent.

There are only two distinguishable sampling schemes, (S0) and (S1), corresponding to the Dobzhansky and Levene estimators for the male protocol. The (S1) sequential sampling scheme is uniformly more efficient in \( r \) and \( \theta \in (0, 1) \) than the fixed sampling scheme (S0). In the left panel of Figure 2, the efficiency of three sequential sampling schemes for the female protocol are graphed as a function of the allele probability \( \theta \) with the sampling plan fixed at \( (r = 3, n = 5, s = 0) \). We see that for a range of intermediate values \([.23, .77]\) that the Levene estimator (S2) outperforms the other two BLUS estimators based on (S1) and (S3), but if the allele probability is outside this interval, then the estimator (S3) beats estimator (S2). Estimator (S3) is uniformly more efficient than (S1) in that they both have the same variance and \( \eta_{r,n}^{(3)} \leq \eta_{r,n}^{(1)} \). As the maximum feasible family size \( n \) is increased, the range over which scheme (S2) is more efficient shrinks (middle panel) and eventually disappears (right panel). Provided values of \( (r, n) \) are chosen so that (S1)–(S3) do not coincide for all \( \theta \), the estimators based on (S1) and (S3) eventually beat the Levene estimator (S2) as the allele probability approaches fixation or loss. (This latter fact is important in estimating rare allele probabilities to measure gene flow.) The behavior of the sampling scheme (S3) suggests that there can be substantial information in the counts of genotypes in a family profile beyond the runs observed. In fact, the behavior of (S3) would suggest that a scheme that stops sampling when \( r \) of a genotype are seen should uniformly outperform schemes (S0)–(S3). The expected number of offspring from such a scheme may be computable (Olkin and Sobel, 1965). Against the increased efficiency of plans such as (S3), we must weigh the simplicity of the Levene estimator. On grounds of simplicity and efficiency, the Levene estimator and its associated sampling scheme (S2) are advocated.

In choosing an efficient plan \( r^* \), there is a trade-off between information gained and cost incurred. The information per family is to first order, proportional to \( \frac{\theta(1 + g)}{\theta(1 + g)} \) by Table 2. This quantity increases as the stopping constant \( r \) increases. On the other hand, the expected number of offspring per family for the Levene estimator also to first order, increases in \( r \). This can be seen by letting the maximum feasible family size \( n \) get large in \( \eta_{r,n}^{(1)} \) from Table 2:

\[
r(\theta^* + \bar{\theta}^*) + 2\theta\bar{\theta}(\theta^2 + \bar{\theta}^2)(\bar{n} + r\bar{g}) + (\theta\bar{\theta})^2[24 + 2r + 4g - 32(\frac{3}{4})r - 32(\frac{3}{4})^\prime].
\]

In order to design sample surveys, the highly efficient sampling plans based on the Levene estimator (S2) in (7) have been tabulated (Table 4). Three quantities are needed to enter Table 4 for the male (female) protocol to extract the efficient sequential sampling plan \( r^* \). The investigator must specify the set-up cost per family and the maximum amount of work that he is willing to invest per family, i.e., the maximum feasible family size \( n \). He
Table 4
Most efficient sampling plans for the male and female protocols

