Am. J. Hum. Genet. 42:506–515, 1988

Multifactorial Analysis of Family Data Ascertained through Truncation: A Comparative Evaluation of Two Methods of Statistical Inference

D. C. Rao,*† R. Wette,* and W. J. Ewens‡

*Division of Biostatistics and †Departments of Psychiatry and Genetics, Washington University School of Medicine, St. Louis; and ‡Department of Mathematics, Monash University, Clayton, Australia, and Department of Biology, University of Pennsylvania, Philadelphia

Summary

When family data are ascertained through single selection based on truncation, a prevailing method of analysis is to condition the likelihood function on the proband's actual phenotypic value. An alternative method conditions the likelihood function on the event that the proband's measurement lies in the truncation region. Both methods are contrasted here by using Monte Carlo simulations; identical sets of data were analyzed using both methods. The results suggest that, under either method, (1) parameter estimates are nearly unbiased and (2) likelihood-ratio tests of null hypotheses are approximately distributed as χ^2. However, conditioning on the proband's actual phenotypic value yields considerably less efficient estimates and reduced power for hypothesis tests. A corresponding result also holds under complete ascertainment. It is argued, therefore, that whenever sufficient information is available on the nature of truncation, the alternative approach should be used.

Introduction

Genetic epidemiology often deals with investigations of familial aggregation of diseases and disease-related traits based on samples of related individuals. Since familial environment is known to exert influence on many diseases and disease-related traits (e.g., Dadone et al. 1984; Lipid Research Clinics Program 1984*a*, 1984*b*; Janus et al. 1985), resolution of familial aggregation into genetic and familial environmental (cultural) inheritance constitutes an important step. Although many such investigations employ random ascertainment of families, nonrandom ascertainment is coming into increased use. The LRC family studies (Lipid Research Clinics Program Family Study Committee 1984) and the Honolulu Heart Study (Gulbrandsen et al. 1977) are typical examples of nonrandom ascertainment. Such studies require appropriate methods of analysis. A systematic maximum-likelihood method for multifactorial analysis of family data ascertained specifically through truncation has recently been presented elsewhere (Rao and Wette 1987). On the other hand, a prevailing method of analysis (Boehnke and Lange 1984), applied especially in variance-components analysis, uses a likelihood function conditional on the proband's actual phenotypic value. The primary purpose of the present article is to present a comparative evaluation of both approaches under single ascertainment by means of analytical and numerical methods.

In what follows, we shall first briefly review the likelihood theory for a random sample of families, a theory that plays an important role in both methods. This will be followed by a discussion of the two methods of analysis, an analytical evaluation, and, finally, the results of comparative simulation studies.

Received June 23, 1987; revision received September 29, 1987.

Address for correspondence and reprints: Dr. D. C. Rao, Division of Biostatistics, Washington University School of Medicine, Box 8067, 660 South Euclid, St. Louis, MO 63110.

0002-9297/88/4203-0014$02.00

The Likelihood Method for a Random Sample

Consider a random sample of n nuclear families with variable sibship sizes. For one nuclear family with s children, let $\underset{\sim}{X}' = (X_{11}, X_{12}, X_{21}, X_{22}, \ldots, X_{2s})$ denote the row vector of phenotypes of father (X_{11}), mother (X_{12}), and children (X_{2k}, $k = 1, \ldots, s$), which are assumed to be adjusted for the effects of concomitant variables such as age and sex. For simplicity, assume a common mean (μ) and a common variance (σ^2) for the phenotypes of fathers, mothers, and children. Define three familial correlations as follows: marital, $\rho_1 = \rho(X_{11}, X_{12})$; parent-child, $\rho_{12} = \rho(X_{1k}, X_{2l})$ for $k = 1, 2$, and $l = 1, \ldots, s$; and sibling, $\rho_2 = \rho(X_{2k}, X_{2l})$ for $k \neq l = 1, \ldots, s$. Denote the covariance matrix of the $s + 2$ variables ($\underset{\sim}{X}$) by Σ. Assuming that the phenotypes $\underset{\sim}{X}$ of the family members jointly follow an $(s + 2)$-variate normal distribution with joint-density function $f(\underset{\sim}{X})$, the log-likelihood function for the family may be written as

$$\ln L = \ln f(\underset{\sim}{X}) = -\tfrac{1}{2}\,[\ln|\Sigma| + (\underset{\sim}{X} - \underset{\sim}{\mu})'\,\Sigma^{-1}\,(\underset{\sim}{X} - \underset{\sim}{\mu})] + \text{constant}, \quad (1)$$

where $\underset{\sim}{\mu}' = (\mu, \ldots, \mu)$ is the mean vector of $\underset{\sim}{X}$, $|\Sigma|$ is the determinant of and Σ^{-1} is the inverse of the covariance matrix Σ. If $\ln L_i$ is used to denote the log likelihood for the ith of n families, the total log-likelihood function for the entire random sample of n families is

