Am. J. Hum. Genet. 43:374–386, 1988

A Resolution of the Ascertainment Sampling Problem. II. Generalizations and Numerical Results

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

The ascertainment problem arises when families are sampled by a nonrandom process and some assumption about this sampling process must be made in order to estimate genetic parameters. Under classical ascertainment assumptions, estimation of genetic parameters cannot be separated from estimation of the parameters of the ascertainment process, so that any misspecification of the ascertainment process causes biases in estimation of the genetic parameters. Ewens and Shute proposed a resolution to this problem, involving conditioning the likelihood of the sample on the part of the data which is "relevant to ascertainment." The usefulness of this approach can only be assessed by examining the properties (in particular, bias and standard error) of the estimates which arise by using it for a wide range of parameter values and family size distributions and then comparing these biases and standard errors with those arising under classical ascertainment procedures. These comparisons are carried out in the present paper, and we also compare the proposed method with procedures which condition on, or ignore, parts of the data.

Introduction

Perhaps the major initial aim of the successful series of Genetic Analysis workshops, held regularly since 1982, was to calibrate various estimation procedures and computer packages used in genetic epidemiology, in particular by noting whether they give correct parameter estimates when used on artificial "data" calculated when the true parameter values are known (MacCluer et al. 1983, 1984, 1985). A further aim is to assess the standard errors of these estimates. In line with this procedure, which we strongly support, we present in the present paper calibration properties of the ascertainment-assumption-free (AAF) estimation method of Ewens and Shute (1986*a*), as well as of analogous estimation methods, when the data analyzed arise by ascertainment sampling.

In any ascertainment procedure, the likelihood from which parameter estimates are obtained is the conditional likelihood of the data (genetic, phenotypic, and, if relevant, proband status) of the individuals in the families sampled, given the fact of ascertainment of each family. This leads to a likelihood of the form

$$L = \prod_m \prod_d \left[\frac{P_m(\text{asc},d)}{P_m(\text{asc})}\right]^{n(m,d)}, \qquad (1)$$

where m denotes the number of children in each family, d is the data in each family, $n(m,d)$ is the number of families in the sample having m children and data d, P_m (asc, d) is the probability that a family having m children is ascertained and has data d, and P_m (asc) is the probability that a family with m children is ascertained.

To calculate these probabilities we must assume some model for the ascertainment process. The "classical" model of Weinberg (1928) and Fisher (1934) employs the concept of a proband, that is, an affected individual who is assumed, independently of any other affected sibling, to report with the disease. The classical model, in its simplest form, assumes that the potential probands are the affected children in each family and that each affected child, irrespective of its birth order or which family it belongs to, has the same (unknown)

Received May 7, 1987; final revision received May 9, 1988.

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

0002-9297/88/4304-0003$02.00

probability π of becoming a proband. Thus, given k affected children in a family, this model assumes that

$$P(\text{asc}|k) = 1 - (1 - \pi)^k , \quad (2)$$

so that the denominator in (1) is

$$P_m(\text{asc}) = \sum_k Q_m(k) [1 - (1 - p)^k] , \quad (3)$$

where $Q_m(k)$ is the population probability that a family having m children has k of these affected. Complete ascertainment is the special case of (2) with $\pi = 1$, while single ascertainment is the special case of (2) with $\pi \cong 0$.

The classical model was attacked in the early European literature as embodying unrealistic proband assumptions (see Stene, in press), by Stene himself, by Greenberg (1986), and by Ewens and Shute (1986*a*). It nevertheless persists in many computer packages, and in assessing the properties of the AAF estimation procedure we will assess also properties of the classical approach.

There are several variants of the classical procedure. The most efficient uses as data the number of probands in each family, leading to a term $\pi^A(1 - \pi)^{K-A}$ in the numerator in (1), where K is the total number of affected children in the sample and A is the total number of probands. A far less efficient approach is to use as a datum only the fact that each ascertained family has at least one proband. This leads to a term of the form

$$\prod_k [1 - (1 - \pi)^k]^{n(k)} \quad (4)$$

in the numerator of (1), where $n(k)$ is the number of families in the sample having k affected children. Our calculations show that use of (4), rather than use of the binomial term $\pi^A(1 - \pi)^{K-A}$, decreases the efficiency of parameter estimation by a factor of 50 or 100: these calculations are confirmed by Stene (in press). Nevertheless, (4) is used in several computer programs, for example, POINTER (Lalouel and Morton 1981, eq. [3]).

An alternative classical approach, partly motivated by the large standard errors which result from use of (4), is to estimate genetic parameters assuming, respectively, $\pi = 1$ (complete ascertainment) and $\pi \cong 0$ (single ascertainment) as limiting cases, in the hope of bracketing true parameter values. However (Ewens and Shute 1986*b*), these are not limiting cases. Complete and single ascertainment are best thought of as corresponding to

$$\text{Prob}(\text{asc}|k) = \text{const} \times k^\alpha , \quad (5)$$

which $\alpha = 0$ for complete ascertainment and $\alpha = 1$ for single ascertainment. But it is quite possible for (5) to hold with, for example, $\alpha = 2$: this would occur for data arrived at after two consecutive rounds of single ascertainment. This appears to be approximately the case for the data analyzed in Genetics Analysis Workshop IV (Ewens et al. 1986), with one round of ascertainment between families and physicians and a second between physicians and Genetic Analysis workshop data collection. Clearly, the case $\alpha = 2$, which we describe as quadratic ascertainment, lies outside the complete-to-single range.

