Am. J. Hum. Genet. 43:387–395, 1988

A Resolution of the Ascertainment Sampling Problem. III. Pedigrees

Nereda C. E. Shute* and W. J. Ewens*,†

*Department of Mathematics, Monash University, Clayton, Victoria, Australia; and †Department of Biology, University of Pennsylvania, Philadelphia

Summary

When nuclear families are sampled by an ascertainment procedure whose properties are not known, biased estimates of genetic parameters will arise if an incorrect specification of the ascertainment procedure is made. Elsewhere we have put forward a resolution of this problem by introducing an ascertainment-assumption-free (AAF) method, for nuclear family data, which gives asymptotically unbiased estimators no matter what the true nature of the ascertainment process. In the present paper we extend this method to cover pedigree data. Problems that arise with pedigrees but not with families—for example, the question of which families in a pedigree are "ascertainable"—are also considered. Comparisons of numerical results for pedigrees and nuclear families are also made.

Introduction

The ascertainment problem has been described in earlier papers in this series (Ewens and Shute 1986*b*; Shute and Ewens 1988). It arises because some assumption must be made about an ascertainment process in analyzing the data it leads to, and an incorrect assumption will lead to biased estimates of genetic parameters. The classical ascertainment procedure of Weinberg (1928) and Fisher (1934) potentially suffers from this defect. In our earlier papers we proposed a resolution of this problem for the case of nuclear families, and in the present paper we extend this to the case of pedigrees and note also various problems with pedigrees not present with nuclear families.

For nuclear families, the ascertainment problem is overcome by conceptually dividing the data d in any family into two components, d_1 and d_2. The component d_1 is that part of the data "relevant to ascertainment" (so that two families with the same d_1 have the same probability of being ascertained), while d_2 is that part of the data "not relevant to ascertainment." In an earlier paper (Ewens and Shute 1986*b*) we show that asymptotically unbiased estimation of genetic parameters is obtained, no matter what the (unknown) ascertainment procedure might be, by using as likelihood the expression L, defined by

$$L = \prod_m \prod_{d_1} \prod_{d_2} \left[P_m(\text{asc},d_1,d_2)/P_m(\text{asc},d_1)\right]^{n(m,d_1,d_2)}, \quad (1)$$

where $n(m,d_1,d_2)$ is the number of families in the sample having m children and data $\{d_1,d_2\}$. In effect, this likelihood is the conditional probability of the data, given the values of the data relevant to ascertainment—i.e., (d_1)—from each family.

Use of this ascertainment-assumption-free (AAF) likelihood results in a loss of information due to the conditioning on d_1 and hence leads to standard errors of parameter estimates larger than those arising when the true ascertainment process happens to be correctly assumed. When a procedure for the estimation of genetic is chosen, there must be a trade-off between (*a*) the bias in the estimates under the "classical" approach if the ascertainment assumptions made under this approach do not hold and (*b*) the increase in standard errors of the estimates when the AAF likelihood (expression [1]) is used. The extent to which one is confident about knowing the true ascertainment process will be important in making this choice, as are the typical values (assessed by Shute and Ewens [1988]) of the sizes of these biases and standard errors.

The claim has been made that a disadvantage of the

Received May 7, 1987; revision received March 21, 1988.

Address for correspondence and reprints: W. J. Ewens, Department of Mathematics, Monash University, Clayton, Victoria, 3168, Australia.

0002-9297/88/4304-0004$02.00

new approach is that one has to be able to specify what part d_1 of the total data is relevant to ascertainment. However, this specification is necessary under every ascertainment procedure. If d_1 is the number of affected children in the family (as is often the case), likelihood (1) arises, in effect, by one's assuming some arbitrary probability $a_m(d_1)$ that a family with m children, d_1 of whom are affected, is ascertained. The classical method assumes a specific mathematical form

$$P(\text{asc}|d_1) = 1 - (1 - \pi)^{d_1} \qquad (2)$$

for this probability, for some unknown parameter π. Use of this function, however, still implies a specific choice of d_1, and the AAF approach arises simply by replacing the specific (and often unreasonable) mathematical function (2) by the arbitrary probability $a_m(d_1)$. We return to this point later in discussing pedigrees. These comments remain true even for particular cases of the classical procedure, in particular the cases $\pi = 1$ (complete ascertainment) and $\pi \cong 0$ (single ascertainment), when (2) takes the forms

$$P(\text{asc}|d_1) = \text{const}, \text{const} \times d_1, \qquad (3)$$

respectively.

