Am J Hum Genet 38:555–566, 1986

The Effects of a Known Family-Size Distribution on the Estimation of Genetic Parameters

W. J. EWENS,[1,2] SUSAN E. HODGE,[3] AND FOO HOOI PING[1]

SUMMARY

We consider the question: In a segregation analysis, can knowledge of the family-size distribution (FSD) in the population from which a sample is drawn improve the estimators of genetic parameters? In other words, should one incorporate the population FSD into a segregation analysis if one knows it? If so, then under what circumstances? And how much improvement may result?

We examine the variance and bias of the maximum likelihood estimators both asymptotically and in finite samples. We consider Poisson and geometric FSDs, as well as a simple two-valued FSD in which all families in the population have either one or two children. We limit our study to a simple genetic model with truncate selection.

We find that if the FSD is completely specified, then the asymptotic variance of the estimator may be reduced by as much as 5%–10%, especially when the FSD is heavily skewed toward small families. Results in small samples are less clear-cut. For some of the simple two-valued FSDs, the variance of the estimator in small samples of one- and two-child families may actually be *increased* slightly when the FSD is included in the analysis.

If one knows only the statistical form of the FSD, but not its parameter, then the estimator is improved only minutely.

Our study also underlines the fact that results derived from asymptotic maximum likelihood theory do not necessarily hold in small samples.

Received April 5, 1985; revised September 17, 1985.

This work was supported by grant GM-21135 from the National Institutes of Health (W. J. E.); and grants AM-31813 and AM-26844 from the National Institutes of Health and National Institutes of Health Research Career Development Award AM-01145 (S. E. H.).

[1] Department of Mathematics, Monash University, Clayton, Victoria 3168, Australia.

[2] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104.

[3] Departments of Biomathematics and Psychiatry, University of California at Los Angeles Medical School, Center for the Health Sciences, 760 Westwood Plaza, Los Angeles, CA 90024. (To whom reprint requests should be addressed.)

We conclude that in most practical applications it is not worth incorporating the FSD into a segregation analysis. However, this practice may be justified under special circumstances where the FSD is completely specified, without error, and the population consists overwhelmingly of small families.

INTRODUCTION

Suppose, in a sampling scheme whose aim is to estimate genetic parameters, information (either complete or partial) is available about the family-size distribution (FSD) in the population from which the sample is drawn. Then this information can be used [1] to modify the estimates of these parameters. To do so requires using a more cumbersome form of the likelihood (the "Grand-Multinomial," or GM, formulation; see below) than is customary. Since this GM formulation uses FSD information, we might expect it to lead to better estimators than the usual one. In this study, we consider the extent to which information about a known FSD improves the estimators of genetic parameters. We do this by examining the large-sample variances and the small-sample means and variances of various estimators. Our analysis of the asymptotic variances is, in part, an extension of the work of Barrai et al. [2]. For finite sample sizes, we use a combination of exact calculations and simulations. We will show that unexpected values for the exact small-sample variance can arise.

In the model we consider, we will also show, as a subsidiary result, that the asymptotic formula for the variance of the estimates of genetic parameters is sometimes not accurate, when families can have only one or two children, even for samples of 100 families or more.

We give particular attention to a simple two-valued FSD in which all families in the population have either one or two children, since we felt, for reasons given in [1], that the advantages of the GM formulation would be greatest in such a population.

Here, we consider only the situation where our knowledge of the FSD, although perhaps incomplete, is *accurate*. Ewens and Asaba [3] showed that if an incorrect assumption is made for the form of the FSD we can expect a small bias in the estimate of p, together with a rather larger bias (up to about 10%) in the estimated asymptotic variance of this estimate.

For simplicity, we limit our calculations to one simple genetic model, namely, that in which for ascertainable families there is one unknown probability p that a child will be affected. We also assume that every family with at least one affected child is ascertained (truncate selection).