$$\ln L = \sum_{i=1}^{n} \ln L_i \,. \quad (2)$$

The five unknown parameters (μ, σ^2, ρ_1, ρ_{12}, and ρ_2) may be estimated simultaneously by maximizing $\ln L$, and tests of hypotheses on the parameters may be performed using likelihood-ratio tests (LRT). Alternatively, the familial correlations (ρ_1, ρ_{12}, and ρ_2) may be expressed as functions of the parameters of a transmission model, such as the simple one defined by the following three parameters: p = correlation between parental phenotypes; h^2 = genetic heritability (defined as the proportion of phenotypic variance due to genetic effects); and c^2 = cultural heritability (defined here as the proportion of phenotypic variance due to a common sibship environment). This model, obtained as a special case of a more general model (Rao et al. 1984), defines unique expectations for the following three correlations: $\rho_1 = p$, $\rho_{12} = h^2\,(1 + p)/2$, and $\rho_2 = h^2\,(1 + ph^2)/2 + c^2$. Estimating mean, variance, and all three model parameters under the full model results in a value of $\ln L$ in equation (2), say $\ln L_5$. Under the null hypothesis of, say, no marital correlation ($p = 0$), setting $p = 0$ and estimating the remaining four parameters yields another $\ln L$ value, say $\ln L_4$. Then,

$$\chi^2_1 = 2\,(\ln L_5 - \ln L_4) \quad (3)$$

provides the LRT statistic for the null hypothesis $p = 0$. This test statistic follows asymptotically a χ^2 distribution with 1 df (McGue et al. 1987). Similar tests can be developed for other hypotheses of interest. Properties of such tests, evaluated using simulated random samples of family data, have been presented elsewhere (McGue et al. 1987), which corroborate the distributional assumption of the LRT.

Single Ascertainment through Direct Truncation

In the context of multifactorial studies, we define *direct truncation* as the case in which probands are ascertained from a certain region of the phenotypic distribution, such as from the upper decile (Rao and Wette 1987). Typically, this selection type is realized when, in a random sample, individuals are first measured for a certain phenotype (e.g., the high-density-lipoprotein cholesterol [HDL-C]), and only those individuals whose phenotypic values are, say, in the upper decile of the age-sex-specific distribution are selected as "probands" for family studies. Examples include the kindreds ascertained for hyperalphalipoproteinemia, defined by the upper-decile HDL-C (Glueck et al. 1975), and for hypoalphalipoproteinemia, defined by the bottom-decile HDL-C (Third et al. 1984).

Two other cases of ascertainment through truncation, in addition to that of direct truncation, were also considered by Rao and Wette (1987). In one case, probands are ascertained from a certain region of the distribution of a correlated quantitative trait, referred to as *indirect truncation*. In the other case, probands are ascertained through an associated disease, referred to as *latent truncation*. The comparative evaluation presented in the present paper pertains only to the case of direct truncation.

Consider a sample of n families ascertained from the upper tail through direct truncation. Denote by $\underset{\sim}{X}$

the vector of phenotypes of all $s + 2$ members of a nuclear family with s children, with the parameterization as defined in the previous section. Recall that, under single ascertainment, each family has only one proband. Let X_p denote the proband's phenotype (which is a component of $\underset{\sim}{X}$) and let $\underset{\sim}{X}_p$ be the reduced vector of phenotypes of all nonproband members. Assume that $\underset{\sim}{X}$ has been sampled from a multivariate normal distribution *only because* $X_p \geq T$ for each family, where T represents the point of truncation.

Generic Method

For statistical inference under this type of selection, Boehnke and Lange (1984) proposed, in what may be termed a "generic" approach, to condition the likelihood function of a particular family with s children on the actual phenotypic value of the proband, i.e.,

$$L_G(\underset{\sim}{X}_p|X_p) = \frac{f_{s+2}(\underset{\sim}{X})}{f_1(X_p)} , \qquad (4)$$

where f_υ is the υ-variate normal density function (see also Cannings and Thompson 1977; Hopper and Mathews 1982). This method is termed generic only in the context of single ascertainment. For other types of ascertainment, this method ceases to be generic, since some desirable properties no longer hold (W. J. Ewens and R. Green, unpublished data).

Note that equation (4) is equivalent to

$$L_G(\underset{\sim}{X}_p|X_p) = f^*_{s+1}(\underset{\sim}{X}_p) , \qquad (5)$$

where the density function is now an $(s + 1)$-variate normal with the $(s + 1)$-dimensional mean vector $\underset{\sim}{\mu}^* = \underset{\sim}{\mu} + (x_p - \mu)\underset{\sim}{\beta}$, where x_p is the proband's actual value and $\underset{\sim}{\beta}$ is the vector of regression coefficients of the nonproband values on the proband value, and with the $(s + 1) \times (s + 1)$–dimensional covariance matrix $\Sigma^* = \Sigma_{s+1} - \underset{\sim}{\beta}\underset{\sim}{\beta}' \sigma^2$, where Σ_{s+1} is the covariance submatrix of $\underset{\sim}{\Sigma}$ for the nonproband variables. In the above, note that x_p enters the likelihood function as a parameter (of known value), not as a variable. Obviously, the form of equation (5) is valid for any arbitrary distribution of the proband values, truncate or otherwise, since it holds equivalently for any $(s + 2)$-dimensional density function $f_{s+2}(\underset{\sim}{X})$ subject to the condition that the density in equation (5) is $(s + 1)$-variate normal. However, $\underset{\sim}{\mu}^*$ and Σ^* as given above are necessary and sufficient for $f_{s+2}(\underset{\sim}{X})$ to be $(s + 2)$-variate normal if $f^*_{s+1}(\underset{\sim}{X}_p)$ is $(s + 1)$-variate normal.