It has been argued that quadratic ascertainment can be allowed for under the classical scheme by allowing negative values of π, for example, $\pi = -1$. There are at least three reasons why we disagree with this view. First, no value of π can make (2) a quadratic function of k. Second, negative π values can only be used if the inefficient likelihood contribution (4), rather than the correct binomial formula, is used. Finally, the real problem, especially for Genetic Analysis workshop data, is not that the data come exactly from a quadratic ascertainment process but rather that they probably come from complex ascertainment processes which cannot be described by simple models such as that leading to (2). The AAF method recognizes this and replaces (2) by an arbitrary function $a_m(k)$, whose mathematical form is unspecified. Thus, no specific ascertainment assumption is made under this approach. The gain in universality of application must be balanced by a potential loss through an increased standard error in parameter estimates compared with those arising when, say, complete ascertainment is the case and is correctly assumed by the investigator. Part of the calibration of the AAF approach, to which we now turn, is to calculate this increase over a wide and representative range of parameter values.

Likelihoods

We have denoted above the number of affected children in a family by k. We also denote the number of affected parents by g and use the symbol i as an index to denote all the genetic data in a family.

Assuming complete ascertainment, the likelihood used for parameter estimation is (1) *Classical likelihood, assuming complete ascertainment*:

$$L_c = \prod_m \prod_k \prod_i \prod_g \left[\frac{Q_m(k,i,g)}{Q_m}\right]^{n(m,k,i,g)} , \quad (6)$$

where $Q_m(k,i,g)$ is the population probability that a family of size m has data $\{k,i,g\}$ and $Q_m = \sum_k \sum_i \sum_g Q_m(k,i,g)$.

Assuming single ascertainment, the appropriate likelihood is (2) *Classical likelihood, assuming single ascertainment:*

$$L_S = \prod_m \prod_k \prod_i \prod_g \left[\frac{k\, Q_m(k,i,g)}{Q_m^*}\right] , \quad (7)$$

where $Q_m^* = \sum_k kQ_m(k)$ and $Q_m(k) = \sum_i \sum_g Q_m(k,i,g)$.

Ewens and Shute (1986*a*) show that the AAF likelihood is (3) *Ascertainment-assumption-free (AAF) likelihood:*

$$L_{AAF} = \prod_m \prod_k \prod_i \prod_g \left[\frac{Q_m(k,i,g)}{Q_m(k)}\right] . \quad (8)$$

In effect, this likelihood uses the conditional probability of the number of affected parents g and the genetic information i in any family, given the number of affected children k. An even more extreme form of conditioning, used by Winter (1980) and Risch (1984), used a likelihood calculated from the conditional probability of the genetic data i, given both g and k. This likelihood is (4) *Conditioning on all phenotypes (CAP) likelihood*:

$$L_{CAP} = \prod_m \prod_k \prod_i \prod_g \left[\frac{Q_m(k,i,g)}{Q_m(k,g)}\right] , \quad (9)$$

where $Q_m(k,g) = \sum_i Q_m(k,i,g)$.

Our aim is to compare properties of estimates arising from the four likelihoods (6), (7), (8), and (9) over a wide and representative range of parameter values. To do this, some form of data must be assumed. We follow here the approach of Morton (1984) and use "deterministic" data, that is, $n(m,k,i,g)$ values which are at their expected values for a specified choice of genetic parameters, family size distribution, and ascertainment procedure. Thus, we have not used simulated data, since this would lead to a great increase in computer time, although we hope to do so later. Thus, by "bias" we mean, from now on, the asymptotic (large-sample-size) bias with such "deterministic" data, by "unbiased" we mean "asymptotically unbiased," and by standard error we mean the asymptotic values calculated from an information matrix. We do not know at what sample sizes these asymptotic values become reasonably accurate, but we suspect that for the models considered below several hundred families are sufficient.

Example

We illustrate the calibration of the AAF ascertainment procedure by considering the estimation of parameters associated with a disease determined in part by the genes at a single "disease susceptibility" locus S linked to the HLA complex, with recombination fraction R between this locus and HLA. The S locus admits a susceptibility allele S and a normal allele s and the penetrances of the three genotypes are:

genotype	SS	Ss	ss
penetrance	x	λx	0 .

The population frequency of the disease allele is denoted p. There are thus four unknown parameters—x, p, λ, and R—to be estimated. The population prevalence of the disease, sometimes known and sometimes unknown, is

$$p^2x + 2p(1-p)\lambda x , \quad (10)$$

and, clearly, when the prevalence is known, only three parameters need be estimated.

We assume that the data relate to nuclear families ascertained through affected children. The numbers k of affected children and g of affected parents in each family are known, while the genetic information in each family is assumed to be the HLA haplotype sharing pattern of the affected children. Each such pattern is described by k and an index number i (a list of these indices and the corresponding HLA sharing pattern among affected sibs is given by Ewens and Clarke [1984]). Since HLA haplotype sharing patterns of affected children only are used, only families with two or more affected children are ascertained.

