Am J Hum Genet 39:274–281, 1986

Brief Communication

A Note on Cannings and Thompson's Sequential Sampling Scheme for Pedigrees

SUSAN E. HODGE[1] AND MICHAEL BOEHNKE[2]

SUMMARY

We consider sequential sampling of pedigrees for genetic analysis. Cannings and Thompson (1977) gave simple, general guidelines for valid sequential sampling schemes. We show that their formulation of the likelihood contains an error, which is, however, easily corrected so as to maintain the validity of the sequential scheme. We also point out that although sequential and fixed-structure pedigree sampling do have the same likelihoods (as Cannings and Thompson showed), and therefore yield the same maximum likelihood point estimates of genetic parameters, they do not necessarily yield the same significance tests or confidence intervals.

INTRODUCTION

In genetic studies, the problem of how to sample extended pedigrees, as opposed to nuclear families, is extremely complex. Yet with the advent and widespread use of computers, large pedigrees are being increasingly analyzed by human geneticists because of the greater genetic information they provide. Thus, the pedigree sampling problem is taking on increasing practical as well as theoretical importance.

Received October 30, 1985; revised March 12, 1986.

This work was supported in part by grants AM-31813, AM-26844, CA-26803, HL-24489, and HD-04612 from the National Institutes of Health and by Research Career Development Award AM-01145 from the National Institutes of Health (S. E. H.).

[1] Departments of Biomathematics and Psychiatry, UCLA School of Medicine, Los Angeles, CA 90024.

[2] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109.

Two distinct sampling issues require attention when pedigree data are analyzed. First, pedigrees are often ascertained through probands. This ascertainment should be taken into account in a genetic analysis. For nuclear families, methods for correcting for ascertainment through probands are well established [1]. Aspects of correcting for ascertainment in pedigrees have been considered by Elston and Sobel [2], Hodge et al. [3], and Lalouel and Morton [4], among others.

Second, pedigrees are often sampled sequentially; that is, rather than deciding in advance that every sampled pedigree will include the siblings, parents, and grandparents (for example) of a set of randomly selected probands, we might instead choose to sample the first-degree relatives of a proband only if the proband is affected, to sample the maternal grandparents of the proband only if the proband and his or her mother are affected, and so forth. Such a sequential approach is consistent with the idea that studying a potentially genetic trait is best carried out in pedigree (branches) in which the trait is segregating.

Cannings and Thompson [5] proposed ingenious and aesthetically appealing methods to deal with both ascertainment and sequential sampling in pedigrees. First, they showed that in cases of single ascertainment—that is, when π is small and there is only one proband in the pedigree—one can correct for ascertainment by simply conditioning on the phenotype of that single proband. Boehnke and Greenberg [6] pointed out that this method does indeed work only for single ascertainment.

Second, Cannings and Thompson [5] defined rules for *sequential* sampling of extended pedigrees that are simple and quite general and that permit realistic sampling schemes. These rules, as stated by Cannings and Thompson, are; "(i) the choice of individuals to be examined next depends only on the phenotypes *already* observed, . . . and (ii) *all* individuals whose types are examined are included in the analysis" (p. 209). If a sequential sampling scheme meets these criteria, then no correction for it in the analysis is required, according to the authors. However, we believe that this aspect of their paper contains one error and one possible source of confusion. The error occurs in their formulation of the likelihood for the sequential sampling scheme. The possible source of confusion could arise when forming confidence intervals or carrying out significance tests on data collected sequentially. Neither the error nor the source of confusion invalidates the sequential scheme itself, but they do show that care must be taken to formulate the likelihood correctly and use it appropriately.

CORRECT FORMULATION OF THE LIKELIHOOD

In this part of our paper, we: (1) restate the notation and terminology used by Cannings and Thompson [5], (2) show where Cannings and Thompson's likelihood formulation is in error, (3) discuss the role of the auxiliary information random variables Y_n, and (4) discuss the implications of the error for sequential sampling of pedigrees.