PRELIMINARY RESULTS: GRAND- AND SEPARATE-MULTINOMIAL FORMULATIONS AND KNOWLEDGE OF THE FSD

Before we approach the large- and small-sample behavior of the estimator, we define what we mean by the grand-multinomial and separate-multinomials

formulations. Suppose data are available, in a sample, from families of various sizes. Common practice is to treat the data for any specific family size as coming from some multinomial distribution and then (for maximum likelihood estimation) to find a total likelihood by simply multiplying the likelihoods for each observed family size. We call this the separate-multinomials (SM) method. However, as pointed out by Hodge [1], if we have information on the FSD, this approach is not correct. The entire dataset must be viewed as coming from one more-complicated "grand" multinomial distribution, whose probabilities are determined in part by the FSD. We call estimation of p using a likelihood calculated from this single-distribution grand-multinomial (GM) estimation. Apart from the increased efficiency implied by use of the FSD, this approach also allows use of one-child families in segregation analysis, something not possible under SM estimation. Hodge [1] showed that estimates of genetic parameters under the two approaches are identical if, in fact, nothing is known about the FSD, but not otherwise. Our main aim is to investigate the increased efficiency in parameter estimation when information on the FSD is available and is used (via a GM estimation procedure).

This information will, in practice, take one of two forms: either the complete FSD is known or else the form of the distribution is known (e.g., Poisson) up to the value of an unknown parameter. We may thus distinguish three cases: case A: FSD completely known (GM estimation); case B: FSD known in form, parameter unknown (GM estimation); case C: FSD unknown (SM estimation).

The SM Likelihood

As is well known, the likelihood of the sample, for SM estimation (case C), is

$$L_C = \text{const} \cdot p^R(1 - p)^{S-R} \prod_i [1 - (1 - p)^i]^{-n_i} .$$

Here, n_{ij} is the number of families in the sample that are of size i and have j affected, $R = \Sigma\Sigma j\, n_{ij}$ is the total number of affected children, $S = \Sigma\Sigma i\, n_{ij}$ is the total number of children in the sample, n_i is the total number of families in the sample of size i, and $\Sigma n_i = n$ is the total number of families in the sample.

The GM Likelihood

Suppose, in cases A and B, that the population probability that a family is of size i is $a_i(\theta)$. [The set of values $a_1(\theta), a_2(\theta), \ldots$ is then the FSD.] This notation allows for the possibility of a parameter θ, known in case A but unknown in case B, characterizing the distribution. The GM likelihood is now (Hodge [1])

$$L_A = L_B = \text{const} \cdot p^R(1 - p)^{S-R} [A(p, \theta)]^{-n} \prod_i [a_i(\theta)]^{n_i} , \qquad (1)$$

where R, S, and n_i are as defined above and $A(p, \theta) = \Sigma a_i(\theta)[1 - (1 - p)^i]$.

LARGE-SAMPLE VARIANCES

Case A

When the form $\{a_i(\theta)\}$ of the FSD is known exactly, standard theory ([4], pp. 43–44) shows from equation (1) that the reciprocal of the large-sample variance σ_A^2 of the maximum likelihood estimator (MLE) $\hat{p}$ of p is given by:

$$(\sigma_A^2)^{-1} = \frac{n\mu}{Ap(1-p)} - n\frac{\Sigma i^2 a_i(\theta)(1-p)^{i-2}}{A} - n\frac{[\Sigma i a_i(\theta)(1-p)^{i-1}]^2}{A^2} . \tag{2}$$

Case B

When θ is unknown, it must be estimated simultaneously with p by maximum likelihood estimation. A second differentiation gives an information matrix, which, when inverted, yields the large-sample variance σ_B^2 of $\hat{p}$. We find

$$(\sigma_B^2)^{-1} = (\sigma_A^2)^{-1} - I_p^2\theta/I_{\theta\theta} , \tag{3}$$

where $I_{p\theta}$ and $I_{\theta\theta}$ are terms in this information matrix whose explicit forms we do not present here.

Case C

As is well known [5], the large-sample variance σ_C^2 of $\hat{p}$ is given by

$$(\sigma_C^2)^{-1} = \frac{n\mu}{Ap(1-p)} - \frac{n\Sigma i^2 a_i(\theta)(1-p)^{i-2}}{A} - \frac{n}{A}\Sigma\frac{i^2 a_i(\theta)(1-p)^{2i-2}}{1-(1-p)^i} . \tag{4}$$

So far as the comparison between σ_A^2, σ_B^2, and σ_C^2 is concerned, we would intuitively expect that $\sigma_A^2 \leq \sigma_B^2 \leq \sigma_C^2$. We can show theoretically (see APPENDIX) that $\sigma_A^2 \leq \sigma_C^2$, with equality holding only if all family sizes are equal. However, we have not been able to prove in general that $\sigma_A^2 \leq \sigma_B^2$ or that $\sigma_B^2 \leq \sigma_C^2$, although we can do so in some specific cases.