Our aim is to find the analogue of (1) for data from pedigrees and to examine various questions of ascertainment sampling that arise for pedigrees but not for nuclear families. We will also compare numerical properties of estimators from pedigrees with those from families.

Theory for Pedigrees

We write the data for any pedigree in the form $\{M, D_1, D_2\}$. M is a vector describing the structure of the pedigree, including the number of families contained in the pedigree, how they are related, the number of children in each of them, and also which of these families are within the geographical area in which the ascertainment sampling takes place. D_1 describes all the pedigree data that are relevant to ascertainment, and D_2 is all the remaining data (phenotypic and genotypic). For example, if pedigrees are ascertained through affected girls living in a certain area, then D_1 is the number of affected girls living in that area. We define the following terms: $P_M(\text{asc},A)$ = the probability that a pedigree having structure M is ascertained and is of type A (A describes the data contained in the pedigree); $Q_M(A)$ = the population probability that a pedigree having structure M is of type A; and $P_M(\text{asc}|D_1)$ = the probability that a pedigree having structure M and data D_1 is ascertained. These definitions imply that

$$P_M(\text{asc}, D_1,D_2) = Q_M(D_1,D_2)P_m(\text{asc}|D_1)\,, \qquad (4a)$$

$$P_M(\text{asc},D_1) = Q_M(D_1)\, P_m(\text{asc}|D_1)\,, \qquad (4b)$$

$$P_M(\text{asc}) = \sum_{D_1} Q_M(D_1)\, P_m(\text{asc}|D_1)\,. \qquad (4c)$$

The likelihood for a sample of pedigrees is now

$$L = \prod_M \prod_{D_1} \prod_{D_2} \left[P_M(\text{asc},D_1,D_2)/P_M(\text{asc})\right]^{n(M,D_1,D_2)}$$
$$= \prod_M \prod_{D_1} \prod_{D_2} \left[\frac{Q_M(D_1,D_2)\, P_m(\text{asc}|D_1)}{\sum_i Q_M(i)P_m(\text{asc}|i)}\right]^{n(M,D_1,D_2)}, \qquad (5)$$

where $n(M,D_1,D_2)$ is the number of pedigrees sampled that have structure M and data (D_1,D_2), i runs through all D_1 values, and the product involving M covers all possible structure vectors. In this expression, the probabilities $P_m(\text{asc}|D_1)$ and $P_m(\text{asc}|i)$ are specified functions (analogous to [2]) of ascertainment parameters under the classical approach or are a set of free parameters (analogous to the $a_m(i)$ for nuclear families) under the AAF approach.

Thus, if, under the classical approach, we assume complete ascertainment (so that $P(\text{asc}|i) = \text{const}$), likelihood (5) becomes

$$L = \prod_M \prod_{D_1} \prod_{D_2} \left[Q_M(D_1,D_2)/\sum_i Q_M(i)\right]^{n(M,D_1,D_2)}, \qquad (6)$$

while if we assume single ascertainment (for which $P(\text{asc}|i) = \text{const} \times i$), (5) becomes

$$L = \prod_M \prod_{D_1} \prod_{D_2} \left[D_1 Q_M(D_1,D_2)/\sum_i i\, Q_M(i)\right]^{n(M,D_1,D_2)}. \qquad (7)$$

Under the AAF approach, no specification of $a_M(i)$ is made and $a_M(i)$ is estimated along with the genetic parameters. Theory analogous to that in one of our earlier papers (Ewens and Shute 1986*b*) then shows that, as for nuclear families, estimation of the $a_M(i)$ values separates from estimation of genetic parameters and that the latter are found by using the likelihood L, defined by

$$L = \prod_M \prod_{D1} \prod_{D2} \left[P_M(\text{asc},D_1,D_2)/P_M(\text{asc},D_1)\right]^{n(M,D1,D2)}, \quad (8)$$

which is the direct analogue of (1). A formula more useful for computation is

$$L = \prod_M \prod_{D1} \prod_{D2} \left[Q_M(D_1,D_2)/Q_M(D_1)\right]^{n(M,D1,D2)}. \quad (9)$$

Use of (8) leads to estimates of genetic parameters that are free of any specific ascertainment assumption but that will have an accompanying increase in their standard errors, compared with those arising from (6) and (7). As with nuclear families, a trade-off between (*a*) the bias possible through incorrect use of (6) or (7) and (*b*) the increase in standing error using (8) might determine the approach used, depending again on the degree of confidence we have in our knowledge of the true ascertainment procedure.