1. *Notation and Terminology*

Let θ denote the genetic model, that is, genetic assumptions plus parameters. We consider only the "case of a random proband," that is, where the first person examined in the pedigree is chosen at random from the population and need not necessarily be affected with the trait of interest. However, the same argument is valid when one considers sequential sampling beginning with probands of a specific phenotype. C_1, C_2, . . . , represent classes (groups) of individuals observed sequentially. The first class, C_1, is assumed to consist of only one individual, the "random proband." We assume that no additional probands appear subsequently. X_n denotes the (ordered set of) phenotype(s) of the individual(s) in C_n. For ease of notation, let $C(n)$ denote $(C_1, \mathrm{C}_2, \ldots, C_n)$ and $X(n)$ denote $(X_1, X_2, \ldots, X_n)$. Then let g denote the rule used to choose the next class of individuals to be sampled: $C_n = g[C(n-1), X(n-1)]$. The stopping rule is that if $g[C(n-1), X(n-1)]$ is the empty set, then the sampling procedure ends.

Cannings and Thompson introduce an additional set of auxiliary information random variables, Y_1, Y_2, . . . , which are assumed to be independent of the model θ. For example, Y_n might reflect the practical availability of subsequent pedigree members or costs of testing or diagnosis. The Y_n are not essential to our main argument; in fact, the conceptual outlines of the discussion will be clearer without them. Thus, we shall omit the Y_n here and return to them later.

Given this notation, the likelihood of the model is simply (proportional to) the joint probability (probability density function in the continuous case) of all the data given the model:

$$\mathrm{L}(\theta) = P_\theta\,[C(N), X(N)]\ , \tag{1}$$

where N is defined as the first integer such that $C(N+1)$ is the empty set. Equation (1) corresponds to the equation on line 12, p. 211, of Cannings and Thompson [5] but without the auxiliary information random vector $Y(N+1)$. They claim that equation (1) is equivalent to (that is, proportional to as a function of θ) the product of conditional probabilities

$$P_\theta(X(1)|C(1)) \prod_{n=2}^{N} P_\theta[X_n|X(n-1), C(n)]\ , \tag{2}$$

which, in turn, equals the conditional probability

$$P_\theta(X_1, \ldots, X_N|C_1, \ldots, C_N),\ \text{that is,}\ P_\theta[X(N)|C(N)]\ . \tag{3}$$

Equation (3) corresponds to the last equation in Cannings and Thompson's section entitled "Case of a Random Proband," and equation (2) corresponds to the equation just above it. Note that these arguments will apply for both discrete and continuous traits.

2. *Error in Likelihood Formulation*

We maintain that equation (1) is indeed proportional to equation (2) and thus that equation (2) is a correct expression of the likelihood of the model, but that equation (3), the final formulation in Cannings and Thompson's derivation, is incorrect. For one thing, a conditional probability and a joint probability are not, in general, equal, unless the event being conditioned on—here, $C(N)$—has probability unity (or, equivalently, constant likelihood with respect to θ). Precisely in a *sequential* scheme this is not the case.

We now demonstrate formally where the error arises, then illustrate with a simple example. To see why equation (2) does not equal equation (3), note that by the definition of conditional probability,

$$P_\theta[X_n|X(n-1), C(n)] = \frac{P_\theta[X(n)|C(n)]}{P_\theta[X(n-1)|C(n)]} \; . \tag{4}$$

Substituting equation (4) into equation (2) and writing out the product gives

$$P_\theta[X(1)|C(1)] \cdot \frac{P_\theta[X(2)|C(2)]}{P_\theta[X(1)|C(2)]} \cdot \frac{P_\theta[X(3)|C(3)]}{P_\theta[X(2)|C(3)]} \cdots \frac{P_\theta[X(N)|C(N)]}{P_\theta[X(N-1)|C(N)]} \; . \tag{5}$$