Results

We have calculated the large-sample variances (2), (3), and (4) for a variety of FSDs and a variety of parameter values. In all cases, we found that the ratio σ_B^2/σ_C^2 is less than, but very close to, unity. This implies that if the general form of the FSD is known (e.g., Poisson), but with the numerical value of the parameter of the distribution unknown, there is essentially no gain in using the more complex estimation procedure of case B rather than the simpler procedure of case C, where no knowledge of the FSD is assumed.

The ratio σ_A^2/σ_C^2 is usually quite close to unity, but not as much so as σ_B^2/σ_C^2. We tabulate σ_A^2/σ_C^2 for the case where the FSD is Poisson in table 1, for the case where the FSD is geometric in table 2, and for the case where the family

TABLE 1

ASYMPTOTIC VARIANCE RATIOS σ_A^2/σ_C^2 FOR THE POISSON FSD

	SEGREGATION RATIO p						
POISSON PARAMETER θ	.01	.05	.10	.20	.25	.50	.90
.5	1.000	.977	.959	.932	.923	.910	.969
1.0	.995	.978	.961	.938	.932	.928	.978
2.0	.995	.979	.964	.949	.946	.955	.990
3.0	.995	.980	.968	.958	.958	.973	.996

size takes the two values 1 and 2, with respective probabilities $1 - \theta$ and θ, in table 3.

The main features of the tables and of our unreported calculations are the following: (1) In every case considered, $\sigma_A^2 < \sigma_B^2 < \sigma_C^2$, confirming our intuitive expectation. (2) As mentioned above, the difference between σ_B^2 and σ_C^2 is minute—in the Poisson case, for example, the difference between the two variances never exceeds 0.4%. (3) The differences between σ_A^2 and σ_C^2 are sometimes more noteworthy. This difference can be as high as 5%–10%; that is, the ratio σ_A^2/σ_C^2 goes down to 95% or even 90%. Such cases occur when the population FSD is skewed toward smaller families, as in the upper portions of the first three tables.

FINITE SAMPLE-SIZE VARIANCES AND BIAS

Large-sample formulas for the variances of maximum likelihood estimates, such as those considered in the preceding sections, are often assumed to give accurate expressions for samples of size 100 or more. Real samples will often be less than 100, and, further, the large-sample formulas might not be accurate even for samples exceeding 100. Also, estimates might be biased in small samples, and if so, a better measure of the accuracy of an estimator is the mean square error rather than the variance. For these reasons, we investigated small-sample properties of our estimators and calculated the biases and mean square errors of estimators, as well as the variances.

An immediate problem in this connection concerns the large amount of com-

TABLE 2

ASYMPTOTIC VARIANCE RATIOS σ_A^2/σ_C^2 FOR THE GEOMETRIC FSD
PROB(FAMILY SIZE = i) = $(1 - \theta)\theta^i$

	SEGREGATION PROBABILITY p						
GEOMETRIC PARAMETER θ	.01	.05	.10	.20	.25	.50	.90
.1	.994	.975	.954	.924	.913	.896	.963
.3	.993	.969	.947	.921	.914	.913	.974
.5	.991	.962	.939	.921	.919	.935	.985
.7	.986	.949	.932	.931	.945	.963	.994
.9	.962	.936	.949	.971	.980	.993	.999

TABLE 3

ASYMPTOTIC VARIANCE RATIOS σ_A^2/σ_C^2 FOR THE TWO-VALUED FSD

FSD		SEGREGATION RATIO p						
1 $(1-\theta,$	2 $\theta)$	.01	.05	.10	.20	.25	.50	.90
.99	.01	.995	.977	.958	.927	.916	.890	.957
.90	.10	.996	.981	.964	.937	.927*	.903*	.961
.75	.25	.997	.986	.973	.952	.944*	.923*	.968
.50	.50	.998	.992	.985	.972	.967*	.952*	.979
.25	.75	.999	.997	.993	.988	.985	.978	.990
.10	.90	1.000	.999	.998	.995	.994	.991	.996

* Also see table 4.

puting necessary for exact calculations. We used a two-pronged approach. We started with the simplest possible FSD, the two-valued distribution with only one- and two-child families. Exact calculations are feasible for this FSD. For more complicated FSDs, we turned to Monte Carlo simulation. We now present the results of these two approaches in turn.