Details of the calculation of $Q_m(k,g,i)$, Q_m, $Q_m(k)$, and $Q_m(k,g)$ for this example (to be used in [6]–[9]) are given by Ewens and Shute (1986*a*) and are not repeated here. The deterministic "data" used in the calibration depend on the numerical values of x, p, λ, and R, on the family size distribution, and on the true (but unknown) ascertainment procedure. We consider various parameter combinations and family size distributions below. So far as the true ascertainment scheme

is concerned, we consider three cases: (i) complete ascertainment, where the sample probability that a family is in category (m,k,g,i) is proportional to the population frequency $Q_m(k,g,i)$; (ii) single ascertainment, where the sample probability is proportional to $kQ_m(k,g,i)$; and (iii) quadratic ascertainment, where the sample probability is proportional to $k^2Q_m(k,g,i)$.

Results

We have estimated parameter values, using, respectively, the four estimation procedures corresponding to the likelihoods (6)–(9), using "deterministic" data corresponding to complete, single, and quadratic ascertainment ([i], [ii], and [iii] above), for approximately 30 parameter combinations, covering all possible values of the parameters x, p, λ, and R but focussing on those values (e.g., small p) which are à priori likely in practice. We describe our initial conclusions in two ways, first by displaying (in table 1) estimates for representative parameter combinations and second by summarizing verbally the conclusions for all 30 combinations.

Table 1 gives the estimates of parameters for six representative parameter combinations when the family size distribution is

m	2	3	4
Prob	.6	.275	.125

(11)

Estimates are given when the prevalence (10) is known and used in the estimation procedure and also when it is unknown.

The main conclusion to be drawn from table 1 is that, in all cases, the AAF approach yields unbiased estimates of all parameters, while the classical approach only yields unbiased estimates when the true ascertainment assumption happens to be made. These conclusions hold for all parameter combinations we considered. The standard errors of parameter estimates are, as expected, smallest when the classical approach is used and the correct ascertainment scheme happens to be assumed in the analysis. However, the standard errors of the AAF estimates are often not much larger than those arising under the classical approach: when the prevalence is known (the most common case in practice), the average increase in standard error, compared with the classical complete ascertainment estimation procedure for complete ascertainment "data", is about 15% for the values in all parameter combinations we have considered, while for quadratic "data" the average bias of the classical complete ascertainment estimator is about 43%. One would surely accept an increase in standard error of about 15% to protect against the possibility of a bias of this order.

Apart from the major conclusion noted in the last paragraph, further, more specific conclusions may be noted from table 1. First, when x is small the standard errors of all estimators are larger than when x is large; however, larger standard errors arise with large values of p, λ, and R than with small values. Second, the largest standard errors (for all methods of estimation) tend to occur for intermediate values of λ, smaller standard errors arising for small and large λ. Finally, when the data arise from a complete ascertainment process but single ascertainment is assumed, then, when $\lambda < .5$, $\hat{x} < x$, $\hat{p} > p$, $\hat{\lambda} > \lambda$, $\hat{R} < R$. When the data arise from a single ascertainment process but complete ascertainment is assumed, $\hat{x} > x$, $\hat{p} < p$, $\hat{\lambda} < \lambda$, $\hat{R} > R$. However, when $\lambda > .5$, all these inequalities are reversed.

The conclusions drawn from table 1 typify those for all parameter combinations we examined—there are no peculiar combinations for which contrary conclusions apply.

Two remarks about the estimates arising when one conditions on all affectedness data, using (9), are in order. First, the standard errors of estimates of x are often very large, particularly when the population prevalence is unknown, being often approximately $10/\sqrt{n}$ when the prevalence is known and $50/\sqrt{n}$ when it is unknown. With these values, one would need, respectively, a sample of 400, and 10,000 families before conventional plus and minus two standard deviation limits lie within the values 0 and 1, which can in any event be set à priori as natural limits for x.

The second comment concerns the standard errors of the estimates of λ in table 1. In a small number of cases the standard errors of these estimates, when one conditions on all affectedness information, are smaller than the corresponding standard errors when one conditions only on the affectedness status of the children. Although there is no mathematical theorem known to us preventing this occurring, it is quite unexpected intuitively that it should. We propose to examine this phenomenon elsewhere.

The conclusions just noted all apply if the family size distribution is as given in (11). We checked that they continue to apply for other family size distributions. Table 2 lists parameter estimates corresponding to those given for data set 1 and data set 2 of table 1, when the family size distribution is

m	2	3	4
Prob	.3	.35	.35

(12)

Table 1

Maximum Likelihood Estimates of x, p, λ, and R (with Standard Errors $\times \sqrt{n}$ of Unbiased Estimators) for Various Parameter Combinations (i.e., Data Sets), Various Ascertainment Processes, and Various Ascertainment Assumptions