For equation (5) to equal equation (3) would require a telescoping product, with each term $P_\theta[X(n-1)|C(n)]$ in the denominator canceling against a term $P_\theta(X(n-1)|C(n-1)]$ in the numerator. However, in this sequential sampling approach, $C_n = g_\theta[C(n-1), X(n-1)]$, so that C_n may well depend on X_{n-1}, and these terms do not cancel; that is:

$$P_\theta[X(n-1)|C(n)] \neq P_\theta[X(n-1)|C(n-1)] \; . \tag{6}$$

As an example, consider the following simple but instructive sampling scheme: (1) Sample a random proband. (2) If the proband is affected, sample his or her parents. (3) Stop. Here C_1 is the random proband, C_2 is either the empty set or the parents of the proband, and C_3 is the empty set. Consider a trait that occurs at random in the population with frequency $0 < \theta < 1$; that is, the trait is not familial, there is no transmission, and all phenotypes are independent. Then for $n = 2$,

$$P_\theta[X(1) = \mathrm{i}|C(1)] = \begin{cases} \theta \text{ for } i = \text{affected} \\ 1 - \theta \text{ for } i = \text{unaffected} \end{cases} ,$$

but

$$P_\theta[X(1) = i|C(2) = A] = \begin{cases} 1 \text{ for } i = \text{affected}, A = \text{parents of proband,} \\ \quad \text{or } i = \text{unaffected}, A = \text{empty set} \\ \\ 0 \text{ for } i = \text{affected}, A = \text{empty set,} \\ \quad \text{or } i = \text{unaffected}, A = \text{parents of proband} \; . \end{cases}$$

Thus, $P_\theta[X(1)|C(1)] \neq P_\theta[X(1)|C(2)]$.

This example illustrates what we pointed out above, namely, that the individuals sampled at stage n, C_n, may depend on the previously collected phenotypes X_{n-1}; that is, after all, the whole point of a sequential sampling scheme. The very fact that C_2, in this example, consists of the parents, and is not the null set, tells us that X_1 *must* be affected. Nor is this demonstration of the inequality (6) due merely to the simplicity of the example. We can imagine more complex and realistic sampling schemes, for example, where one continues sampling first-degree relatives of whoever is affected among those sampled so far. Then the fact that, say, the maternal grandparents of the proband were sampled would necessarily imply that the proband and the proband's mother must be affected.

3. *The Role of the* $\mathrm{Y_n}$

Let us return now to the Y_n, which represent auxiliary information. Some care must be taken in defining them. It is not really sufficient to say, as Cannings and Thompson [5] do, that "we observe a random variable Y_n which is independent of the underlying model θ" (p. 210). One must also specify explicitly, as part of the assumptions of the sampling scheme, that the probability of each X_n, conditioned on $X(n - 1)$ and $C(n - 1)$, is independent of $Y(n)$; that is,

$$P_\theta[X_n|Y(n), C(n), X(n - 1)] = P_\theta[X_n|C(n), X(n - 1)] \; . \qquad (7)$$

Otherwise, the final formulation of the likelihood, equation (2), is not correct as it stands, and the terms inside the product sign must be replaced by terms including $Y(n)$, as on the left side of equation (7). For example, Y_n could indicate where geographically the candidates for C_n live; this geographic location could influence not only availability of family members for study but also the nature or the frequency of the disease itself, as with multiple sclerosis [7]. Yet the probability of Y_n itself would still be independent of θ, as required. Of course, in this example, geographic location could have been included in the original model θ, which would circumvent this difficulty. Our point is simply that the requirement "Y_n is independent of the model θ" is not mathematically as restrictive as equation (7), and that equation (7) is what is needed to validate equation (2).

4. *Implications of the Error*

We maintain that the sequential sampling scheme can still be valid, as long as equation (2) is used, not equation (3) (and subject to the limits on interpretation to be discussed in SEQUENTIAL VS. FIXED-STRUCTURE SAMPLING, below). Moreover, we surmise that most users (investigators, computer programmers, etc.) probably have implemented the correct formulation, equation (2) not equation (3), whether or not they are aware of this fact. We now justify these statements.

The reason this error does not invalidate the sequential sampling scheme is that the error apparently arose out of ambiguities in the notation, not out of the sequential sampling concept as such. Simply, equation (3) incorrectly collapses the *chain* of conditional probabilities into a *single* conditional probability, by conditioning all the phenotypes on the *whole* pedigree structure, instead of conditioning each stage of phenotypes on the structure sampled *so far*. This latter conditioning process is what equation (2) correctly expresses.