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Male</td>
<td></td>
<td></td>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10.90</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>3.</td>
<td>3.</td>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>3.</td>
<td>3.</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.9</td>
<td>2.9</td>
<td>2.9</td>
<td>2.9</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.2</td>
<td>3.5</td>
</tr>
<tr>
<td>0.0990</td>
<td>0.0675</td>
<td>0.0562</td>
<td>0.0562</td>
<td>0.0562</td>
<td></td>
<td>0.0450</td>
<td>0.0316</td>
<td>0.0272</td>
<td>0.0270</td>
<td>0.0268</td>
<td></td>
</tr>
<tr>
<td>20.80</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>3.</td>
<td>3.</td>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>3.</td>
<td>3.</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.4</td>
<td>3.7</td>
</tr>
<tr>
<td>0.1600</td>
<td>0.1200</td>
<td>0.1000</td>
<td>0.1000</td>
<td>0.1000</td>
<td></td>
<td>0.0800</td>
<td>0.0581</td>
<td>0.0499</td>
<td>0.0492</td>
<td>0.0486</td>
<td></td>
</tr>
<tr>
<td>30.70</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>4.</td>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>3.</td>
<td>3.</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.8</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.4</td>
<td>3.9</td>
</tr>
<tr>
<td>0.2100</td>
<td>0.1575</td>
<td>0.1312</td>
<td>0.1181</td>
<td>0.1181</td>
<td></td>
<td>0.1050</td>
<td>0.0777</td>
<td>0.0669</td>
<td>0.0655</td>
<td>0.0645</td>
<td></td>
</tr>
<tr>
<td>38.62</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>4.</td>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>3.</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.8</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>0.2356</td>
<td>0.1767</td>
<td>0.1472</td>
<td>0.1325</td>
<td>0.1325</td>
<td></td>
<td>0.1178</td>
<td>0.0879</td>
<td>0.0758</td>
<td>0.0697</td>
<td>0.0726</td>
<td></td>
</tr>
<tr>
<td>40.60</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>4.</td>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>3.</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.8</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.9</td>
<td>3.9</td>
</tr>
<tr>
<td>0.2400</td>
<td>0.1800</td>
<td>0.1500</td>
<td>0.1350</td>
<td>0.1350</td>
<td></td>
<td>0.1200</td>
<td>0.0897</td>
<td>0.0773</td>
<td>0.0711</td>
<td>0.0740</td>
<td></td>
</tr>
<tr>
<td>50.50</td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>4.</td>
<td></td>
<td>1.</td>
<td>2.</td>
<td>3.</td>
<td>4.</td>
<td>3.</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>2.8</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.9</td>
<td>4.0</td>
</tr>
<tr>
<td>0.2500</td>
<td>0.1875</td>
<td>0.1562</td>
<td>0.1406</td>
<td>0.1406</td>
<td></td>
<td>0.1250</td>
<td>0.0937</td>
<td>0.0809</td>
<td>0.0744</td>
<td>0.0772</td>
<td></td>
</tr>
</tbody>
</table>

The top number is the stopping constant $r^*$ for the sequential plan. The middle number is the plan's expected number of offspring examined per family. The bottom number is $N$ times the variance of the Levene estimator. The set-up cost is 3.8 offspring per collected parent.
Table 5

Familial data on parents collected from Davis Mountains State Park, Texas, on July 24–25, 1982

<table>
<thead>
<tr>
<th>Fathers</th>
<th></th>
<th>Mothers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Family profile</td>
<td>Score</td>
<td>Event</td>
<td>Weight</td>
</tr>
<tr>
<td>TL TL TL PP</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>AR AR PP</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>AR AR AR AR</td>
<td>0</td>
<td>$\Lambda_0$</td>
<td>0</td>
</tr>
<tr>
<td>PP AR</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>PP PP PP PP</td>
<td>2</td>
<td>$\Lambda_2$</td>
<td>1</td>
</tr>
<tr>
<td>PP AR</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>AR PP</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>PP AR</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>PP AR</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>PP PP PP PP</td>
<td>2</td>
<td>$\Lambda_2$</td>
<td>1</td>
</tr>
<tr>
<td>PP PP PP PP</td>
<td>2</td>
<td>$\Lambda_2$</td>
<td>1</td>
</tr>
<tr>
<td>TL TL TL TL</td>
<td>0</td>
<td>$\Lambda_0$</td>
<td>0</td>
</tr>
<tr>
<td>AR AR AR AR</td>
<td>0</td>
<td>$\Lambda_0$</td>
<td>0</td>
</tr>
<tr>
<td>PP PP PP AR</td>
<td>1</td>
<td>$\Lambda_1$</td>
<td>1/2</td>
</tr>
<tr>
<td>PP PP PP PP</td>
<td>2</td>
<td>$\Lambda_2$</td>
<td>1</td>
</tr>
<tr>
<td>PP PP PP PP</td>
<td>2</td>
<td>$\Lambda_2$</td>
<td>1</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>ARPP PPPP ARPP</td>
<td>1 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
<tr>
<td>PPP PPP PPP</td>
<td>2 + 2</td>
<td>$\Lambda_{12}$</td>
<td>3/4</td>
</tr>
</tbody>
</table>

The initial estimate $\theta = .62$ and maximum family size $n = 5$ yielded an efficient plan $r^* = 4$ for males and $r^* = 3$ for females from Table 4. These efficient plans were used to collect the data above. The *scoring procedure* is that of Dobzhansky (Arnold and Morrison, 1985, p. 790).
must also have some preliminary estimate \( \hat{\theta} \) of the allele probability from previous samples. If he truly has no prior information about the allele probability, he can entertain a least favorable situation by guessing \( \theta = \frac{1}{2} \) (see Result 4) in choosing a sequential sampling plan \( r^* \) from Table 4. Reading from the appropriate row (\( \theta \)) and column (\( n \)) yields the efficient sampling plan \( r^* \). Below each plan \( r^* \), the table also reports the expected number of offspring per family \( \eta_{r^*,n}^{(2)} \) in Table 2 and \( N \) times the variance to be expected under the plan \( r^* \). This last variance entry in each cell can also be used to read out \( N \) times the estimated variance for the Levene estimate, once data are obtained.