Exact Calculations: The Two-Valued FSD

Under this FSD, a family has either two children, with population probability θ, or one child, with probability $1 - \theta$. Moreover, there are only three classes of families in the dataset, with respective numbers n_{11}, n_{21}, and n_{22}, where, using the notation established above, n_{ij} is the number of i-child families with exactly j children affected.

First, the estimators of p in cases B and C are identical, namely:

$$\hat{p} = 2n_{22}/(n_{21} + 2n_{22}) \ . \tag{5}$$

This estimator ignores, as it must, the one-child families, and this fact causes some difficulties in estimation theory, since in a small sample there is a nonnegligible probability that each family in the sample has only one child. No estimation of p would, or could, be undertaken for such a sample. Therefore, we decided to consider, in principle, only those samples with at least one two-child family. Probabilities for such samples can be found by a simple conditional probability argument; for samples of 20 or more, no significant amendment is necessary in this recalculation.

In case A (the FSD known, i.e., θ known), it follows from the likelihood (2) that the MLE of p is

$$\begin{aligned}\hat{p} = \{&\theta(n_{11} - n_{22}) - n_{21} - n_{22} \\ &+ [(\theta n_{22} - \theta n_{11} + n_{21} + n_{22})^2 + 4\theta(1 + \theta)n_{11}n_{22}]^{1/2}\}/2\theta n_{11}\end{aligned} \tag{6}$$

when n_{11} is nonzero and

$$\hat{p} = (1 + \theta)n_{22}/[(1 + \theta)n_{22} + n_{21}] \tag{7}$$

when n_{11} is zero. Note that when $\theta = 1$ (all families of size 2), the estimator (7) is identical to estimator (5), as we expect, since then there can be no difference between SM and GM estimation.

For any given sample size n, and for any given values of the parameters θ and p, we can calculate the probability of each possible value of the triplet (n_{11}, n_{21}, n_{22}), conditioned on $n_{11} + n_{12} > 0$. For each such triplet, we can estimate p using estimator (5) for SM estimation and estimator (6) or (7) for GM estimation. The exact mean value of the estimate of p for each mode of estimation can then be calculated by taking the weighted average (the weights being the probabilities for each triplet) of these estimates, and, similarly, the exact variance and the exact mean square error of each estimate can be found.

Table 4 shows the mean square error (MSE) for various sample sizes from $n = 20$ families up to $n = 500$. (We do not show the variance, since its behavior is very similar to that of the MSE for these values of n.) Table 5 gives the biases of the two estimators.

Our intuitive prior convictions would be that GM estimation of p is in all respects superior to SM estimation, since it uses the known form for the FSD. In some aspects of the tables, this intuition is confirmed: the bias in GM estimation is always less than for SM estimation for the numerical values we considered in table 5, and, as the sample size increases, the two variances

TABLE 4

EXACT MSE RATIOS (MSE_A/MSE_C) FOR THE TWO-VALUED FSD, AS A FUNCTION OF SAMPLE SIZE n

FSD			SEGREGATION RATIO p	
$(1-\theta,\ \theta)$		No.	.25	.50
.90	.10.....	20	.951	.795
		40	1.027	.848
		60	1.044	.881
		80	1.037	.885
		100	1.022	.883
		200	.970	.889
		500	.941	.897
		∞	.927*	.903*
.75	.25.....	20	1.054	.923
		40	1.028	.913
		60	.997	.907
		80	.980	.909
		100	.971	.911
		∞	.944*	.923*
.50	.50.....	20	1.026	.952
		40	.993	.944
		60	.983	.946
		80	.978	.947
		100	.976	.948
		∞	.967*	.952*

* The asymptotic values are taken from table 3.