Moreover, the formulation in equation (2) corresponds to what investigators presumably actually do when they put together the likelihood of a pedigree, whether sequentially or not. To illustrate, consider again the simple example presented above. Consider the event (C_1 = random proband, C_2 = parents of proband; X_1 = affected, X_2 = affected × unaffected). Following equation (2), one finds first the probability $P_\theta(X_1|C_1)$, which is θ; then since the trait occurs at random in the population, $P_\theta[X_2|X(1), C(2)] = 2\theta(1 - \theta)$, so the likelihood of the model, equation (2), is $2\theta^2(1 - \theta)$. However, by equation (3), the incorrect formulation, the conditional probability $P_\theta(X_1,X_2|C_1,C_2)$ would have to be *conditioned* on the fact that the proband's parents were included and would be only $2\theta(1 - \theta)$. This differs from the correct probability, equation (2), by a factor of θ—due in this example to conditioning X_1 on C_1 *and* C_2. Again, the validity of the counter-example does not rely on the simplicity of the sampling strategy.

We believe most investigators would almost "instinctively" calculate the correct $2\theta^2(1 - \theta)$ for this example, not $2\theta(1 - \theta)$, even if they thought they were implementing equation (3), as given by Cannings and Thompson. Thus, we do not anticipate that computer programs implementing Cannings and Thompson's sequential scheme will need to be corrected. However, clearly this is something that users of the sequential sampling scheme will need to check themselves.

SEQUENTIAL VS. FIXED-STRUCTURE SAMPLING

The possible source of confusion to which we alluded in the INTRODUCTION arises when we turn from the likelihood and maximum likelihood point estimation to the use of likelihood methods in significance tests and confidence intervals.

Consider a given set of phenotypic observations on a pedigree. Whether that pedigree was obtained under a sequential sampling scheme as discussed above, or under a fixed-pedigree-structure scheme, the likelihood (2) [not equation (3), as discussed above] will be identical. So will the maximum likelihood estimate (MLE) of θ. If one wishes only to determine what the data reveal about the genetic hypotheses and parameters—for example, if one is interested only in the maximum likelihood parameter estimate—then one need not be concerned with how the data were sampled.

However, error and confusion can arise if one proceeds to form classical Neyman-Pearson confidence intervals about the estimate $\hat{\theta}$ or to carry out significance tests on genetic hypotheses. Such calculations incorporate not only the information yielded by the data themselves, and thus contained in the likelihood (1) or (2), but also assumptions about the sampling space, that is, about what *could have* happened had we repeated the same experiment a large number of times. Clear exposition of this point can be found in Edwards [8]. For confidence intervals or significance levels, knowledge of how the pedigree was sampled, whether by a sequential or fixed-structure scheme, is relevant.

The width of a confidence interval and the significance level of a test will differ, depending on the sampling scheme.

We do not need to belabor the point. It has been amply discussed in numerous classical books on mathematical statistics and on the philosophy of scientific inference (e.g., [8–11]); moreover, constructing simple examples to illustrate the point is easy. Nor are we advocating *either* that one use only the likelihood *or* that one also form confidence intervals and perform significance tests. We merely want to remind the reader that the fact that two likelihoods are equivalent does not imply that all other statistical applications are necessarily identical.

We also want to make two further comments, one concerning significance tests in finite samples and the other concerning the asymptotic behavior of significance tests. By "finite" we mean that the number of pedigrees sampled, which we denote m, is finite, whereas by asymptotic we mean as m approaches infinity. Note that whether m is finite or infinite is a separate issue from how n, the number of individuals or classes of individuals in the pedigree, is determined.

Significance Tests in Finite Samples

To determine these exactly for a sequential sampling scheme would be nontrivial and would no doubt require extensive simulations. The elegant theory of sequential sampling pioneered by Wald [12] and carried further by Ghosh [13] and other writers would not be applicable in the exceedingly complex situations presented by pedigrees. However, in fairness, exact finite-sample significance levels have not been determined for fixed-structure sampling schemes, either, in most situations. Rather, investigators generally take advantage of the asymptotic properties of the MLE and the likelihood-ratio (LR) statistic (see next paragraph).

