Sampling Considerations in the Gathering and Analysis of Pedigree Data

R. C. Elston[1] and E. Sobel[2]

SUMMARY

A general expression for the likelihood of a set of phenotypic observations on a randomly sampled pedigree, suitable for a wide variety of genetic models, has been previously modified to allow for independent ascertainments via probands. In this paper, further allowance is made for the fact that a pedigree usually contains some individuals who, whatever their phenotype, could never be probands, and we derive the limiting form of the likelihood appropriate for single ascertainment. The case when the sampling frame is ill-defined is discussed, and suggestions made for how to proceed in such a case.

A general expression for the likelihood of a set of phenotypic observations on a randomly sampled simple pedigree, suitable for a wide variety of genetic models, was derived by Elston and Stewart [1] and later extended to pedigrees of arbitrary structure by Lange and Elston [2]. Under certain circumstances, this likelihood can be used as a basis for analyzing pedigrees ascertained through probands, even though it is strictly appropriate only for randomly sampled pedigrees. If the number of probands is small compared to the total number of individuals with similar phenotypes, biases in the estimated segregation ratios will be small; biases in estimates of gene frequency and heritability, however, may remain relatively large. Although the estimated gene frequency was not appropriate for the population as a whole, this likelihood was successfully used to demonstrate segregation of a major gene in a large pedigree ascertained through four probands [3]. If, however, the number of probands in the sample is relatively large, as is usually the case whenever smaller pedigrees (such as

Received August 10, 1978.

This research was supported in part by a Public Health Service Research Scientist Award (MH31732) and research grant (MH26721) from the National Institute of Mental Health, and by a research grant (GM16697) from the National Institute of General Medical Sciences.

[1] Department of Biostatistics and the Genetics Curriculum, University of North Carolina, Chapel Hill, North Carolina.

[2] Statistical Consultant, Francis and Sobel Associates, P.O. Box 5940, Santa Fe, New Mexico 87502.

nuclear families, each ascertained via a proband), are pooled for analysis, it is essential to allow for the method of sampling in constructing the likelihood.

Traditionally, segregation analysis has been based on the likelihood of independently sampled sibships conditional on parental phenotypes, and methods of allowing for ascertainment in special cases of this situation are well known [4–8]. It was shown by Go et al. [9], however, that use of the unconditional likelihood of nuclear families and of larger pedigree structures leads to more efficient parameter estimates. Previous work [10, 11] has developed a method of modifying the unconditional likelihood of a pedigree to allow for independent ascertainments, assuming a given functional relationship between the probability of being a proband and an individual's phenotype. This method, however, has two drawbacks. First, it ignored the fact that a pedigree usually contains some individuals who, whatever their phenotype, could never be probands, for example, either because (1) they do not live within the sampling frame, or (2) their position in the pedigree and the method of sampling precludes them from being probands. Second, the expression given for the modified likelihood tends to the indefinite form 0/0 as the probability of becoming a proband becomes small (single ascertainment). The purpose of this paper is to show how these two drawbacks can be corrected when we have a rigidly defined sampling frame. It then becomes apparent that, with certain exceptions, there are inherent difficulties whenever the sampling frame is loosely defined. We discuss these difficulties and their implications for the sampling of pedigree data.

COMPLETELY KNOWN SAMPLING FRAME AND ASCERTAINMENT FUNCTION

Suppose there is a well-defined population in which every member has been measured for trait z. Although it is not necessary to measure every member, provided the potential for measurement is there, we nevertheless assume the measurement has been made; this allows a clear distinction between members and nonmembers of this population. If, for example, the probands are drawn from school children between the ages of 6 and 12, only such individuals are in the population, and we assume that no other child has been measured. The trait may be measured on a continuous scale, as the quantitative level of an enzyme, or as a discrete entity, that is, a dichotomous variable, affected ($z = 1$) vs. unaffected ($z = 0$). At this point, our statistical population or universe consists of individuals measured for phenotype z, and only those individuals can become probands. Furthermore, suppose there is a known function $\pi(z)$, which specifies the sampling probability that an individual with phenotype z in the population becomes a proband. In the traditional dichotomous model, $\pi(z)$ takes the form

$$\pi(z) = \begin{cases} \pi \text{ if } z = 1 \\ 0 \text{ if } z = 0, \end{cases} \tag{1}$$

where π is called the ascertainment probability. More generally, for a quantitative variable we can let

$$\pi(z) = \begin{cases} \pi \text{ if } z \geq T \\ 0 \text{ if } z < T \end{cases}$$

for some threshold value T; or $\pi(z)$ may be, for example, an exponential function of z (see reference [11]).