Specific Method

In contrast to the above method, Rao and Wette (1987) proposed, in what may be termed a "specific" approach, to condition the likelihood function on the event $(X_p \geq T)$ that a proband value is contained in the specific region from which the probands are selected under direct truncate selection; i.e.,

$$L_S(\underset{\sim}{X}_p|X_p \geq T) = \frac{f_{s+2}(\underset{\sim}{X})}{Q(X_p \geq T)} , \qquad (6)$$

where $Q(X_p \geq T) = Q[(T - \mu)/\sigma]$ is the upper-tail probability of the standard normal distribution function corresponding to the normal deviate $Z = (T - \mu)/\sigma$. Note that when the actual value of Q is known, the denominator in equation (6) enters the likelihood function only as a constant and is of no further interest. However, the value of Z is also then known (as the value Z of the standard normal deviate corresponding to Q), which fact imposes the linear constraint

$$\mu = T - \sigma Z \qquad (7)$$

on the relationship among μ, σ^2, T, and Z. Thus, μ is not estimated as an independent parameter but only according to equation (7).

For either method, the total log-likelihood function for a sample of n families is obtained by summing the log-likelihood functions over the n individual families. In comparing the two approaches, may it first be noted that the specific approach is explicitly geared to the particular type of selection in that it stipulates that probands are sampled from the upper tail $X_p \geq T$ of a univariate normal distribution and is therefore not applicable to other sampling schemes. On the other hand, the generic approach does not involve such a restriction; and it follows that statistical inference based on it is validly applicable not only to direct truncation but to a wider class of situations (see, e.g., Simpson et al. 1981; Hopper and Mathews 1982; Beaty and Liang 1987), so long as we confine it to single ascertainment. By the same token, one may expect that, when used in a particular situation, the (or any) generic approach will be less efficient for estimation and less powerful for hypothesis testing than the (or any) specific approach tailored to that particular situation. Whether the loss in efficiency and power is of practical significance is, of course, a different issue, which will be addressed in a later section.

A Theoretical Evaluation

The likelihood functions L_G and L_S, defined respectively by equations (4) and (6) for the generic and specific approaches, clearly satisfy the relation

$$L_S = L_G \times [f_1(X_p)/Q(X_p \geq T)] \ . \qquad (8)$$

This relation provides the information needed to compare the variances of the estimators under the two approaches, the key observation being that the term in brackets on the right-hand side of equation (8) is a proper density function (viz., the conditional density function of the proband's value, given that this value exceeds T). Thus, equation (8) can be rewritten as

$$L_S = L_G \times L_p \ , \qquad (9)$$

where L_S, L_G, and L_p are all proper likelihood functions; and this observation continues to be true when the sample contains many families and we form likelihood functions from the products of likelihood functions over individual families. We may now reinterpret the terms in equation (9) as these products.

Since the "information" matrix for any maximum-likelihood procedure can be found by taking expectations of appropriate second derivatives (with respect to the various parameters) of the total log-likelihood function, equation (9) yields the following result on the information matrices from the three terms in equation (9):

$$I_S = I_G + I_p \ . \qquad (10)$$

Because each L in equation (9) is a proper likelihood function, all three I matrices in equation (10) are positive definite, and this implies that more information is extracted from the data under the specific method than under the generic method. The inverse of any I matrix provides the asymptotic variances and covariances of parameter estimates, and the relationship in equation (10) proves (Rao 1973) that the variance of the specific estimator of any parameter is necessarily less than that of the generic estimator. To this extent we would prefer the specific estimation procedure, although, as will be noted later, this result applies only when the selection region ($=Q$) is well defined.

Both the specific and the generic estimation approaches are maximum-likelihood procedures, so we expect standard maximum-likelihood theory to apply for both. In particular, both procedures should lead to asymptotically (as $n \to \infty$) unbiased estimators. There is no theory to indicate whether, in small samples, the bias arising under one approach is smaller than that arising under the other, so we would expect in practice that there would be an unsystematic pattern of bias comparison between the two approaches.

Although the theory outlined above shows that the variance—and hence the SE—of the estimator of any parameter under the specific approach is less than that under the generic approach, it does not indicate how large the difference between the SEs might be. To get some information on this point we resort to Monte Carlo simulations. The results of these simulations are described in the next section.

Finally, the foregoing result—i.e., that the specific approach will always produce parameter estimates with smaller SEs than those produced under the generic method—relies on the fact that the bracketed term on the right-hand side in equation (8) is a proper density function. A corresponding result holds under complete ascertainment as well (W. J. Ewens and R. Green, unpublished data).