The efficient sequential plans \( r^*(\theta) \) for each sampling scheme, their expected family size \( \eta_{r^*,n}^{(0)} \), and their information per unit cost of observation all vary in a simple way with the allele probability, as described in Result 5.

**Result 5.** For each sampling scheme (\( Si \)) and protocol, the efficient plan \( r^* \) is a concave step function, symmetric about its maximum at \( \theta = \frac{1}{2} \). For each sampling scheme (\( Si \)) and protocol, the expected family size \( \eta_{r^*,n}^{(0)} \) for the efficient plan \( r^* \) is a concave, piecewise smooth function, symmetric about its maximum at \( \theta = \frac{1}{2} \) and with minimum of \( r^* \) at \( \theta = 0 \) or 1.

The properties of the efficient plan \( r^* \), its expected relative cost \( \eta_{r^*,n}^{(0)} \), and its information per unit cost of observation from Result 4 are all summarized in Figure 1. It is of interest to the investigator to know the maximum amount of work he will be asked to perform in the least favorable situation when the allele probability \( \theta \) is \( \frac{1}{2} \). Using a set-up cost of 3.8, we find the efficient plans and their expected cost per family in Table 4. In each case we find that the sequential plans never require us to observe a run of more than 5 offspring (\( r^* \leq 5 \)). This fact is to be contrasted with the earlier recommendations of at least 8 in Dobzhansky et al. (1963) for the female protocol. Dobzhansky’s recommendation would appear to correspond to a set-up cost of 8 or more. If the set-up cost is 3 offspring or less per family, then the recommendation of Morris and Spieth (1978) \( (r^* = 1) \) is correct, using stopping rules (\( S0 \)) and (\( S2 \)) for the female protocol and using (\( S0 \)) and (\( S1 \)) for the male protocol.

When populations are monomorphic under (\( S1 \)), the efficient plan \( r^* \) requires waiting for only one observation \( (r^* = 1) \); the expected relative cost of sampling is smallest, namely close to 4; and the information per unit cost of observation is highest (Figure 1). As heterozygosity increases, it is more efficient to sample more offspring \( (r^* > 1) \); the expected relative cost increases; and the information per unit cost of observation goes down.

**4. An Example**

Every 10 years since the 1930s, a survey of the inversion polymorphism in *Drosophila pseudoobscura* has been made (Dobzhansky and Powell, 1975). In the 1980s, 48 collection sites throughout the western United States were sampled. For example, on July 24–25, 1982, J. R. Powell collected 16 males and 26 females from a site in the Davis Mountains State Park, Texas, polymorphic predominantly for two inversions, Arrowhead (\( AR \)) and Pike’s Peak (\( PP \)). L. B. Klaczko karyotyped larvae from each collected parent to estimate the probability that a IIIrd chromosome carries PP. The basic design question is how many offspring to examine from each adult. This translates into a choice of length of run \( r^* \) of one genotype leading to a cessation of sampling. To select this efficient offspring sampling plan \( r^* \), Powell and Klaczko need three numbers: (i) a preliminary estimate of the inversion probability; (ii) the set-up cost per adult collected; and (iii) the maximum feasible family size. First, from the 1970s survey (Anderson et al., 1975), Powell and Klaczko have a preliminary estimate of the inversion probability, i.e., \( \hat{\theta} = .62 \). Second, an experienced worker can prepare and examine 8 offspring per hour. The time (including travel to and
from the site) to collect parents is 16 hours. (If multiple sites had been visited, the time to reach a given site would be part of the given site's set-up cost; return time to the laboratory at the end of collecting could be equally divided between sites.) In the Davis Mountains collection, 42 adults were obtained. To set up 42 lines would involve preparing a rack of vials and the transfer of 42 parents to their individual vials: 4 hours of work. The total set-up cost is 20 hours or, equivalently, 160 = 8 × 20 offspring identified. The set-up cost per family is 3.8 = 160/42 offspring per family. This set-up cost will vary from collection to collection and with the polymorphism being studied (Clegg, unpublished Ph.D. dissertation, University of California, Davis, 1972; Morris and Spieth, 1978; and Highton, 1975). Third, they decide that they are not willing to examine more than 5 offspring per adult (n = 5).