We remark that, as with nuclear families, it is not a valid criticism of the AAF procedure that it requires one to specify the data relevant to ascertainment. This specification must also be made under the classical approach (see, e.g., the probabilities [6] and [7]) and indeed under any approach. We do, however, agree that in practice it might be very difficult to make this specification. The theory in one of our earlier papers (Ewens and Shute 1986*b*) shows that asymptotically unbiased estimates will arise if one conditions, if possible, on the affectedness status of all individuals, although this will imply large standard errors of estimates of absolute penetrance values.

An Example

We illustrate the theory with a very simple (and very unrealistic) example which is free of the complexities of a more realistic case. In this way the properties of the AAF method can be highlighted, as well as various problems specific to pedigrees but not to families. Application of the theory to the more realistic segregation analysis model of HLA-linked diseases is currently being examined.

Consider a disease that is determined by the genes at a single disease susceptibility locus *S*, admitting a susceptibility allele *S* and a normal allele *s*. The penetrances for the three possible genotypes are known, and are here assumed to be

genotype	*SS*	*Ss*	*ss*
penetrance	.7	.1	0

The objective is to estimate the population frequency p of the disease allele *S* by using data deriving from pedigrees. We shall use the simplest pedigree in all the following work, again to highlight the points at issue.

Consider two brothers (B1 and B2) and their unrelated wives (W1 and W2), yielding family 1 and family 2. Suppose there are two children in each family (C1, C2, C3, and C4), the sex of whom is irrelevant. The parents of the two brothers (G1 and G2) are not observed, although their existence is of course required for various calculations. All other individuals in the pedigree are observed. For simplicity, all pedigrees in the sample are assumed to have this structure, so we discard the suffix M (in the formulas of the previous section) in our calculations.

The four "independent" individuals in this pedigree are G1, G2, W1, W2, and we assume that each of these is independently *SS*, *Ss*, *ss*, with respective probabilities p^2, $2p(1-p)$, and $(1-p)^2$. Given this, we can calculate the probabilities for the joint genetic constitutions of all individuals observed. Knowing the penetrances of the disease, we can also calculate the probabilities of each of the 2^8 sets of affectedness statuses of the eight individuals observed. The population probability of the pedigree having affectedness status configuration j is denoted $Q(j)$ (where $j = 1, \ldots 256$ is an index of the 2^8 affectedness statuses), such that

$$\begin{aligned} Q(1) &= \text{Prob } \{\text{W1, W2, B1, B2, C1, C2, C3, C4 all affected}\} \\ Q(2) &= \text{Prob } \{\text{W1, W2, B1, B2, C1, C2, C3 affected; C4 not affected}\} \\ &\vdots \\ &\vdots \\ Q(256) &= \text{Prob } \{\text{no one affected}\}. \end{aligned} \quad (10)$$

These are all functions of p and can be computed easily.

Now assume that a pedigree is ascertained through affected children and that all the sibs, parents, aunts, uncles, and cousins of the affected individuals are examined. In this case, the data relevant to ascertainment (D_1) is (in a way made more precise below) the number of affected children in the pedigree and D_2 consists

of the remaining data, namely, the affectedness status of parents, aunts, uncles, and (possibly) cousins.

The first problem specific to pedigrees but not to nuclear families concerns the location of the families. The pedigree under consideration consists of two distinct families, so it is possible that the families live in geographical areas remote from each other and that ascertainment is possible through only one family, the other family not being within the geographical scope of the investigation. Such a case will obviously result in different probabilities of ascertainment than those arising when it is possible for the pedigree to be ascertained through either family. In the following calculations we shall consider both of these cases. It is possible also that the investigator—or the computer program he is using—might incorrectly specify which families are ascertainable. We consider below all four possible combinations.

A. Ascertainment through One Family Only

If ascertainment is possible through only one family of the pedigree (say, family 1), then D_1 (that part of the data relevant to ascertainment) is the number of affected children in the family (which we denote K_1) and D_2 (the data not relevant to ascertainment) is all the remaining affectedness data in both families. D_2 consists of $\{K_2,D_3\}$, where K_2 is the number of affected children in family 2 and D_3 is the affectedness data in both sets of parents. We thus write $\{D_1,D_2\}$ as $\{K_1,(K_2,D_3)\}$. We suppose initially that the investigation correctly allows for the fact that ascertainment is possible only through family 1.