Comparative Evaluation of the Methods by Means of Simulation

The performance of each of the two methods discussed above for parameter estimation and hypothesis testing was evaluated by means of the Monte Carlo method, which was implemented on a Harris 100 computer system (for details, see Rao and Wette 1987). Each family consisted of two parents and a variable number (s = 2, 3, or 4) of offspring (a minimum of two offspring ensures that all families contribute information on the sibling correlation). The sibship-size distribution, generated according to a geometric distribution, was held constant over multiple replications of a simulation condition so as not to disturb the sampling properties of estimators and test statistics by extraneous variation (see Rao and Wette 1987). The proband's position in the family was determined at random, with each family member being equally likely to become proband. The family data for such a family were generated so that the proband's selection variable (X_p) was in the upper $100Q\%$ tail, with Q = .10, .05, or .01, and for $\mu = 0, \sigma^2 = 1, p = 0, h^2 = 0.6$, and $c^2 = 0.2$ (values for p, h^2, and c^2 were chosen so as to correspond to highly heritable phenotypes such as some lipid variables).

Samples of n families, each generated this way,

were analyzed by maximizing the total log likelihood under each of the two methods. Specifically, ln L was maximized to estimate all relevant parameters under the alternative (full-model) hypothesis, as well as under the null hypothesis $p = 0$, in which only the remaining parameters were estimated. The LRT statistic was then evaluated for testing the null hypothesis. The entire process of simulation of a sample of n families and analysis of the data was replicated 1,000 times for each value of Q. The sampling properties of the parameter estimates and of the LRT statistic were then evaluated over the N replications.

In every experimental condition, more than the targeted number of replicate samples had to be generated since some of them were regarded as unsuitable for an automated analysis. A particular replication was unacceptable—and therefore the generated sample was rejected and replaced with a new sample—if any of the following situations pertained to either of the two methods of analysis: (1) a boundary solution (i.e., ± 1 for p, and 0 or 1 for h and c) resulted despite two attempts at numerical optimization; (2) optimization did not converge owing to numerical difficulties; (3) the variance-covariance matrix of parameter estimates was not positive definite; or (4) the LRT statistic was not positive. The latter three situations were responsible for 97% of all sample rejections in our simulations. In all, 8,673 replications were generated to yield 8,500 acceptable replications (4,000 in table 1, 3,000 in table 3, 1,000 in table 4, and 500 in table 5), with an overall sample rejection rate of 2.0%. For any of the nine simulation conditions presented in tables 1 and 3–5, this rate ranged between 0.8% and 4.1%. For an automated procedure involving large numbers of replications, these sample rejection rates are regarded as negligible. In any case, they could not have affected our results to any considerable degree.

Parameter Estimates and Hypothesis Tests

Table 1 presents the main results obtained using the two methods both in small samples ($n = 50$) at three levels of truncation ($Q = .01, .05$, and .10) and in large samples ($n = 200$) at an extreme level of truncation ($Q = .01$). In each case, the three model parameters h, c, and p (in addition to other relevant parameters) were estimated by means of maximum-likelihood iteration, and the three familial correlations (ρ_1, ρ_{12}, and ρ_2) were computed as functions of the estimates of h, c, and p. Although parameter estimates under the null hypothesis of $p = 0$ might be more appropriate to present (unless the null hypothesis is rejected), we chose to present the estimates under the alternative hypothesis so as to present estimates of all parameters. In any case, parameter estimates and SEs were very comparable under the two hypotheses (except for p). Results presented include (1) the average parameter estimates; (2) two types of SD, where SD1 is the observed SD of estimates among replications and SD2 is the root-mean-squared average of the asymptotic SEs obtained from the likelihood method; and (3) the average LRT statistic (reported as a χ^2 with 1 df) for testing the null hypothesis of $p = 0$. It may be noted that the SE of an *average* parameter estimate is given by SD/$\sqrt{N}$ and not by SD itself. We prefer to present the SD as the appropriate measure of precision for family studies because it is (an estimate of) the "SE" of a parameter estimate as obtained from just one such study. It may be noted that SD1, which is not available in single studies, never exceeds SD2 appreciably, which fact indicates that the SE obtained under the likelihood methods is a valid estimate of the sampling variation of the estimates even in small samples.

Although bias, as estimated by the difference between average estimate and true value of a parameter, is systematic and often significant under both methods, especially in small samples, it should pose no real concern, for the following two reasons: First, it is small compared with the SE of the parameter estimate from single studies (e.g., SD1 in table 1). Second, the significance of bias is artificial in the sense that it depends on the number N chosen. Beyond a certain N (such as 100), the effect of increasing the N has been seen to reduce primarily the SE but not the bias, thus rendering the otherwise unimportant bias "significant" (Rao and Wette 1987). We employed a large N primarily to investigate the distribution of the LRT statistic as validly as possible. We conclude that the parameter estimates are, under either method, sufficiently accurate even in small samples. The large bias observed by Boehnke and Lange (1984) for the generic method perhaps may be attributable to random variation, since they used small N values.