Asymptotic Significance Levels

Under fixed-structure sampling, and assuming that all the pedigrees are chosen to be of the same structure, then the pedigree phenotype vectors are independent and identically distributed (i.i.d.). It is well known that under mild regularity conditions, the MLE is asymptotically normally distributed, and the LR statistic is asymptotically distributed as chi-squared (e.g., [11]). As mentioned above, most investigators testing genetic hypotheses utilize this fact, even though no one really knows how large a sample (m) is needed to be considered "asymptotic." However, under a sequential sampling scheme, the pedigrees are clearly not i.i.d., and it is not clear to us that these asymptotic properties continue to hold. Even if they do, requisite sample sizes may not be the same as under a fixed-structure scheme.

DISCUSSION

In conclusion, we have shown that the likelihood (3) given by Cannings and Thompson [5] for a sequential sampling scheme is incorrect, but that an earlier equation (2) in their derivation is correct. Fortunately, we believe that readers and users will probably have used the correct formulation, although this supposition remains to be confirmed. We have also pointed out that equivalence in likelihoods under different sampling schemes does not necessarily imply equal significance tests or confidence intervals. Specifically, significance tests derived under the assumption of fixed-pedigree-structure sampling may not be appropriate for pedigrees sampled sequentially. Thus, Cannings and Thompson's claim that no correction in the analysis is required under a properly defined sequential sampling scheme is correct only if understood in the right context.

Finally, by no means do we wish to denigrate the value of sequential sampling schemes of the type suggested by Cannings and Thompson [5]. For extended pedigrees, fixed-structure schemes are simply impractical for most studies, however tractable their statistical properties may be. Sequential schemes, in contrast, are practical. Moreover, they coincide with the way that many genetic studies are actually done. The two rules specified by [5] (see INTRODUCTION) are straightforward and easily implemented, and Cannings and Thompson have done a service in stating them explicitly. By correcting the proof in [5] and pointing out that their result extends to point estimates but not to confidence intervals or significance tests, we hope to have clarified complicated but important issues in the genetic analysis of complex pedigrees.

ACKNOWLEDGMENTS

Our thanks go to Dr. Kenneth Lange, for encouraging us in this work and critically reading the manuscript, and to Dr. C. Cannings for helpful suggestions on an earlier version of the paper. Avis Williams typed the manuscript with speed and accuracy.

REFERENCES

1. MORTON NE: Genetic tests under incomplete ascertainment. *Am J Hum Genet* 11:1–16, 1959
2. ELSTON RC, SOBEL E: Sampling considerations in the gathering and analysis of pedigree data. *Am J Hum Genet* 31:62–69, 1979
3. HODGE SE, GREENBERG DA, ROTTER JI, LANGE KL: Second-order approximations of ascertainment probabilities. *Biometrics* 36:27–33, 1980
4. LALOUEL JM, MORTON NE: Complex segregation analysis with pointers. *Hum Hered* 3:312–321, 1981
5. CANNINGS C, THOMPSON EA: Ascertainment in the sequential sampling of pedigrees. *Clin Genet* 12:208–212, 1977
6. BOEHNKE M, GREENBERG DA: The effects of conditioning on probands to correct for multiple ascertainment. *Am J Hum Genet* 36:1298–1308, 1984
7. HAILE RW, HODGE SE, ISELIUS L: Genetic susceptibility to multiple sclerosis: a review. *Int J Epidemiol* 12:8–16, 1983
8. EDWARDS AWF: *Likelihood.* Cambridge, England, Cambridge Univ. Press, 1984
9. HACKING I: *Logic of Statistical Inference.* Cambridge, England, Cambridge Univ. Press, 1965
10. COX DR, HINKLEY DV: *Theoretical Statistics.* London, Chapman and Hall, 1974
11. KENDALL MG, STUART A: *The Advanced Theory of Statistics,* vol 2. New York, Hafner, 1974
12. WALD A: *Sequential Analysis.* New York, Dover, 1973 (originally published by John Wiley in 1947)
13. GHOSH BK: *Sequential Tests of Statistical Hypotheses.* Reading, Mass., Addison-Wesley, 1970