We consider data arising from three different true ascertainment schemes: (i) *complete ascertainment*—where the probability of ascertainment is independent of the number of affected children, K_1, in the ascertainable family (family 1); (ii) *single ascertainment*—where the probability of ascertainment is proportional to K_1; and (iii) *quadratic ascertainment*—where the probability of ascertainment is proportional to K_1^2. In the case of nuclear family data, complete and single ascertainment define limiting cases if the probability of ascertainment is defined as in (2). However, it is possible to devise a quadratic process of ascertainment in which the probability of ascertainment of a family is proportional to the square of the number of affected children in that family, i.e. (see [3]), $P(\text{asc}|d_1) = \text{const} \times d_1^2$. This quadratic case cannot be described by (2) and is outside the limits defined by the single-to-complete range. This problem has been discussed fully by use elsewhere (Ewens and Shute 1986*a*; Shute and Ewens 1988). The 1985 Genetic Analysis Workshop presented data of an HLA-linked disease that appears to have been gathered using a type of "quadratic" ascertainment process (Ewens et al. 1986). It is therefore reasonable to consider the quadratic case as a possible type of ascertainment process for pedigrees as well as for families.

The method we use for finding properties of different estimation procedures, for the three forms of data above, was suggested by Morton (1984) and was used by us for family data (Ewens and Shute 1986*b*, 1988). A particular value of the parameter p is chosen and one of the ascertainment schemes (i)–(iii) given above is assumed to hold. The mean of each $n(D_1,D_2)$ combination is calculated for this value of p and the chosen ascertainment process, assuming a sample of n pedigrees. The data are now taken to be exactly as these mean values. These data are now inserted, as the $n(D_1,D_2)$ values, in each of the likelihoods (6), (7), and (9), and the maximum likelihood estimates are calculated. The values obtained are the asymptotic (i.e., large-sample-size) mean values of the maximum likelihood estimates of p, and a standard information-theory formula gives the asymptotic standard error of each asymptotically unbiased estimate. From now on we often omit the word "asymptotic" when referring to biases and standard errors arrived at in this way.

In the example we consider, the likelihoods (6), (7), and (9) simplify, on putting

$$Q = \sum_{K1=1}^{2} \sum_{K2=0}^{2} \sum_{D3} Q(\mathrm{K_1},\mathrm{K_2},\mathrm{D_3}) ,$$

and

$$Q^* = Q(K_1 = 1) + 2Q(K_1 = 2)$$

with

$$Q(K_1 = j) = \sum_{K2=0}^{2} \sum_{D3} Q(j,K_2,D_3) ,$$

to

$$L = \prod_{K=1}^{2} \prod_{K2=0}^{2} \prod_{D3} \left[Q(K_1,\mathrm{K_2},\mathrm{D_3})/Q\right]^{n(K_1,K_2,D3)} , \tag{11}$$

for the case where complete ascertainment is assumed, to

$$L = \prod_{K1=1}^{2} \prod_{K2=0}^{2} \prod_{D3} \left[K_1 Q(K_1,K_2,D_3)/Q^*\right]^{n(K_1,K_2,D3)}, \quad (12)$$

when single ascertainment is assumed, and to

$$L = \prod_{K1=1}^{2} \prod_{K2=0}^{2} \prod_{D3} \left[Q(K_1,K_2,D_3)/Q(K_1)\right]^{n(K_1,K_2,D3)},$$

$$= \prod_{K2=0}^{2} \prod_{D3} \left[Q(1,K_2,D_3)/Q(K_1=1)\right]^{n(1,K_2,D3)}$$

$$\times \left[Q(2,K_2,D_3)/Q(K_1=2)\right]^{n(2,K_2,D3)}. \quad (13)$$

for the AAF approach.

We thus have three possible true ascertainment processes ([i]–[iii], given above) and three estimation procedures (corresponding to [11]–[13]), namely, classical assuming complete ascertainment, classical assuming single ascertainment, and AAF. There are thus nine different combinations of true and assumed ascertainment schemes to be examined. For each of these combinations the maximum likelihood estimate of p can be obtained, along with the accompanying standard error of unbiased estimates; the results are shown in table 1 for the case where the true value of p is .1.

The main conclusion that we draw from table 1 is that, as in the case of nuclear families, the classical method gives correct estimates only if the correct ascertainment process is assumed, while the AAF method always gives unbiased estimates, no matter what the true ascertainment process.