Although bias is negligible under both methods, the SEs are considerably larger under the generic method than under the specific method. They are often almost twice as large—and, in fact, four times as large for the mean. A clear exception is that of the estimate of c^2 (which parameter enters only the ρ_2

Table 1

Parameter Estimates, SDs, and Tests of Hypotheses, Using the Specific (eq. [6]) and Generic (eq. [5]) Likelihood Methods

Parameter (true value) and Estimate	Q: .01 (upper 1%), n = 50 Families[a]		Q: .01 (upper 1%), n = 200 Families[b]		Q: .05 (upper 5%), n = 50 Families[c]		Q: .10 (upper 10%), n = 50 Families[d]	
	Specific	Generic	Specific	Generic	Specific	Generic	Specific	Generic
p (.0):								
Average	−.019	−.011	−.004	.001	−.017	−.009	−.019	−.018
SD1	.091	.180	.043	.085	.097	.172	.108	.172
SD2	.086	.168	.043	.085	.096	.167	.103	.165
h^2 (.6):								
Average	.578	.559	.595	.594	.577	.561	.578	.557
SD1	.092	.170	.044	.073	.103	.170	.114	.176
SD2	.091	.174	.043	.072	.102	.171	.110	.182
c^2 (.2):								
Average	.196	.195	.200	.199	.192	.190	.192	.192
SD1	.049	.054	.023	.025	.061	.065	.066	.070
SD2	.050	.055	.024	.026	.061	.067	.071	.072
ρ_1 (.0):								
Average	−.019	−.011	−.004	.001	−.017	−.009	−.019	−.018
SD1	.091	.180	.043	.085	.097	.172	.108	.172
ρ_{12} (.3):								
Average	.286	.293	.297	.301	.286	.293	.286	.287
SD1	.056	.119	.027	.058	.059	.115	.062	.109
ρ_2 (.5):								
Average	.488	.494	.498	.501	.485	.491	.486	.488
SD1	.059	.091	.029	.045	.063	.092	.069	.089
μ (0):								
Average	.038	−.062	.010	−.020	.032	−.042	.024	−.016
SD1	.143	.565	.071	.236	.097	.427	.076	.330
SD2	NA	.647	NA	.236	NA	.443	NA	.349
σ^2 (1):								
Average	.977	1.005	.994	1.001	.977	.999	.984	.997
SD1	.122	.170	.060	.076	.116	.161	.116	.148
SD2	.121	.205	.061	.078	.119	.178	.119	.167
LRT of H_0: p = 0:								
Average χ^2_1	1.15	1.14	1.02	.96	1.04	1.07	1.11	1.12
Distribution of N = 1,000 χ^2_2	10.32	9.68	.37	4.55	2.65	3.22	6.72	9.52
LRT values: χ^2 vs. gamma P	.006	.008	.83	.10	.27	.20	.03	.01

NOTE.—The number of replications is 1,000 for each simulation condition. SD1 = the square root of the empirical variance among the N estimates of a parameter. SD2 = the root-mean-squared average of the asymptotic SEs obtained on the basis of the likelihood method. NA = not available (since μ was calculated using eq. [7]).

[a] In each replication, there were 28, 12, and 10 families with 2, 3, and 4 children, respectively.

[b] In each replication, there were 110, 63, and 27 families with 2, 3, and 4 children, respectively.

[c] In each replication, there were 25, 15, and 10 families with 2, 3, and 4 children, respectively.

[d] In each replication, there were 26, 16, and 8 families with 2, 3, and 4 children, respectively.

correlation). A plausible explanation is that, even when a child is the proband, sibs of the proband (for $s > 2$) still provide information on c^2; and if a parent is the proband, the full sibship contributes information on c^2. Perhaps this is more readily evident from the SEs of the correlation estimates; that is, whereas the SE for ρ_1 or ρ_{12} is nearly doubled, that of ρ_2 is only 50% larger.

When each method is considered on its own merits, the relative efficiency of the parameter estimates is high, as indicated by the closeness of SD1 and SD2; the latter may underestimate the actual SE slightly in small samples. In comparison, however, the specific approach evidently provides, when the sampling is done under direct truncation, considerably higher efficiency than does the generic approach. This fact substantiates the argument forwarded earlier—namely, that a method specifically geared to direct truncation should be more efficient.

We note that all of these conclusions agree with the theoretical predictions made in the previous section. The SEs of all generic estimators are always larger than those of the corresponding specific estimators, the simulations showing that the excess is usually by a factor of approximately two. Further, as predicted theoretically, there is no systematic pattern to the bias, with sometimes one estimator and sometimes the other having the smaller bias. Two further general theoretical predictions are also confirmed by the results in table 1. First, under both estimation procedures the bias decreases as n increases. Second, the SEs of parameter estimates (again under both estimation procedures) obtained from $n = 200$ should be half of those from $n = 50$, and this is again confirmed by the simulations.