A. Data Set 1: $x = .50$, $p = .04$, $\lambda = .08$, $R = .025$

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (3.44)	.04 (.48)	.08 (.40)	.025 (.38)
	(7)	.312	.074	.104	.001
	(8)	.50 (7.46)	.04 (.84)	.08 (.76)	.025 (.66)
	(9)	.50 (31.4)	.04 (1.09)	.08 (.79)	.025 (.77)
Single	(6)	.754	.024	.060	.042
	(7)	.50 (2.97)	.04 (.44)	.08 (.37)	.026 (.36)
	(8)	.50 (6.74)	.04 (.76)	.08 (.70)	.025 (.63)
	(9)	.50 (28.9)	.04 (1.03)	.08 (.73)	.025 (.71)
Quadratic	(6)	1.000	.019	.043	.060
	(7)	.752	.025	.059	.043
	(8)	.50 (5.96)	.04 (.67)	.08 (.64)	.025 (.59)
	(9)	.50 (28.5)	.04 (1.03)	.08 (.69)	.025 (.66)
Prevalence known (= .00387):					
Complete	(6)	.50 (2.03)	.04 (.16)	.08 (.27)	.025 (.30)
	(7)	.401	.045	.090	.017
	(8)	.50 (2.80)	.04 (.18)	.08 (.32)	.025 (.34)
	(9)	.50 (13.5)	.04 (.70)	.08 (.74)	.025 (.73)
Single	(6)	.630	.036	.069	.036
	(7)	.50 (1.87)	.04 (.16)	.08 (.26)	.025 (.29)
	(8)	.50 (2.76)	.04 (.18)	.08 (.32)	.025 (.34)
	(9)	.50 (12.0)	.04 (.62)	.08 (.68)	.025 (.68)
Quadratic	(6)	.820	.034	.054	.053
	(7)	.645	.036	.068	.037
	(8)	.50 (2.71)	.04 (.19)	.08 (.32)	.025 (.34)
	(9)	.50 (10.3)	.04 (.63)	.08 (.61)	.025 (.63)

B. Data Set 2: $x = .50$, $p = .04$, $\lambda = .40$, $R = .025$

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (9.00)	.04 (.91)	.40 (7.76)	.025 (.85)
	(7)	.402	.046	.474	.019
	(8)	.50 (9.71)	.04 (.95)	.40 (8.31)	.025 (.88)
	(9)	.50 (38.4)	.04 (1.23)	.40 (10.0)	.025 (1.46)
Single	(6)	.936	.023	.211	.045
	(7)	.50 (7.83)	.04 (.83)	.40 (6.78)	.025 (.78)
	(8)	.50 (8.54)	.04 (.86)	.40 (7.32)	.025 (.82)
	(9)	.50 (34.0)	.04 (1.12)	.40 (9.03)	.025 (1.27)
Quadratic	(6)	1.00	.036	.162	.050
	(7)	.940	.024	.211	.045
	(8)	.50 (7.36)	.04 (.78)	.40 (6.34)	.025 (.75)
	(9)	.50 (30.7)	.04 (1.01)	.40 (8.06)	.025 (1.12)

(*continued*)

Table 1 (continued)

B. Data Set 2: $x = .50$, $p = .04$, $\lambda = .40$, $R = .025$					
Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence known (= .01616):					
Complete	(6)	.50 (3.56)	.04 (0.13)	.40 (3.93)	.025 (.26)
	(7)	.420	.042	.455	.022
	(8)	.50 (3.82)	.04 (0.13)	.40 (4.09)	.025 (.27)
	(9)	.50 (22.0)	.04 (1.09)	.40 (9.02)	.025 (1.32)
Single	(6)	.671	.041	.284	.030
	(7)	.50 (3.29)	.04 (.13)	.40 (3.69)	.025 (.26)
	(8)	.50 (3.61)	.04 (.13)	.40 (3.88)	.025 (.27)
	(9)	.50 (19.3)	.04 (.97)	.40 (8.14)	.025 (1.19)
Quadratic	(6)	.890	.047	.177	.043
	(7)	.702	.041	.270	.031
	(8)	.50 (3.35)	.04 (.13)	.40 (3.62)	.025 (.27)
	(9)	.50 (16.1)	.04 (.82)	.40 (7.10)	.025 (1.03)

C. Data Set 3: $x = .10$, $p = .04$, $\lambda = .08$, $R = .025$					
Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.10 (1.70)	.04 (1.16)	.08 (.86)	.025 (.79)
	(7)	.061	.078	.102	.001
	(8)	.10 (3.86)	.04 (2.30)	.08 (1.72)	.025 (1.56)
	(9)	.10 (66.1)	.04 (2.40)	.08 (1.66)	.026 (1.75)
Single	(6)	.168	.019	.059	.041
	(7)	.10 (1.42)	.04 (1.02)	.08 (.76)	.025 (.71)
	(8)	.10 (3.44)	.04 (2.04)	.08 (1.54)	.025 (1.41)
	(9)	.10 (64.1)	.04 (2.26)	.08 (1.49)	.025 (1.55)
Quadratic	(6)	.273	.009	.043	.051
	(7)	.170	.019	.059	.041
	(8)	.10 (3.16)	.04 (1.86)	.08 (1.42)	.025 (1.36)
	(9)	.10 (63.1)	.04 (2.17)	.08 (1.35)	.025 (1.36)
Prevalence known (= .00077):					
Complete	(6)	.10 (.88)	.04 (0.28)	.08 (.35)	.025 (.37)
	(7)	.078	.047	.086	.019
	(8)	.10 (1.17)	.04 (.34)	.08 (.40)	.025 (.41)
	(9)	.10 (7.67)	.04 (2.08)	.08 (1.61)	.025 (1.52)
Single	(6)	.127	.035	.074	.031
	(7)	.110 (.81)	.04 (.27)	.08 (.34)	.025 (.36)
	(8)	.10 (1.17)	.04 (.34)	.08 (.40)	.025 (.41)
	(9)	.10 (6.85)	.04 (1.86)	.08 (1.45)	.025 (1.37)
Quadratic	(6)	.169	.029	.066	.038
	(7)	.133	.034	.073	.032
	(8)	.10 (1.16)	.04 (.34)	.08 (0.40)	.025 (.41)
	(9)	.10 (5.90)	.04 (1.60)	.08 (1.26)	.025 (1.21)