The ratio of AAF standard errors to standard errors derived from the classical method tends to be smaller for pedigree data than for nuclear family data. For example, when the data are derived from a complete ascertainment process, the ratio of these standard errors for pedigrees is 1.08, compared with 1.19 for nuclear families. Similarly, when the data are from a single ascertainment scheme, the ratio is 1.30 for nuclear families and 1.13 for pedigrees. Hence, when a trade-off is made between the potential estimation bias arising by using the classical method (which assumes either complete or single ascertainment) and increased standard error in the AAF method, there is not as large a sacrifice of information in the conditioning process implied by the AAF method when using pedigree data as opposed to nuclear family data. As a result one might be more inclined to use the AAF method to analyze pedigree data.

The last column in table 1 gives the standard errors of the unbiased estimates for nuclear families (Ewens and Shute 1986*b*) under the same true and assumed ascertainment scheme combinations as are considered for pedigree data. It is of interest to compare the standard errors of the estimates from the nuclear family data and those of the pedigree data, since in both instances ascertainment occurs through one particular family. Should we expect that a pedigree containing two families will contain twice as much information as one nuclear family? Since the inverse of the variance is used as a measure of information, then doubling the infor-

Table 1

Maximum Likelihood Estimates and Standard Errors of the Unbiased Estimates of p, (True Value .1), Using Likelihoods (11)–(13), for "Data" from Three True Ascertainment Processes (10), Correctly Assuming Ascertainment Is Only through Family 1 in the Pedigree

True Ascertainment Process and Likelihood	$\hat{p}$	se($\hat{p}$) × $\sqrt{n}$	$\hat{p}$	se($\hat{p}$) × $\sqrt{n}$ (Nuclear Families)
Complete ascertainment:				
(11) (Assumes complete)	.1000	.2324	.1000	.3402
(12) (Assumes single)	.0816	. . .	.0641	. . .
(13) (AAF)	.1000	.2504	.1000	.4057
Single Ascertainment:				
(11) (Assumes complete)	.1197	. . .	.1407	. . .
(12) (Assumes single)	.1000	.2184	.1000	.3009
(13) (AAF)	.1000	.2565	.1000	.3910
Quadratic ascertainment:				
(11) (Assumes complete)	.1501	. . .	.2009	. . .
(12) (Assumes single)	.1287	. . .	.1534	. . .
(13) (AAF)	.1000	.2403	.1000	.3699

NOTE.—Nuclear family data are from Ewens and Shute (1986*b*).

mation is equivalent to halving the variance (or to dividing the standard error by $\sqrt{2}$). Table 1 shows that, except for one case (when single ascertainment is true and assumed), all the standard errors of $\hat{p}$ when the pedigree data are used are less than $1/\sqrt{2}$ times those when only nuclear family data are used. In these cases the pedigree contains more than double the information of the nuclear family, suggesting that it is more beneficial to obtain pedigree data than to double the sample number of nuclear families. Thus, for the estimation of p in this model (but not necessarily for estimation of other parameters in more complex models), there is more information contained in two related families than there is in two unrelated (or independently ascertained) families. However, in the case of single ascertainment, a pedigree appears not to contain as much information on p as do two nuclear families. This presumably occurs because there must be at least one affected child in each nuclear family if both families are to be ascertained, whereas a pedigree only requires one affected child in the two families comprising the pedigree if the pedigree is to be ascertained. This biases the type of families (and pedigrees) in the sample as compared with families in the population as a whole. The extra information from families implied by this fact presumably outweighs the extra information provided by the pedigree structure.

It is interesting to note that, in the above example, the size of the biases when pedigree data are used are approximately only one-half the size of the biases arising when nuclear family data are used (shown in table 1). Since a pedigree contains more information than a nuclear family, knowledge of how the pedigree is ascertained is relatively less important than similar knowledge on a nuclear family. Hence, as expected, making an incorrect assumption about ascertainment of a pedigree results in smaller biases than if an incorrect assumption is made about the ascertainment of a nuclear family.