The LRT statistic for testing the null hypothesis $p = 0$ was computed in each of the replications. Its sampling distribution was investigated to see whether it attains the χ^2 distribution with 1 df, as predicted on the basis of asymptotic theory. This was done by testing for χ^2 against a two-parameter gamma distribution with an LRT yielding a χ^2 with 2 df (e.g., see Rao and Wette 1987). The results are given at the bottom of table 1. Under the specific method, the null distribution of the LRT statistic approximates a χ^2 distribution reasonably well even in samples as small as 50 families—except, perhaps, for extreme truncation (i.e., $Q = .01$). A similar conclusion holds under the generic method also, with, perhaps, one exception. Under extreme truncation, the parameter estimates are accurate under either method and almost fully efficient under the specific method even in small samples, whereas it may appear that the sampling distribution of the LRT statistic under either method attains a χ^2 distribution only in larger samples. It should be noted, however, that the LRT is likely to be sensitive against small deviations from the χ^2 distribution because it is based on a large N.

Empirical Rejection Rates and Power

The empirical rejection rates at several nominal significance levels (α) are presented in table 2 for each of the four conditions of table 1. The empirical rates correspond rather closely to the nominal levels under either method.

As noted earlier, one may expect reduced power for the generic method. Power was evaluated under each method by generating data corresponding to the following three combinations of h^2 and n: $h^2 = .6$ with $n = 50$ and $h^2 = .2$ with $n = 100$ and $n = 200$. The hypothesis $h^2 = 0$ was tested using the LRT statistic described above. The estimated power, given by the empirical rejection rate obtained from 1,000 replications each, is presented in table 3 for $Q = .05$. It is evident that, as anticipated, the generic method is considerably less powerful than the specific one in all situations.

Table 2

Empirical Rejection Rates (i.e., Observed Type I Error) Corresponding to Several α's under the Specific and Generic Likelihood Methods

Q (n) and Method	Empirical Level Corresponding to an α of .01	.05	.10
.01 (50):			
Specific	.020	.065	.123
Generic	.014	.063	.127
.01 (200):			
Specific	.012	.062	.097
Generic	.007	.042	.087
.05 (50):			
Specific	.008	.064	.105
Generic	.016	.059	.105
.10 (50):			
Specific	.017	.064	.120
Generic	.015	.064	.124

NOTE.—The SEs of empirical rejection rates, evaluated at nominal levels (on the basis of 1,000 replications) are .003, .007, and .010, respectively, at α values of .01, .05, and .10.

Table 3

Empirical Power of the LRT for the Null Hypothesis $h^2 = 0$, Evaluated under Both the Specific and the Generic Likelihood Methods, for $Q = .05$ and Using $N = 1{,}000$

h^2 (n) and Method	Power Corresponding to an α of		
	.01	.05	.10
.2 (100[a]):			
Specific	0.50	0.74	0.84
Generic	0.18	0.38	0.51
.2 (200[b]):			
Specific	0.86	0.95	0.98
Generic	0.39	0.64	0.75
.6 (50[c]):			
Specific	0.99	1.00	1.00
Generic	0.75	0.89	0.94

NOTE.—The results for $h^2 = .2$ and $n = 50$, which resulted in many boundary solutions, are regarded as unreliable and are therefore not presented. The sample rejection rate due to boundary solutions, which never exceeded 0.2% in any other simulation condition, was >50% for this case.

[a] In each replication, there were 53, 32, and 15 families with 2, 3, and 4 children, respectively.

[b] In each replication, there were 110, 59, and 31 families with 2, 3, and 4 children, respectively.

[c] In each replication, there were 26, 15, and 9 families with 2, 3, and 4 children, respectively.

Truncation Region Unknown

We have seen that whenever the exact value of Q is known, the specific method yields more efficient parameter estimates and provides greater power for hypothesis testing than does the generic method. Since knowledge of the actual value of Q plays a fundamental role in formulating the likelihood function under the specific method (see eq. [7]), it is important to investigate the performance of the method when Q is unknown (or known with error), for the following reasons: First, if truncation is applied to raw phenotypes irrespective of concomitant variation, the actual value of Q will be unknown for the phenotype after adjustment for concomitant variation. Second, even if a specified level of truncation is postulated at the level of, say, age-sex-specific phenotypic distributions, deviations from strict adherence, as have occurred in the Lipid Research Clinics Program (1984*a*), will result in somewhat ambiguous values of Q for the final data. Third, at a theoretical level, one may question how much of the increased efficiency and power associated with the specific method are attributable to the knowledge of the actual value of Q. For these reasons, we performed another simulation experiment, generating samples of $n = 200$ with actual values of $Q = .05$, $h^2 = .6$, $c^2 = .2$, and $p = 0$. Each of the 1,000 replications was analyzed in seven different ways: once by using the generic method; once by using the specific method without making use of the actual value of Q (i.e., by maximizing the likelihood function in eq. [6] but without using the constraint given in eq. [7]); and in the remaining five cases by using the specific method with one of five different values of Q (.03, .04, .05, .06, and .07). The latter five cases, which include the true value of .05 for comparison, enable an assessment of the performance of the specific method when a precise value of Q is unknown.