TABLE 5

EXACT BIASES IN THE ESTIMATION OF p FOR CASES A AND C, FOR THE TWO-VALUED FSD

FSD			GENETIC PARAMETER p			
			.25		.50	
$(1-\theta,\ \theta)$		No.	Case (A)	Case (C)	Case (A)	Case (C)
.90	.10	20	−.017	−.052	−.044	−.080
		40	−.010	−.027	−.014	−.040
		60	−.008	−.018	−.007	−.025
.75	.25	20	−.015	−.023	−.017	−.032
		40	−.006	−.011	−.006	−.015
		60	−.004	−.007	−.003	−.010
.50	.50	20	−.010	−.013	−.010	−.016
		40	−.004	−.006	−.005	−.008
		60	−.003	−.004	−.003	−.005

NOTE: Bias is $E[\hat{p}] - p$.

approach their asymptotic values, for which, as noted above, the GM value is always less than the SM value. However, the tables present several curious features. We would expect, both from common sense and from the large-sample comparison, that the MSE for estimation of p would be smaller under GM estimation than for SM, no matter what the sample size. This, however, is not always the case. For $\theta = .5$ and $p = .25$, GM estimation has a lower MSE only for sample sizes of 35 or more, while when $\theta = .25$ and $p = .25$, the sample size must be 60 or more before GM estimation has the lower MSE. For $\theta = .1$ and $p = .25$, we require a sample of 120 or more. On the other hand, when $p = .5$, GM estimation has a smaller MSE for all numerical values we considered.

The second curious feature is that the ratios of the GM to SM MSEs do not necessarily smoothly increase (or smoothly decrease) as the sample size is increased: the ratio can, for example, increase and then decrease (and, then, as for the case $\theta = .5$, $p = .5$, increase again). On the other hand, in the majority of cases, the ratio is smoothly increasing (or decreasing, as the case may be) as soon as the sample size exceeds about 30.

Both these curious features are true for the variance as well as the MSE.

Simulations

As stated earlier, it is extremely difficult to calculate exact variances of estimates of p in finite samples when the FSD allows families of size larger than two. This is because of the very large number of possible ways that the families can be distributed. We therefore conducted some simulations for the Poisson FSD, to assess whether the conclusions just reached for one- and two-child families continue to hold. The results of this simulation are reported in table 6. We note the following conclusions: first, SM and GM estimates have variances that are quite close: second, for small samples, it is possible that the SM

TABLE 6

SIMULATED VARIANCE RATIOS σ_A^2/σ_C^2 AND BIASES FOR CASES A AND C, FOR THE POISSON FSD (SEGREGATION RATIO $p = .25$)

		POISSON PARAMETER θ					
		0.5			3.0		
M*	No.†	Variance ratio	Bias (A)	Bias (C)	Variance ratio	Bias (A)	Bias (C)
750	20	1.03	−.016	−.025	.97	−.006	−.007
300	50	.99	−.009	−.012	.96	−.005	−.005
150	100	.91	−.009	−.009	.95	−.005	−.004
	∞	.92‡			.96‡		

* M = no. datasets simulated.
† No. = no. families per dataset.
‡ From table 1.

estimate has a smaller variance than the GM estimate (although this cannot be completely verified because of the randomness in the simulation procedure); third, that the SM and the GM variances have both effectively reached their large-sample values when the sample size is 100 or more. Finally, the biases of the two estimators are close, and both approach zero as n increases. Thus, for the more "spread-out" Poisson family-size distribution, the asymptotic theory seems rather more accurate than for the less "spread-out" one- and two-child family case.

DISCUSSION

Here, we have examined the effect that knowledge of the population family-size distribution (FSD) can have on the estimation of genetic parameters. We have considered three cases: where we know the FSD completely (case A); where we know the statistical form of the FSD but not its parameter θ (case B); and where we know nothing about the FSD (case C). We have examined a variety of FSDs, and of these, we have reported here results for the Poisson, the geometric, and a simple two-valued FSD in which all families have either one or two children. We have considered both the asymptotic (large-sample) variances and the variances and biases for finite-sample sizes.

Briefly, we found the following for the asymptotic variances. Comparing cases A and C: the asymptotic variance for case A, σ_A^2, cannot exceed σ_C^2. The ratio σ_A^2/σ_C^2 is close to unity in many cases. However, for selected examples of the FSDs we examined, this ratio can be lower than .95, or even .90. These lower ratios occur when the segregation parameter p is around .25 to .50 (i.e., within a reasonable range for genetic models) and when the FSD is skewed toward smaller families (e.g., $\theta = 0.5$ in the Poisson; and several cases in the two-valued FSD). Comparing cases B and C, however, there is little decrease in the variance, and the ratio σ_B^2/σ_C^2 is close to unity for all values of the parameters considered.