(continued)

Table 1 (continued)

D. Data Set 4: $x = .50, p = .09, \lambda = .08, R = .025$					
Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (2.85)	.09 (.80)	.08 (.53)	.025 (.58)
	(7)	.351	.134	.093	.000
	(8)	.50 (6.14)	.09 (1.40)	.08 (.79)	.025 (1.03)
	(9)	.50 (24.7)	.09 (2.30)	.08 (.92)	.025 (1.25)
Single	(6)	.740	.059	.057	.055
	(7)	.50 (2.46)	.09 (.72)	.08 (.51)	.025 (.54)
	(8)	.50 (5.47)	.09 (1.25)	.08 (.76)	.025 (.97)
	(9)	.50 (22.5)	.09 (2.25)	.08 (.90)	.025 (1.13)
Quadratic	(6)	.985	.046	.041	.077
	(7)	.738	.060	.055	.058
	(8)	.50 (4.78)	.09 (1.09)	.08 (.72)	.025 (.91)
	(9)	.50 (20.5)	.09 (2.18)	.08 (.88)	.025 (1.04)
Prevalence known (= .01060):					
Complete	(6)	.50 (1.84)	.09 (.40)	.08 (.45)	.025 (.41)
	(7)	.404	.101	.089	.012
	(8)	.50 (2.41)	.09 (.42)	.08 (.47)	.025 (.45)
	(9)	.50 (11.4)	.09 (1.23)	.08 (.83)	.025 (1.23)
Single	(6)	.628	.081	.070	.040
	(7)	.50 (1.70)	.09 (.39)	.08 (.44)	.025 (.39)
	(8)	.50 (2.35)	.09 (.42)	.08 (.47)	.025 (.44)
	(9)	.50 (9.85)	.09 (1.06)	.08 (.78)	.025 (1.13)
Quadratic	(6)	.823	.073	.057	.059
	(7)	.643	.080	.068	.042
	(8)	.50 (2.27)	.09 (.41)	.08 (.46)	.025 (.44)
	(9)	.50 (8.50)	.09 (.91)	.08 (.74)	.025 (1.04)

E. Data Set 5: $x = .50, p = .04, \lambda = .08, R = .12$					
Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (4.89)	.04 (.76)	.08 (.61)	.12 (.60)
	(7)	.297	.079	.118	.090
	(8)	.50 (12.9)	.04 (1.42)	.08 (1.71)	.12 (1.32)
	(9)	.50 (56.6)	.04 (1.92)	.08 (1.94)	.12 (1.51)
Single	(6)	.727	.027	.055	.140
	(7)	.50 (4.37)	.04 (.70)	.08 (.57)	.12 (.56)
	(8)	.50 (11.3)	.04 (1.24)	.08 (1.60)	.12 (1.26)
	(9)	.50 (52.4)	.04 (1.85)	.08 (1.87)	.12 (1.44)
Quadratic	(6)	.976	.021	.038	.156
	(7)	.717	.027	.054	.142
	(8)	.50 (10.9)	.04 (1.18)	.08 (1.58)	.12 (1.27)
	(9)	.50 (49.2)	.04 (1.79)	.08 (1.82)	.12 (1.39)

(*continued*)

Table 1 (continued)

E. Data Set 5: $x = .50$, $p = .04$, $\lambda = .08$, $R = .12$

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence known (= .00387):					
Complete	(6)	.50 (2.54)	.04 (.33)	.08 (.60)	.12 (.60)
	(7)	.444	.035	.109	.100
	(8)	.50 (2.88)	.04 (.35)	.08 (.86)	.12 (.78)
	(9)	.50 (25.4)	.04 (1.24)	.08 (1.80)	.12 (1.50)
Single	(6)	.585	.044	.056	.142
	(7)	.50 (2.46)	.04 (.32)	.08 (.57)	.12 (.56)
	(8)	.50 (2.85)	.04 (.35)	.08 (.85)	.12 (.76)
	(9)	.50 (22.7)	.04 (1.10)	.08 (1.70)	.12 (1.44)
Quadratic	(6)	.761	.042	.042	.156
	(7)	.597	.043	.056	.142
	(8)	.50 (2.81)	.04 (.34)	.08 (.83)	.23 (.73)
	(9)	.50 (20.1)	.04 (.96)	.08 (1.61)	.12 (1.39)