The results are summarized in table 4, which presents the average estimates of the three correlations, their SD1's, average LRT χ^2 values for testing the null hypothesis $p = 0$, and the empirical rejection rates corresponding to $\alpha = .05$. Several important features emerge. First, when a specific value of Q is not assumed—and therefore the constraint in equation (7) not utilized—both methods give very similar results (see the first two rows in table 4). Second, so long as a specific value of Q—and therefore equation (7)—is utilized, the specific method yields considerably more efficient parameter estimates (as judged by SD1) even if an incorrect value of Q is used. Third, use of an incorrect value of Q results in a considerably higher type I error and introduces appreciable (systematic) bias into the parameter estimates; the bias is within ~1 SD if the error in Q is $\leq 20\%$ ($Q = .04$ or .06), and it is within ~2 SDs if the error in Q is $\leq 40\%$ ($Q = .03$ or .07). Fourth, the absolute bias and the average likelihood-ratio χ^2 value are both U-shaped as functions of Q, with the minimum values occurring at the true value of $Q = .05$; it is not known whether this observation would also apply to single studies, which reflect greater sampling variability, as opposed to averages over many replications, which are more stable. To investigate these properties under more extreme truncation, we repeated the experiment with a true value of $Q = .01$ and using only 500 replications. As seen in table 5, essentially the same conclusions hold even under extreme truncation.

Discussion

We have discussed two alternative maximum-likelihood methods, both based on the assumption of

Table 4

Average Estimates of the Three Familial Correlations, Their SD1's, Average χ^2_1 for Testing the Null Hypothesis $p = 0$, and Empirical Rejection Rate at $\alpha = .05$ in 1,000 Replications

Method of Analysis and Q	Average Estimates (SD1[a])			(LRT of H_0: $p = 0$) Average χ^2_1	Empirical Rejection Rate at $\alpha = .05$
	ρ_1	ρ_{12}	ρ_2		
Generic	−.000 (.083)	.299 (.056)	.499 (.045)	0.95	.046
Specific:					
Unknown	.000 (.080)	.299 (.054)	.499 (.044)	0.93	.038
.03	.118 (.046)	.381 (.027)	.555 (.029)	8.00	.754
.04	.052 (.047)	.335 (.028)	.523 (.031)	2.29	.208
.05	−.003 (.048)	.297 (.029)	.497 (.032)	1.02	.047
.06	−.050 (.049)	.264 (.029)	.476 (.033)	2.03	.164
.07	−.090 (.050)	.236 (.030)	.458 (.034)	4.24	.435

Note.—Each replication, consisting of $n = 200$ families, contained 107, 54, and 39 families with 2, 3, and 4 children, respectively. Each replication, generated with $\rho_1 = 0$, $\rho_{12} = .3$, $\rho_2 = .5$ and $Q = .05$, was analyzed using each of seven methods of analysis.

[a] See Note to table 1.

multivariate normality, for multifactorial analysis of family data ascertained through truncation on the phenotypic distributions of probands. The generic method simply conditions the likelihood function of the phenotypes of nonprobands on the actual phenotypic value(s) of the proband(s) (see, e.g., Boehnke and Lange 1984). The *specific* method, on the other hand, conditions the likelihood function on the actual event that the proband's value is in a specific Q, such as beyond a certain threshold (Rao and Wette 1987).

We have seen that, when ascertainment is based on truncation and the actual value of Q is known, both methods yield nearly (asymptotically) unbiased estimates of parameters; and the empirical rejection rates (estimates of type I error rates) under either method are comparable to the nominal levels. On the other hand, the SEs of parameter estimates are approximately twice as large under the generic method—with a corresponding decrease in the power of hypothesis tests—as compared with those under the specific method. However, these conclusions apply only when the actual value of Q is known, for (1) when this value is not known, we have seen that both methods perform equally well and (2) when an inaccurate value is assumed, depending on the extent of error involved, the specific method can yield appreciable bias and high empirical rejection rates (type I error). The foregoing conclusions hold under both moderate and extreme levels of truncation. Finally,

Table 5

Average Estimates of the Three Familial Correlations, Their SD1's Average χ^2_1 for Testing the Null Hypothesis $p = 0$, and Empirical Rejection Rate at $\alpha = .05$ in 500 Replications

Method of Analysis and Q	Average Estimates (SD1[a])			Average χ^2_1 (LRT of H_0: $p = 0$)	Empirical Rejection Rate at $\alpha = .05$
	ρ_1	ρ_{12}	ρ_2		
Generic	.001 (.088)	.301 (0.060)	.499 (.044)	1.04	.042
Specific:					
Unknown	−.000 (.085)	.300 (0.058)	.499 (.044)	1.05	.052
.005	.104 (.039)	.373 (0.025)	.549 (.026)	7.64	.702
.0075	.042 (.041)	.329 (0.026)	.519 (.027)	2.01	.176
.01	−.005 (.042)	.297 (0.026)	.496 (.028)	0.94	.036
.0125	−.042 (.042)	.270 (0.027)	.479 (.028)	1.82	.156
.0150	−.074 (.042)	.248 (0.027)	.464 (.029)	3.64	.392

Note.—Each replication, consisting of $n = 200$ families, contained 105, 62, and 33 families with 2, 3, and 4 children, respectively. Each replication, generated with $\rho_1 = 0$, $\rho_{12} = .3$, $\rho_2 = .5$ and $Q = .01$, was analyzed using each of seven methods of analysis.