For finite-sample sizes, the results are not so straightforward. Whether we

look at the variance or the MSE, we find that the ratio σ_A^2/σ_C^2 occasionally *exceeds* unity; moreover, this ratio does not always approach its asymptotic value smoothly. In particular, we can make these statements with certainty for the simple two-valued FSD (table 4), the only FSD for which we did exact calculations. Simulations for the Poisson FSD (table 6) also yielded a ratio greater than unity for at least one sample size, although with simulations we cannot rule out the possibility that this observation is due simply to random variation.

The fact that the ratio σ_A^2/σ_C^2 can exceed unity in finite samples seems counterintuitive to us. *If* the FSD is known, then the GM estimator (6) or (7) represents the MLE, whereas the SM estimator (5) can be viewed as "arbitrarily" throwing out some information. Intuitively, we expected the MLE to have lower variance and MSE than any other estimator. However, in fact, such a relationship is not guaranteed theoretically, since in this case the MLE of p is not a sufficient statistic for p (see [4]).

The simple two-valued FSD interested us for two reasons. First, due to its simplicity, it is amenable to exact calculations. (The number of different ways that n families can be distributed among k categories is $(n + k - 1)!/[n! \cdot (k - 1)!]$; see [6]. Thus, for example, if the FSD includes families up to size 4, then there are 10 different types of families in the dataset. Now even a sample of only 10 families includes approximately 92,000 different configurations, and even for a [small] sample of 30 families, there are approximately 2×10^8 different family configurations.) The second appeal of the two-valued FSD was our intuitive expectation that, to the extent that case A *was* superior to case C, this superiority should be greatest in populations with a preponderance of small families, and, indeed, our findings did support this intuition.

We turn now to the question of why we used exact calculations for finite sample sizes, as opposed to relying exclusively on simulations. The quantity we were interested in—namely, the ratio of two variances—is difficult to simulate precisely without extraordinarily large numbers of families. To illustrate, consider the entry for sample size $n = 50$ in table 6, with the Poisson parameter $\theta = 0.5$. Three hundred of these datasets were simulated. We can use F tables to form an approximate 95% two-sided confidence interval about the simulated ratio of .99 shown in the table. With (300,300) degrees of freedom, the confidence interval is approximately (0.79, 1.24). This interval is nowhere near narrow enough to settle the question of interest, that is, whether the true variance ratio is greater or less than unity. Yet 15,000 families were used. To obtain a sufficiently narrow confidence interval would require tens of thousands, perhaps hundreds of thousands, of simulated families. At least for larger n ($n = 100$ in table 6, for example), we can see that the simulated variance ratio is fairly close to the asymptotic value. But for smaller sample sizes ($n = 20$ and $n = 50$), we cannot be sure the simulated variances are near their asymptotic values.

Although we were primarily interested in the behavior of MLEs of a genetic parameter, a secondary finding is the need for caution in the use of asymptotic

large-sample statistical properties for populations where the family size is limited to one or two children. Our exact calculations reveal that even samples of 100 such families can have variances that are far from their asymptotic limits (table 4). Moreover, although the ratio σ_A^2/σ_C^2 is always less than unity asymptotically, it can be greater than unity in finite samples, even those as large as 60, 80, or 100 families. It is true that these are 60 or 100 very small families, and our simulations suggest that this problem would be less severe in studies involving larger families. Nevertheless, we should not apply asymptotic results too casually to small samples, whatever the family-size distribution.

CONCLUSION

The main conclusion we reach is that when estimating genetic parameters from data from families of various sizes, the most reasonable estimation procedure in most situations is that involving SM estimation, that is, where any information one has concerning the family-size distribution (FSD) is ignored. The reason for this is as follows: If the FSD, with all its parameter values, is known and used in the estimation procedure, one can expect at best only a small gain in precision for estimation of genetic parameters (and, in some cases, for samples of size 30–50, there may even be a loss in precision), compared to the SM estimation procedure where information on the FSD is ignored. When the form of the FSD is known (e.g., Poisson), but its parameter values are unknown (and must be estimated along with the genetic parameters), at best a minute increase in precision over the SM approach will occur. Since, also, there are possible biases involved if an incorrect form for the FSD is assumed, and since the calculations are rather simpler in the SM case, we believe that, on the whole, the best practical estimation procedure is SM estimation, that is, estimation where one simply ignores any information about, or one makes no assumption about, the form of the family-size distribution.