F. Data Set 6: $x = .50$, $p = .09$, $\lambda = .18$, $R = .12$

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (6.53)	.09 (2.25)	.18 (1.57)	.12 (.95)
	(7)	.299	.170	.293	.054
	(8)	.50 (9.49)	.09 (2.74)	.18 (2.86)	.12 (1.61)
	(9)	.50 (48.5)	.09 (3.79)	.18 (3.58)	.12 (2.50)
Single	(6)	.756	.059	.106	.161
	(7)	.50 (5.97)	.09 (2.11)	.18 (1.43)	.12 (.88)
	(8)	.50 (8.66)	.09 (2.56)	.18 (2.65)	.12 (1.50)
	(9)	.50 (43.4)	.09 (3.53)	.18 (3.27)	.12 (2.18)
Quadratic	(6)	.997	.048	.068	.188
	(7)	.737	.063	.103	.164
	(8)	.50 (7.82)	.09 (2.38)	.18 (2.42)	.12 (1.39)
	(9)	.50 (39.4)	.09 (3.31)	.18 (3.02)	.12 (1.92)
Prevalence known (= .01879):					
Complete	(6)	.50 (1.89)	.09 (.71)	.18 (1.50)	.12 (.58)
	(7)	.441	.073	.277	.105
	(8)	.50 (1.96)	.09 (.82)	.18 (2.08)	.12 (.71)
	(9)	.50 (26.2)	.09 (3.36)	.18 (3.54)	.12 (2.41)
Single	(6)	.530	.113	.113	.149
	(7)	.50 (1.87)	.09 (.69)	.18 (1.40)	.12 (.56)
	(8)	.50 (1.95)	.09 (.81)	.18 (2.03)	.12 (.70)
	(9)	.50 (22.6)	.09 (2.96)	.18 (3.18)	.12 (2.13)
Quadratic	(6)	.646	.110	.086	.170
	(7)	.538	.114	.110	.152
	(8)	.50 (1.93)	.09 (.80)	.18 (1.96)	.12 (.69)
	(9)	.50 (19.2)	.09 (2.59)	.18 (2.86)	.12 (1.88)

NOTE.—See text for details. Family size distribution is (11).

Table 2

Maximum Likelihood Estimates of x, p, λ, and R (with Standard Errors $\times \sqrt{n}$ of Unbiased Estimators) for Family Size Distribution (12)

A. Data Set 1: $x = .50$, $p = .04$, $\lambda = .08$, $R = .025$

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (2.95)	.04 (.43)	.08 (.36)	.025 (.35)
	(7)	.311	.072	.105	.001
	(8)	.50 (6.81)	.04 (.74)	.08 (.71)	.025 (.61)
	(9)	.50 (29.7)	.04 (1.11)	.08 (.74)	.025 (.69)
Single	(6)	.747	.026	.060	.042
	(7)	.50 (2.55)	.04 (.39)	.08 (.34)	.025 (.34)
	(8)	.50 (5.99)	.04 (.65)	.08 (.64)	.025 (.58)
	(9)	.50 (27.5)	.04 (1.06)	.08 (.68)	.025 (.64)
Quadratic	(6)	1.000	.021	.043	.061
	(7)	.740	.026	.059	.044
	(8)	.50 (5.26)	.04 (.57)	.08 (.59)	.025 (.55)
	(9)	.50 (25.6)	.04 (1.02)	.08 (.64)	.025 (.60)
Prevalence known (= .00387):					
Complete	(6)	.50 (1.90)	.04 (.16)	.08 (.26)	.025 (.29)
	(7)	.384	.046	.092	.015
	(8)	.50 (2.97)	.04 (.18)	.08 (.33)	.025 (.34)
	(9)	.50 (11.8)	.04 (.60)	.08 (.69)	.025 (.66)
Single	(6)	.655	.036	.067	.038
	(7)	.50 (1.73)	.04 (.16)	.08 (.25)	.025 (.28)
	(8)	.50 (2.90)	.04 (.18)	.08 (.33)	.025 (.34)
	(9)	.50 (10.2)	.04 (.52)	.08 (.62)	.025 (.62)
Quadratic	(6)	.868	.034	.051	.057
	(7)	.663	.036	.066	.039
	(8)	.50 (2.80)	.04 (.18)	.08 (.33)	.025 (.34)
	(9)	.50 (8.83)	.04 (.45)	.08 (.56)	.025 (.58)

B. Data Set 2: $x = .50$, $p = .04$, $\lambda = .40$, $R = .025$

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	(6)	.50 (7.89)	.04 (.81)	.40 (6.83)	.025 (.74)
	(7)	.388	.045	.482	.019
	(8)	.50 (8.63)	.04 (.84)	.40 (7.39)	.025 (.77)
	(9)	.50 (38.9)	.04 (1.16)	.40 (9.13)	.025 (1.32)
Single	(6)	.928	.026	.213	.043
	(7)	.50 (6.72)	.04 (.74)	.40 (5.84)	.025 (.69)
	(8)	.50 (7.45)	.04 (.76)	.40 (6.40)	.025 (.72)
	(9)	.50 (34.8)	.04 (1.06)	.40 (8.09)	.025 (1.16)
Quadratic	(6)	1.000	.042	.155	.052
	(7)	.901	.027	.223	.043
	(8)	.50 (6.40)	.04 (.69)	.40 (5.51)	.025 (.67)
	(9)	.50 (30.6)	.04 (.96)	.40 (7.05)	.025 (1.03)

(*continued*)

Table 2 (continued)

Data	Likelihood	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
B. Data Set 2: $x = .50$, $p = .04$, $\lambda = .40$, $R = .025$					
Prevalence known (= .01616):					
Complete	(6)	.50 (3.50)	.04 (.12)	.40 (3.83)	.025 (.26)
	(7)	.397	.043	.472	.020
	(8)	.50 (3.89)	.04 (.13)	.40 (4.07)	.025 (.26)
	(9)	.50 (19.1)	.04 (.93)	.40 (8.32)	.025 (1.15)
Single	(6)	.725	.041	.260	.032
	(7)	.50 (3.13)	.04 (.12)	.40 (3.51)	.026 (.26)
	(8)	.50 (3.59)	.04 (.12)	.40 (3.77)	.025 (.26)
	(9)	.50 (16.4)	.04 (.81)	.40 (7.34)	.025 (1.02)
Quadratic	(6)	.966	.047	.162	.049
	(7)	.738	.041	.260	.033
	(8)	.50 (3.24)	.04 (.12)	.40 (3.44)	.025 (.26)
	(9)	.50 (13.8)	.04 (.70)	.40 (6.35)	.025 (.91)

It is clear that the general conclusions of table 1 continue to apply: we have again checked this result for a wide range of parameter combinations.