B. Ascertainment through Both Families

If ascertainment of a pedigree is possible through both families in the pedigree, then the data relevant to ascertainment (D_1) is the number of affected children in each family of the pedigree, i.e., $D_1 = \{K_1, K_2\}$. D_2 is all the remaining affectedness data of the pedigree (i.e., the affectedness status of the parents in both families). The pedigree can be ascertained only if it contains at least one affected child; so $K_1 + K_2 \geq 1$. We suppose initially that the investigator correctly notes that ascertainment is possible through both families. We now consider three true ascertainment schemes: (i) complete ascertainment—where the probability of ascertainment of the pedigree is independent of the number of affected children in the pedigree; (ii) "single" ascertainment—where the probability of ascertainment is proportional to the total number of affected children in the pedigree ($K_1 + K_2$); and (iii) "quadratic" ascertainment—where the probability of ascertainment is proportional to $(K_1 + K_2)^2$. Note that our definitions of "single" and "quadratic" ascertainment make an assumption that might be unreasonable in practice, namely, that the probability of ascertainment depends on K_1 and K_2 only through their sum, $K_1 + K_2$. Consider the case of a pedigree with two affected children. If $K_1 = K_2 = 1$ ($K_1 + K_2 = 2$), then it is possible from a practical point of view that the pedigree will be less likely to be ascertained than if $K_1 = 2$ and $K_2 = 0$ (with, again, $K_1 + K_2 = 2$), since, in the latter case, family 1 may seek medical aid because it has two affected children, whereas, in the former case, knowledge that a cousin is affected by the same disease as one of their own affected children might not cause either family to feel concerned enough to seek advice. Thus it is possible that different weightings for the probability of ascertainment for these two cases apply; ascertainment estimation procedures allowing for this will be investigated in a later paper. In the present paper the definitions presented above will be assumed. If the investigator correctly assumes that ascertainment can be through either one or both families in the pedigree, he will use the following classical likelihoods for estimating p from pedigree data.

Classical Likelihood Assuming Complete Ascertainment

Under a complete ascertainment scheme $P(\text{asc},D_1,D_2)$—i.e., $P(\text{asc},K_1,K_2,D_2)$—and $P(\text{asc})$ are proportional to $Q(K_1,K_2,D_2)$ and

$$Q = \sum_{\substack{K_1=0 \\ K_1+K_2\geq 1}}^{2} \sum_{K_2=0}^{2} \sum_{D_2} Q(K_1,K_2,D_2) ,$$

respectively, so that the likelihood used, when one assumes complete ascertainment, is

$$L = \prod_{\substack{K_1=0 \\ K_1+K_2\geq 1}}^{2} \prod_{K_2=0}^{2} \prod_{D_2} \left[Q(K_1,K_2,D_2)/Q\right]^{n(K_1,K_2,D_2)} . \quad (14)$$

Classical Likelihood Assuming Single Ascertainment

If the investigator assumes a single ascertainment process, $P(\text{asc},K_1,K_2,D)$ and $P(\text{asc})$ are proportional to

$(K_1 + K_2)Q(K_1,K_2,D_2)$ and Q^*, respectively, where $Q^* = Q(K_1 + K_2 = 1) + 2Q(K_1 + K_2 = 2) + 3Q(K_1 + K_2 = 3) + 4Q(K_1 + K_2 = 4)$ with

$$Q(K_1 + K_2 = j) = \sum_{D_2} \sum_{\substack{K_1 = 0 \\ K_1 + K_2 = j}}^{2} \sum_{K_2 = 0}^{2} Q(K_1,K_2,D_2) .$$

The likelihood then becomes

$$L = \prod_{K_1 = 0}^{2} \prod_{K_2 = 0}^{2} \prod_{D_2} \left[\frac{(K_1 + K_2)Q(K_1,K_2,D_2)}{Q^*}\right]^{n(K,K_2,D_2)} . \quad (15)$$

Two approaches to estimation are possible under the AAF approach. The first is the more conservative one, in which the investigator assumes only that ascertainment depends on $\{K_1,K_2\}$ in some unspecified manner. Under the second approach the investigator is willing to assume that ascertainment depends, albeit in an unspecified way, on $K_1 + K_2$. These approaches lead to likelihoods (16) and (17) below, respectively.

AAF Likelihoods

(*a*) *Conservative.*—The likelihood is

$$L = \prod_{\substack{K_1 = 0 \\ K_1 + K_2 \geq 1}}^{2} \prod_{K_2 = 0}^{2} \prod_{D_2} \left[\frac{Q(K_1,K_2,D_2)}{Q(K_1,K_2)}\right]^{n(K,K_1,D_2)} , \quad (16)$$

where

$$Q(K_1,K_2) = \sum_{D_2} Q(K_1,K_2,D_2) .$$

(*b*) *Ascertainment Depends on $K_1 + K_2$.*—The likelihood is

$$L = \prod_{\substack{K_1 = 0 \\ K_1 + K_2 \geq 1}}^{2} \prod_{K_2 = 0}^{2} \prod_{D_2} \left[\frac{Q(K_1,K_2,D_2)}{Q(K_1 + K_2)}\right]^{n(K_1 + K_2,D_2)} , \quad (17)$$

where

$$Q(K_1 + K_2) = \sum_{K_1' + K_2' = K_1 + K_2} Q(K_1',K_2') .$$

As before, there are three possibilities ([i]–[iii], as given above) for the true ascertainment process, and now there are four estimation procedures, corresponding to the likelihoods (14)–(17). Deterministic data are calculated as described above, and from these the estimates and their standard errors, corresponding to the 12 combinations of various true and assumed ascertainment processes, may be calculated. The results are given in table 2 for the case $p = .1$.