[a] See Note to table 1.

although the results presented here pertain only to direct truncation on the phenotypic distribution, we verified that the conclusions apply to the case when truncation is applied not to the phenotypic distribution directly but to that of a correlated quantitative trait (i.e., indirect truncation; Rao and Wette 1987).

The results must be put into perspective. The specific method clearly provides more efficient parameter estimates and increased power for hypothesis tests, so long as truncation is strictly enforced and the value of the truncation region is known accurately. However, the generic method is easier to implement and is therefore more appealing, especially when truncation is applied to correlated traits. As a more general method capable of handling a wider range of situations under single ascertainment, the generic method is clearly more robust to deviations from the purported sampling method. At a practical level, actual family studies are less likely to enforce truncation strictly and accurately, thus giving rise to potentially inaccurate estimates of the truncation region. In conclusion, then, the specific method should be reserved for those cases in which the truncation region is known with minimum error. Finally, the simulations presented here are somewhat limited in scope, especially since we used only one set of parameter values. Therefore, extrapolation of the results to other situations may not readily apply.

Acknowledgments

The authors are grateful to Mike Boehnke, Terri Beaty, and Ken Lange for helpful comments on an earlier draft of the manuscript. This work was partly supported by U. S. Public Health Service grants GM 28719, HL 33973, and MH 31302.

References

Beaty, T. H., and K. Y. Liang. 1987. Robust inference for variance components models in families ascertained through probands. I. Conditioning on proband's phenotype. Genet. Epidemiol. 4:203–210.

Boehnke, M., and K. Lange. 1984. Ascertainment and goodness of fit of variance component models for pedigree data. Pp. 173–192 *in* D. C. Rao, R. C. Elston, L. H. Kuller, M. Feinleib, C. Carter, and R. Havlik, eds. Genetic epidemiology of coronary heart disease: past, present, and future. Alan R. Liss, New York.

Cannings, C., and E. A. Thompson. 1977. Ascertainment in the sequential sampling of pedigrees. Clin. Genet. 12:208–212.

Dadone, M. M., S. J. Hasstedt, S. C. Hunt, J. B. Smith, K. O. Ash, and R. R. Williams. 1984. Genetic analysis of sodium-lithium countertransport in 10 hypertension prone kindreds. Am. J. Med. Genet. 17:565–577.

Glueck, C. J., R. W. Fallat, F. Millet, P. Gartside, R. C. Elston, and R. C. P. Go. 1975. Familial hyper-alpha-lipoproteinemia: studies in 18 kindreds. Metabolism 24:1243–1265.

Gulbrandsen, C. L., N. E. Morton, G. G. Rhoads, A. Kagan, and R. Lew. 1977. Behavioral, social and physiological determinants of lipoprotein concentrations. Soc. Biol. 24:289–293.

Hopper, J. L., and J. D. Mathews. 1982. Extensions of multivariate normal models for pedigree analysis. Ann. Hum. Genet. 46:373–383.

Janus, E. D., N. T. Phillips, and R. W. Carrell. 1985. Smoking, lung function, and α_1-antitrypsin deficiency. Lancet 1:152–154.

Lipid Research Clinics Program. 1984*a*. The Lipid Research Clinics Coronary Primary Prevention Trial Results. I. Reduction in incidence of coronary heart disease. JAMA 251:351–364.

———. 1984*b*. The Lipid Research Clinics Coronary Primary Prevention Trial Results. II. The relationship of reduction in incidence of coronary heart disease to cholesterol lowering. JAMA 251:365–374.

Lipid Research Clinics Program Family Study Committee. 1984. The Collaborative Lipid Research Clinics Program Family Study. I. Design and description of data. Am. J. Epidemiol. 119:931–943.

McGue, M., R. Wette, and D. C. Rao. 1987. A Monte Carlo evaluation of three statistical methods used in path analysis. Genet. Epidemiol. 4:129–155.

Rao, C. R. 1973. Linear statistical inference and its applications. 2d ed. John Wiley & Sons, New York.

Rao, D. C., and R. Wette. 1987. Non-random sampling in genetic epidemiology: maximum likelihood methods for truncate selection. Genet. Epidemiol. 4:357–376.

Rao, D. C., M. McGue, R. Wette, and C. J. Glueck. 1984. Path analysis in genetic epidemiology. Pp. 35–81 *in* A. Chakravarti, ed. Human population genetics: the Pittsburgh Symposium. Van Nostrand Reinhold, Stroudsburg, PA.

Simpson, J. M., P. J. Brennan, C. A. McGilchrist, and R. B. Blacket. 1981. Estimation of environmental and genetic components of quantitative traits with application to serum cholesterol levels. Am. J. Hum. Genet. 33:293–299.

Third, J. L. H. C., J. Montag, M. Flynn, J. Freidel, P. Laskarzewski, and C. J. Glueck. 1984. Primary and familial hypoalphalipoproteinemia. Metabolism 33:136–146.