Because family size distribution (12) tends to emphasize larger families more than does family size distribution (11), we would expect smaller standard errors of (asymptotically unbiased) estimators to arise under (12), for all methods of estimation. The effect of family size distribution on bias appears to be quite complex, although when the population prevalence in known, larger biases uniformly arise for the family size distribution (12). We would expect this conclusion since for family size distribution (11) approximately 96% of families sampled have two affected children, so that assumptions about the ascertainment process are relatively unimportant.

We considered next the effects of two other procedures used in parameter estimation in ascertainment sampling. The first of these concerns conditioning on parental affectedness data. The AAF procedure described in preceding sections is a specific case of the procedure described by Ewens and Shute (1986*a*), where the AAF procedure conditions on the "data relevant to ascertainment." In the above it has been assumed that these data concern the number of affected children in each family, so that the AAF procedure described above conditions (see [8]) on the number of affected children.

Some estimation procedures condition on data different from that "relevant to ascertainment" (e.g., POINTER; see Lalouel and Morton 1981). In these procedures the conditioning is on parental phenotype, with the (reasonable) aim of removing any potential bias due to fertility differences between affected and nonaffected individuals. If ascertainment is through affected children and the investigator conditions on parental phenotype only, then the ascertainment problem will not be removed and biased estimates will be obtained if an incorrect ascertainment assumption is made.

The likelihood when conditioning is on the number of affected parents is

$$L = \prod_m \prod_k \prod_g \prod_i [P_m(\text{asc},k,g,i)/P_m(\text{asc},g)]^{n(m,k,g,i)}, \quad (13)$$

where $P_m(\text{asc},g)$ is the probability that a family of size m has g affected parents and is ascertained. The conditioning in (13) is not on the data "relevant to ascertainment" if these data concern the number of affected children in the family, and thus it can be expected to lead to biased estimates under incorrect ascertainment assumptions. We confirmed this by calculating deterministic "data" from complete, single, and quadratic ascertainment processes, assuming the family size distribution (11) and parameter values

$$x = .5,\ p = .04,\ \lambda = .08,\ R = .025. \quad (14)$$

These data were then analyzed using the likelihood (13),

Table 3

Estimates of x, p, λ, and R (True Values Given in [14]), Together with Standard Errors of Unbiased Estimates, When Ascertainment Is through Children and Likelihood (13) (Conditioning on Parental Affectedness Status) Is Used

Data	Likelihood (13) Assuming	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$	$\hat{R}$
Prevalence unknown:					
Complete	Complete	.50 (3.75)	.04 (.95)	.08 (.79)	.025 (.62)
	Single	.349	.037	.080	.023
Single	Complete	.698	.046	.081	.028
	Single	.50 (3.28)	.040 (.85)	.08 (.73)	.025 (.59)
Quadratic	Complete	.931	.059	.084	.032
	Single	.697	.046	.081	.028
Prevalence known (= .00387):					
Complete	Complete	.50 (2.90)	.04 (.21)	.08 (.27)	.025 (.30)
	Single	.322	.053	.092	.013
Single	Complete	.746	.032	.069	.036
	Single	.50 (2.49)	.040 (.20)	.08 (.26)	.025 (.29)
Quadratic	Complete	1.000	.028	.057	.049
	Single	.743	.032	.069	.036

assuming, respectively, single and complete ascertainment. The results are presented in table 3.

As expected, the genetic parameter estimates are biased when an incorrect ascertainment assumption is made. When the prevalence is known, the biases in x and p arising from (13) are larger than those arising from the classical likelihoods (6) and (7). The standard errors of unbiased estimates are slightly higher than those arising from (6) and (7) when the prevalence is known but considerably higher for p, λ, and R when it is unknown. However, as expected, these standard errors are much smaller than those arising from the CAP likelihood (9), which conditions on children in addition to parents. Similar results are obtained when "data" arising from family size distribution (12) are analyzed and when other parameter combinations are considered.

We finally considered estimation procedures which ignore part of the data. Consider the data vector (m,k,g,i). The only part of the data that gives information of R is i. Suppose we decide that R is not of major importance and that we want to estimate x, p, and λ only, ignoring i in the data. The data vector then becomes $(m,k,g) = \sum_i (m,k,g,i)$ and, similarly, $P_m(k,g) = \sum_i P_m(k,g,i)$.

Suppose, having estimated x, p, and λ by using data of the form (m,k,g), that a separate estimation procedure for R is carried out, using the estimated x, p, and λ values as "true" values. Is this method for estimating x, p, λ, and R any more or less accurate than the approach which estimates all four parameters simultaneously?

The likelihood equivalent to the AAF likelihood (8), when i is ignored, is

$$L_{ii} = \prod_m \prod_k \prod_g \left[\frac{Q_m(k,g)}{Q_m(k)}\right]^{n(m,k,g)} \qquad (15)$$

(the suffix "ii" denoting "ignoring i"), where $n(m,k,g) = \sum_i n(m,k,i,g)$. Like the likelihood (8), (15) does not depend on a specific ascertainment assumption, so the estimates derived from it should be unbiased whatever the true ascertainment process. Our calculations showed that this is so. We now compare the standard errors arising from this likelihood with the standard errors obtained when i is used and R is also estimated.