Table 2

Maximum Likelihood Estimates and Standard Errors of the Unbiased Estimates of p (True Value .1), Using Likelihoods (14)–(17), for Three True Ascertainment Processes (14), Correctly Assuming Ascertainment Is through Both Families in the Pedigree

True Ascertainment Process and Likelihood	$\hat{p}$	$\text{se}(\hat{p}) \times \sqrt{n}$
Complete ascertainment:		
(14) (Assumes complete)	.1000	.2519
(15) (Assumes single)	.0653	. . .
(16) (AAFa)	.1000	.3120
(17) (AAFb)	.1000	.3111
Single ascertainment:		
(14) (Assumes complete)	.1431	. . .
(15) (Assumes single)	.1000	.2184
(16) (AAFa)	.1000	.3027
(17) (AAFb)	.1000	.3012
Quadratic ascertainment:		
(14) (Assumes complete)	.2085	. . .
(15) (Assumes single)	.1543	. . .
(16) (AAFa)	.1000	.2890
(17) (AAFb)	.1000	.2868

As in the case of ascertainment through family 1 only, it is only when the true ascertainment scheme is correctly assumed that the classical method produces unbiased estimates. Both AAF approaches provide unbiased estimates, irrespective of the true ascertainment scheme. Note that, in contrast to the case when ascertainment is possible through one family only, here the asymptotic biases for pedigree data are approximately the same as or slightly higher than those for nuclear families. This shows that, in contrast to the previous case, a correct specification of the ascertainment process is as important for pedigree data as for nuclear family data.

So far as standard errors of unbiased estimators are concerned, the loss of information in the conditioning process leading to the AAF procedures implies that AAF estimators have standard errors higher than those for the classical approach when the correct ascertainment scheme happens to be assumed. This increase is, however, not large, and one may often be prepared to ac-

cept this increase in the interests of having parameter estimates that will always be asymptotically unbiased. Note also that, as expected, the standard errors of the estimators derived from the conservative likelihood, (16), are larger than those of the estimators derived from (17).

Incorrect Assumptions about the Number of Ascertainable Families

What happens to the estimates in the previous sections if the investigator now makes an incorrect assumption about whether ascertainment is possible through one or both families in the pedigree? There are two possibilities. The first arises when ascertainment is truly possible through both families in the pedigree but the investigator assumes that only one of the families is ascertainable. It is in a way inconceivable that this error can be made, since the investigator should notice pedigrees ascertained through the assumed nonascertainable families. We nevertheless consider this case because data collected by one person are often analyzed by another, who may be unaware of the true ascertainment process and who may use a "black box" computer program using one or other of the likelihoods, (11)–(13). Note in this connection that these likelihoods in effect only use a subset of all families (those where $K_1 > 0$) and that the above problem will not be highlighted by an error message from the computer.

Table 3

Maximum Likelihood Estimates and Standard Errors of the Unbiased Estimates of *p* (True Value .1), Using Likelihoods (11)–(13), for Three True Ascertainment Processes (14), Incorrectly Assuming Ascertainment Is Possible through Only One Family in the Pedigree

True Ascertainment Process and Likelihood[a]	$\hat{p}$	se($\hat{p}$) × $\sqrt{n}$
Complete ascertainment:		
(11) (Assumes complete)	.1000	.3218
(12) (Assumes single) . . .	.0816	. . .
(13) (AAF)	.1000	.3467
Single ascertainment:		
(11) (Assumes complete)	.1505	. . .
(12) (Assume single)	.1285	. . .
(13) (AAF)	.1372	. . .
Quadratic ascertainment:		
(11) (Assumes complete)	.2225	. . .
(12) (Assumes single) . . .	.1967	. . .
(13) (AAF)	.1919	. . .

[a] Both families ascertainable.