From Table 4, the plan recommended is to stop sampling when: (i) a definitive score is made; (ii) a run of 4 males (or 3 females) of one genotype is seen; or (iii) 5 offspring are examined. According to Table 4, the expected number of offspring examined, namely $N_{1.5}$, will be (16) × (3.4) offspring from fathers and (26) × (3.9) offspring from mothers for a total of 156 offspring. The cost under the fixed plans (Arnold, 1981; Arnold and Morrison, 1985) would have been (16)(5) + (26)(5) = 210, a saving of 54 offspring examined. The anticipated precision in the gene arrangement's probability estimates in males and females is also available from Table 4, namely $\sqrt{.1325/16} = .09$ for males and $\sqrt{.0726/26} = .05$ for females.

The family profiles in Table 5 were collected according to the most efficient sampling plan under the sampling schemes of the Levene estimator (5). For the familial data in Table 5, the Levene estimates are

$$\hat{\theta} = (N_2 + \frac{1}{2}N_1)/N = [5 + \frac{1}{2}(8)]/(16) = .56$$

for fathers and

$$\hat{\theta} = (N_2 + \frac{3}{4}N_{12} + \frac{1}{2}N_{02} + \frac{1}{2}N_{11} + \frac{1}{4}N_{01})/N$$

$$= [6 + \frac{3}{4}(15) + \frac{1}{2}(1) + \frac{1}{2}(2) + \frac{1}{4}(2)]/(26) = .74$$

for mothers. Consulting Table 4 yields their standard error estimates of $\sqrt{.1368/16} = .09$ and $\sqrt{.0618/26} = .05$ for males and females, respectively. The number (131) of offspring examined is actually less than the expected number, 156. A FORTRAN-77 program is available from J. Arnold for computing efficient sequential plans.

ACKNOWLEDGEMENTS

J. Arnold wishes to thank J. C. Avise for fruitful discussions. We also wish to thank P. E. Smouse, C. E. Taylor, J. R. Powell, A. H. D. Brown, A. Clark, I. R. Savage, M. Ball, A. F. MacRae, and J. Felsenstein for their detailed criticisms of this manuscript leading to its substantial improvement. This material is based on work supported by the National Science Foundation under grant BSR-8315821.

RÉSUMÉ

Nous avons, dans un travail antérieur, développé plusieurs modèles permettant l'analyse génétique de populations naturelles. Des parents, de genotype inconnu, sont tirés d'une population naturelle, polymorphe à un seul locus. On observe les génotypes d'un certain nombre de descendants issus de ces N parents. Le problème est de choisir un plan d'échantillonnage des descendants, efficace pour estimer la fréquence d'un allèle dans la population des parents à partir des structures génétiques des N familles de descendants. Un critère, appelé information par coût unitaire d'observation, est introduit pour évaluer des plans d'échantillonnage séquentiels dans lesquels le nombre de descendants par famille est aléatoire. On a introduit quelques règles simples et commodes pour arrêter l'échantillonnage des descendants d'un parent donné. On arrête l'échantillonnage lorsque (i) pour la 1re fois, les
descendants permettent de conclure sur le genotype des parents; (ii) sur un nombre déterminé de descendants, on n’observe qu’un seul genotype; (iii) on a examiné le maximum de descendants. Nous recommandons cette procédure. Pour chaque plan d’échantillonnage, nous fournissons le meilleur estimateur linéaire non biaisé et l’estimateur séquentiel du maximum de vraisemblance de la fréquence de l’allèle. La connaissance des moments de ces estimateurs permet de fournir des plans d’échantillonnage séquentiels, meilleurs (au sens de l’information par coût unitaire), aussi simples et moins coûteux que les procédures non séquentielles usuelles.

REFERENCES


Sequential Alternatives to the Dobzhansky Estimator


Received April 1985; revised November 1985.

APPENDIX

Almost all of the results in this paper can be derived from Table 6 for the female protocol and simpler analog for the male protocol. Calculations for the sampling scheme (S2) are presented for the female protocol. The calculations for the score probabilities in Table 1 are briefly indicated in the proof of Result 3.