However, we might consider amending these conclusions for a population meeting two very specific requirements. First, census data must be available giving an accurate description of the FSD—not as following a particular statistical distribution, such as the Poisson, but simply as a multinomial probability for each family size. Second, the population must consist almost exclusively of very small families. As we have shown here, when the population consists of 75% one-child and 25% two-child families, or even 90% one-child and 10% two-child, then using the GM formulation (case A) can decrease the variance of the estimator by approximately 10%—with reasonable sample sizes that could realistically be available (table 4). Although these two requirements will probably not be met very often, they might be fulfilled currently in the People's Republic of China.

ACKNOWLEDGMENT

Our thanks to Dr. David A. Greenberg for providing the computer program used to generate all our simulated families.

APPENDIX

PROOF THAT $\sigma_A^2 \leq \sigma_C^2$

We will use the expressions (2) and (4) to prove that $\sigma_A^2 \leq \sigma_C^2$, equality holding if and only if all families are of the same size.

The first two terms on the two respective right-hand sides of expressions (2) and (4) are equal, so that the desired inequality will hold if we can show that $[\Sigma\, ia_i(\theta) \cdot (1 - p)^{i-1}]^2/A^2 \leq [\Sigma\, i^2\, a_i(\theta)(1 - p)^{2i-2}]/A$, or

$$[\Sigma\, ia_i(\theta)(1 - p)^{i-1}]^2 \leq \left[\Sigma\, \frac{i^2 a_i(\theta)(1 - p)^{2i-2}}{1 - (1 - p)^i}\right]\{\Sigma\, a_i(\theta)[1 - (1 - p)^i]\}\ ,$$

whatever the form of $\{a_1(\theta), a_2(\theta), \ldots\}$. The coefficient of $a_i^2(\theta)$ on both left- and right-hand sides is $i^2(1 - p)^{2i-2}$. The coefficient of $a_i(\theta)a_j(\theta)$ on the left-hand side is

$$2ij(1 - p)^{i+j-2}\ , \tag{A1}$$

and on the right-hand side is

$$\frac{i^2(1 - p)^{2i-2}[1 - (1 - p)^j]}{1 - (1 - p)^i} + \frac{j^2(1 - p)^{2j-2}[1 - (1 - p)^i]}{1 - (1 - p)^j}\ . \tag{A2}$$

Thus, the desired inequality will follow if expression (A1) < (A2) for $i \neq j$. Dividing both expressions by the positive quantity $ij(1 - p)^{i+j-2}$, this will be the case if

$$2 < \frac{i(1 - p)^i[1 - (1 - p)^j]}{j(1 - p)^j[1 - (1 - p)^i]} + \frac{j(1 - p)^j[1 - (1 - p)^i]}{i(1 - p)^i[1 - (1 - p)^j]}\ .$$

But this inequality is always true for $i \neq j$, since the right-hand side is of the form $x + x^{-1}$, which exceeds 2 unless $x = 1$, which implies $i = j$.

REFERENCES

1. Hodge SE: Family-size distribution and Ewens' equivalence theorem. *Am J Hum Genet* 37:166–177, 1985
2. Barrai I, Mi MP, Morton NE, Yasuda N: Estimation of prevalence under incomplete selection. *Am J Hum Genet* 17:221–235, 1965
3. Ewens WJ, Asaba B: Estimating parameters of the family-size distribution in ascertainment sampling schemes: numerical results. *Biometrics* 40:367–374, 1984
4. Kendall MG, Stuart A: *The Advanced Theory of Statistics*, vol 2. London, Griffin, 1961
5. Elandt-Johnson RC: *Probability Models and Statistical Methods in Genetics*. New York, John Wiley, 1971
6. Feller W: *An Introduction to Probability Theory and Its Applications*, vol 1. New York, John Wiley, 1957