The information matrix from which the variances of the new estimates are obtained can be calculated using deterministic "data" and (15) or by a simple subtraction of information matrices of likelihoods that have already been examined. If we denote the information matrices corresponding to use of (8), (9), and (15) by I(8), I(9), and I(15), respectively, it is a simple consequence of the identity

$$\frac{Q_m(k,g)}{Q_m(k)} = \frac{Q_m(k,i,g)}{Q_m(k)} \div \frac{Q_m(k,i,g)}{Q_m(k,g)}$$

that $I(15) = I(8) - I(9)$. Thus, there is less information obtained by using the likelihood (15) than by using (8), and the standard errors of the estimates of the former should be larger than those of the "AAF" estimates arising from the likelihood (8).

This is verified by the results of table 4, which considers both the case where the population prevalence is known and the case where it is unknown. Table 4 shows the standard errors of the estimates of x, p, and λ arising from the likelihood (15), using "data" which assumes the family size distribution (11) and parameter values given in (14). The standard errors arising from (15) are significantly larger than those arising from (8), (table 1) (i.e., significantly larger than those arising when i is used as part of the data). Thus, although x, p, and λ can be estimated without using the information contained in i, the information in i is important in the estimation of x, p, and λ, as well as of R, particularly when the population prevalence is unknown. Hence, estimating x, p, and λ first and then subsequently estimating R using the x, p, and λ estimates as "true" values could produce inaccurate estimates for all parameters, particularly for R (since R is estimated assuming potentially inaccurate estimates of x, p, and λ). Such a procedure for estimating the four genetic parameters is therefore not recommended.

Conclusion

The AAF method produces asymptotically unbiased estimation of all genetic parameters, no matter what the ascertainment procedure might be. The price paid for this universality of application is a modest increase in standard error compared with cases where the true ascertainment process is known and used in the estimation of genetic parameters, averaging some 15% in the wide and representative set of cases we have considered. However, biases often exceeding 50% are possible when an incorrect ascertainment procedure is assumed. These conclusions apply for a wide range of family size distributions.

Conditioning likelihoods on the number of affected parents (in order to remove fertility deficit problems) will not correct for ascertainment if ascertainment is through children—indeed, biases in parameter estimates are increased by this conditioning.

The effects of using subsets of the data to estimate subsets of the parameters is considered. This method leads to estimates with greatly increased standard errors.

Table 4

Standard Errors ($\times \sqrt{n}$) of Estimators Derived from the AAF "Ignoring i" Likelihood (15)

True Ascertainment Process	$\hat{x}$	$\hat{p}$	$\hat{\lambda}$
Prevalence known:			
Complete.....	40.95 (7.46)	5.88 (.84)	4.27 (.76)
Single	55.77 (6.74)	7.92 (.76)	4.09 (.66)
Quadratic	41.48 (5.96)	5.93 (1.03)	3.81 (.59)
Prevalence unknown:			
Complete.....	4.15 (2.80)	1.64 (.18)	4.22 (.32)
Single	4.13 (2.76)	1.55 (.18)	3.97 (.32)
Quadratic	4.16 (2.71)	1.49 (.19)	3.76 (.32)

NOTE.—The numbers in parentheses are corresponding values from table 1 (when i is not ignored). The family size distribution is (11), and the true parameters are (14).

References

Ewens, W. J., and C. P. Clarke. 1984. Maximum likelihood estimation of genetic parameters of HLA-linked diseases using data from various sizes. Am. J. Hum. Genet. **36**: 858–872.

Ewens, W. J., and N. C. E. Shute. 1986*a*. A resolution of the ascertainment sampling problem. I. Theory. Theor. Popul. Biol. **30**:388–412.

———. 1986*b*. The limits of ascertainment. Ann. Hum. Genet. **50**:399–402.

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. [Suppl.] **1**:319–322.

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

Greenberg, D. A. 1986. The effect of proband designation on segregation analysis. Am. J. Hum. Genet. **39**:329–339.

Lalouel, J. M., and N. E. Morton. 1981. Complex segregation analysis with pointers. Hum. Hered. **31**:312–321.

MacCluer, J. W., C. T. Falk, R. S. Spielman and D. K. Wagener. Genetic Analysis Workshop II: summary. Genet. Epidemiol. **1**:147–159.

McCluer, J. W., C. T. Falk, and D. K. Wagener. Genetic Analysis Workshop III: summary. Genet. Epidemiol. **2**:185–198.

MacCluer, J. W., D. K. Wagener, and R. S. Spielman. 1983. Genetic Analysis Workshop I: segregation analysis of simulated data. Am. J. Hum. Genet. **35**:784–792.

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.

Risch, N. 1984. Segregation analysis incorporating linkage

markers. I. Single-locus models with an application of type 1 diabetes. Am. J. Hum. Genet. 36:363–386.
Stene, J. The incomplete, multiple ascertainment model: assumptions, applications and alternative models. Genet. Epidem. (in press).
Weinberg, W. 1928. Mathematische Grundlagen der Probandenmethode. Z. Indukive Abstammungs-und Vererbungslehre 48:178–228.
Winter, R. M. 1980. The estimation of phenotype distributions from pedigree data. Am. J. Med. Genet. 7:537–542.