Proof of Result 1. If the estimator \( \hat{\theta} \) in (7.7) is to be unbiased for any sampling scheme leading to (6.6), then we must have

\[
K_{22} + w_{12}K_{12} + \cdots + w_{11}K_{11} = \theta = \theta (\theta + \bar{\theta})^3
\]

or

\[
K_{22} + w_{12}K_{12} + \cdots + w_{11}K_{11} = \theta^4 + 3\theta^3\bar{\theta} + 3\theta^2\bar{\theta}^2 + \theta\bar{\theta}^3. \quad (A.1)
\]

The right-hand side (RHS) and left-hand side (LHS) are polynomials in \( \theta^4, \theta^3\bar{\theta}, \theta^2\bar{\theta}^2, \) and \( \theta\bar{\theta}^3. \) Matching the coefficients on the RHS using Table 6 with those on the LHS yields a linear system of equations in the weights:

\[
2p \ w_{02} + 2g + 4hw_{12} = 3,
\]

\[
2(1 + p) \ w_{02} + g^2 + 4hw_{01} + 4hw_{12} + 4fw_{11} = 3,
\]

\[
2p \ w_{02} + 0 + 4bw_{01} = 1.
\]

Solving these equations for \( w_{12}, w_{11}, \) and \( w_{01} \) in terms of \( w_{02} \) yields the family of linear sequential unbiased estimators in Result 1. These expressions for the weights \( w \) hold provided a sampling scheme leads to (6) and \( r > 1. \)

Proof of Result 2. Provided a sampling scheme leads to (6.6) under the male protocol, there is one and only one linear sequential unbiased estimator. Provided a sampling scheme leads to (6.6) under the female protocol, the variance of a linear sequential unbiased estimator \( \hat{\theta} \) can be calculated from the well-known properties of a list of multinomial counts, such as \( N, \) using the unbiasedness condition:

\[
K_{22} + w_{12}K_{12} + \cdots + w_{11}K_{11} = \theta.
\]
Table 6
Score probabilities for female protocol

<table>
<thead>
<tr>
<th>Event</th>
<th>Coefficient</th>
<th>$\theta^4$</th>
<th>$\theta^3\bar{y}$</th>
<th>$\theta^2\bar{y}^2$</th>
<th>$\theta\bar{y}^3$</th>
<th>$\bar{y}^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{00}$</td>
<td>0</td>
<td></td>
<td>$g^2$</td>
<td>$2g$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$A_{02}$</td>
<td>$w_{02}$</td>
<td></td>
<td>$2p$</td>
<td>$2(1 + p)$</td>
<td>$2p$</td>
<td></td>
</tr>
<tr>
<td>$A_{22}$</td>
<td>1</td>
<td>$l$</td>
<td>$g^2$</td>
<td>$4h$</td>
<td>$4b$</td>
<td></td>
</tr>
<tr>
<td>$A_{01}$</td>
<td>$w_{01}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_{12}$</td>
<td>$w_{12}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$A_{11}$</td>
<td>$w_{11}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(The weights that might be attached to $N_{22}$ or $N_{00}$, i.e., $w_{22}$ and $w_{00}$, are forced to be 1 and 0 by the unbiasedness condition, as from Table 6.) We have:

$$\text{var}^*(\hat{\theta}) = \left[K_{22}\bar{y} + w_{12}K_{12}(w_{12} - \theta) + \cdots + w_{01}K_{01}(w_{01} - \theta)\right]/N.$$  \hfill (A.2)

The information per unit cost of observation is $(s = 0)$:

$$1/[\text{var}^*(\hat{\theta})_N]_s.$$  \hfill (A.2)

Since the expected number of offspring per family $\eta_{r,s}^{(x)}(x)$, is independent of the weight $w_{02}$, the information per unit cost of observation is maximized when the variance $\text{var}^*(\hat{\theta})$ is minimized.

Each of the weights $w_{12}$, $w_{11}$, and $w_{01}$ is linear in $w_{02}$, the only freely varying weight. The variance $\text{var}^*(\hat{\theta})$ is positively and linearly related to $w_{02}(w_{02} - 1)$, which takes its minimum value at $w_{02} = 1/2$. Thus, the variance has a minimum when $w_{02} = 1/2$. Substituting $w_{02} = 1/2$ in the other weights yields the best linear sequential unbiased estimator. For sampling scheme (S2) in Table 1, we have

$$w_{12} = 1/4, \quad w_{11} = 1/2, \quad w_{01} = 1/4, \quad \text{and} \quad w_{02} = 1/2.$$  \hfill (A.2)

Proof of Result 3. We calculate the expected number of offspring per family under sampling scheme (S2). The result is obtained from Table 6.