We follow the procedures indicated above to assess properties of estimators in this situation. The results are given in table 3. We note that all estimation procedures (except for the complete-ascertainment case) result in biased estimates, since the assumption of one ascertainable family omits information connected with the second family in the pedigree. In an earlier paper (Ewens and Shute 1986b) we have shown that the AAF method must condition on (at least) *all* the data relevant to ascertainment to produce unbiased parameter estimates, but in the present paper, owing to the investigator's choice of ascertainment assumption, the likelihood conditions on only a subset of the data relevant to ascertainment. Thus, biased estimates are produced in this case, even under the AAF method. This observation is relevant to our earlier remarks concerning specification of the data relevant to ascertainment.

We now turn to the case where only one family is ascertainable but the investigator or his computer program assumes that it is possible for both families in the pedigree to be ascertained. For reasons analogous to those outlined above, properties of estimators in this case should also be examined. The conclusions reached are illustrated by the numerical values shown in table 4. The conservative AAF procedure always leads to asymptotically unbiased estimators. The less conservative approach does not do so, since it does not condi-

Table 4

Maximum Likelihood Estimates and Standard Errors of the Unbiased Estimates of *p* (True Value .1), Using Likelihoods (14)–(17), for Three True Ascertainment Processes (10), Incorrectly Assuming Ascertainment Is Possible through Both Families in the Pedigree

True Ascertainment and Likelihood[a]	$\hat{p}$	se($\hat{p}$) × $\sqrt{n}$
Complete ascertainment:		
(14) (Assumes complete)	.1210	. . .
(15) (Assumes single)	.0816	. . .
(16) (AAFa)	.1000	.3094
(17) (AAFb)	.1000	. . .
Single ascertainment:		
(14) (Assumes complete)	.1431	. . .
(15) (Assumes single)	.1000	.2184
(16) (AAFa)	.1000	.3027
(17) (AAFb)	.1000	.3012
Quadratic ascertainment:		
(14) (Assumes complete)	.1772	. . .
(15) (Assumes single)	.1287	. . .
(16) (AAFa)	.1000	.2925
(17) (AAFb)	.0950	. . .

[a] Only one family ascertainable.

tion on all the data relevant to ascertainment. Of the classical approaches, only in the case where single ascertainment is correctly assumed are asymptotically unbiased estimators obtained.

We note from tables 3 and 4 that, in some cases, asymptotically unbiased estimators are achieved by classical estimators (even though these tables refer to cases where the data relevant to ascertainment are not correctly specified). The reason for this is discussed in detail by one of us elsewhere (Shute 1988).

Conclusion

The methods described here suggest a possible resolution to the ascertainment problem for pedigree data by extending the methods that we have published elsewhere (Ewens and Shute 1986*b*), that is, by conditioning the likelihood contribution from each pedigree on that part of the data in the pedigree that is relevant to ascertainment. We agree that in practice it might be difficult to identify this part of the data; however, this is a difficulty common to every ascertainment correction procedure. The ultimate choice of the method used for estimating genetic parameters will depend on the size of bias or standard error the investigator is willing to accept and on his knowledge of the true ascertainment procedure. While the example used here to illustrate the theory is not realistic, it gives information concerning the relative sizes of these biases and standard errors and about various problems that can arise when sampling pedigrees (rather than nuclear families). A more realistic example, considering HLA-linked diseases, will be considered in a later paper.

References

Ewens, W. J., and N. C. E. Shute. 1986*a*. The limits of ascertainment. Ann. Hum. Genet. **50**:399–402.

———. 1986*b*. A resolution of the ascertainment sampling problem. I. Theory. Theor. Popul. Biol. **30**:388–214.

Ewens, W. J., N. C. E. Shute, N. J. Cox, R. A. Price, and R. S. Spielman. 1986. Ascertainment considerations in the analysis of affected sib shared haplotype data. Genet. Epidemiol. **1**:[Suppl.] 319–322.

Fisher, R. S. 1934. The effect of methods of ascertainment upon the estimation of frequencies. Ann. Eugenics **6**:13–25.

Morton, N. E. 1984. Trials of segregation analysis. Pp. 83–107 *in* A. Chakravarti, ed. Human population genetics. The Pittsburgh symposium. Van Nostrand, New York.

Shute, N. C. E. 1988. Statistical and ascertainment problems in human genetics. Ph.D. thesis, Monash University, Melbourne.

Shute, N. C. E., and W. J. Ewens. 1988. A resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am. J. Hum. Genet. **43**:374–386.

Weinberg, W. 1928. Mathematische Grundlagen der Probandenmethode. Z. Indukive Abstammungs-und Vererbungslehre **48**:178–228.