Let $x$ be the time we stop sampling a family. If the family profile falls in $A_{22}$, then we stop sampling with $x = r$ under (S2). The event $A_{22}$ makes a contribution,

$$r(\theta^4 + 2g\theta^3\bar{y} + g^2\theta^2\bar{y}^2),$$

to $\eta_{r,s}^{(x)}$ from Table 6. Similarly, the contribution of events $A_{00}$ and $A_{02}$ can be computed with stopping occurring at $x = r$ again under (S2). If events $A_{01}$ or $A_{12}$ occur, we continue sampling until $x = n$. For example, the contribution of $A_{01}$ to $\eta_{r,s}^{(x)}$ is

$$n(4h\theta^2\bar{y}^2 + 4b\theta\bar{y}^3).$$

The only challenging event for which to calculate a contribution is $A_{11}$. Assume we end on an offspring homozygous for the allele (i.e., the first time $n_2 > 0$). Let $x$ be the first time we see an offspring homozygous for the allele, leading to a definitive score. We have already seen at least one offspring homozygous for the other allele so that $n \in A_{11}$.

Case (a). If $x \leq r$, the probability of stopping at offspring $x$ is

$$\frac{4(\bar{y})^{x-1} - (\bar{y})^{x-1}}{4(\bar{y})^{x-1} - (\bar{y})^{x-1}},$$

where the subtracted term ensures that we see at least one offspring homozygous for the other allele in the first $x - 1$ offspring and where the $\frac{1}{4}$ ensures we see no offspring homozygous for the allele until $x$.

Case (b). If $x > r$, the probability of stopping at offspring $x$ is

$$\frac{(\bar{y})^{x-1}[1 - (\bar{y})^{x-1} - (\bar{y})^{x-1}]}{(\bar{y})^{x-1}[1 - (\bar{y})^{x-1} - (\bar{y})^{x-1}]}.$$  \hfill (A.4)

The last two terms subtracted are the probabilities of stopping prematurely with a run of $r$ offspring homozygous for the other allele or heterozygous for both alleles. The terms (A.3) and (A.4) need to be multiplied by 2 for ending on an offspring homozygous for the other allele and by 4, the coefficient of $\theta^2\bar{y}^2$ in Table 6.
Summing these two terms (multiplied by $\theta^2\bar{\theta}^2$) and the contribution of $\lambda_0, \ldots, \lambda_{12}$ in Table 6 yields

$$
\eta_{r,n}^{(2)} = r(\theta^4 + \bar{\theta}^4) + 4\theta\bar{\theta}(\theta^2 + \bar{\theta}^2)(\hat{g}n + rg) \\
+ (\theta\bar{\theta})^2 \left[ 8nh + 2r(g + 1) + 2rg^2 \\
+ \left( 2 \cdot \frac{4}{4} \right) \sum_{x=2}^{n} x \left( \frac{3}{4} \right)^{x-1} - \left( 2 \cdot \frac{4}{4} \right) \sum_{x=2}^{r} x \left( \frac{1}{2} \right)^{x-1} \\
- \left( 2 \cdot \frac{4}{4} \right) \sum_{x=2}^{n} x \left( \frac{3}{4} \right)^{x-1} \left( \frac{1}{4} \right)^{x-1} - \left( 2 \cdot \frac{4}{4} \right) \sum_{x=2}^{r} x \left( \frac{3}{4} \right)^{x-1} \left( \frac{1}{2} \right)^{x-1} \right].
$$

Using the closed form for the derivative of a partial sum of a geometric series ($0 < y < 1$),

$$
\frac{d}{dy} \left( \sum_{x=0}^{n} y^x \right) = \frac{1 - y^{n+1}}{y^2}, \quad y = 1 - y,
$$

we obtain the expression $\eta_{r,n}^{(2)}$ in Table 2. If the quantities $r, n$, and $x$ explicit are removed, the use of a partial sum of a geometric series yields $K_{11}$ for the (S2) scheme in Table 1.

The concavity can be verified by taking derivatives. The symmetry is evident by the interchangeability of $\theta$ and $\bar{\theta}$.