Once we have a set of probands, the complete sample is formed by bringing in phenotypic measures on their relatives; some of these relatives were perhaps in the original population, and others not (i.e., were not among those on whom we originally had phenotypic measures z); these latter are a sample of those who will be denoted simply as "who could not be probands," as opposed to the former, who could be (whether or not they actually were). Thus the phrase "who could not be probands" is taken to be synonymous with "who were not in the proband sampling frame."

We now assume that all the relevant distributional properties of z are the same for the two populations, those who could be probands and those who could not. If z depends on age (and/or sex), this is allowed for in the analysis, and so the relevant distributions are conditional on age (and/or sex); in this way it does not matter if there are age or sex differences between the two populations. Indeed, we may restrict the population of individuals who could be probands to one sex or age group. Similarly, all probabilities and likelihoods quoted are conditional on pedigree size and structure. It will be assumed that the distribution of z conditional on genotype does not depend upon pedigree size or structure, or on the individual's position in the pedigree; these are not trivial assumptions, but without them any rigorous analysis becomes very difficult. Under these assumptions, and conditional on the sampling method, we wish to derive the likelihood of observing the phenotypes of the members of a pedigree; the likelihood for more than one pedigree will then simply be the product of the likelihoods for each pedigree.

A pedigree occurs in the sample if and only if it contains at least one proband, so we want the likelihood of the pedigree phenotypes given that the pedigree does so. Writing this as $L(\text{pedigree}|\geq 1 \text{ proband})$, we have

$$L(\text{pedigree}|\geq 1 \text{ proband}) = \frac{L(\text{pedigree})L(\geq 1 \text{ proband}|\text{pedigree})}{L(\geq 1 \text{ proband})} \tag{2}$$

where $L(\text{pedigree})$ is the likelihood of the pedigree phenotypes assuming it to be randomly sampled from all pedigrees of the same size and structure, $L(\geq 1$ proband|pedigree) is the likelihood (or probability) that the pedigree, given the values of z for those who could be probands, contains at least one proband, and $L(\geq 1$ proband) is the likelihood (or probability) that an arbitrary pedigree of the same size and structure contains at least one proband among those whose particular positions are occupied by individuals in the original population of possible probands. Suppose there are n individuals in the pedigree who could be probands and m who could not; denote the phenotypes of the former z_i $(i = 1,2, \ldots, n)$, and those of the latter $z_j{}^*$ $(j = 1,2, \ldots, m)$. We know how to express L (pedigree) as a function of these $m + n$ values [1], so we concentrate here on expressions for the other two factors on the right side of equation (2).

Let the proband status of the ith individual who could be a proband be $b_i : b_i = 1$ if that individual is a proband, $b_i = 0$ if not (proband status is not defined for individuals who could not be probands). Now define an ascertainment function $\alpha(z, b)$ for each individual who could be a proband as follows: $\alpha(z,0) = 1 - \pi(z)$, $\alpha(z,1) = \pi(z)$ {i.e., $\alpha(z,b) = \pi(z)^b\,[1 - \pi(z)]^{1-b}$}. Then the joint probability of all the proband statuses observed, conditional on the phenotypes in the pedigree, is (assuming independent ascertainments)

$$L(\text{proband statuses}|\text{pedigree}) = \prod_{i=1}^{n} \alpha(z_i,b_i), \tag{3}$$

and

$$L(\geq 1 \text{ proband}|\text{pedigree}) = 1 - \prod_{i=1}^{n} \alpha(z_i,0) = 1 - \prod_{i=1}^{n} [1-\pi(z_i)]. \tag{4}$$

The denominator in equation (2) is obtained by summing the numerator over all possible phenotypes for each individual. It is given by

$$L(\geq 1 \text{ proband}) = \sum_{z_1} \cdots \sum_{z_n} \sum_{z_1^*} \cdots \sum_{z_m^*} L(\text{pedigree})L(\geq 1 \text{ proband}|\text{pedigree}). \tag{5}$$

(If z is continuous, the summations are replaced by integrations; this is also true in all that follows.) Now let $g_u(z)$ be the probability density function for the phenotype of an individual who has genotype u; if there are k genotypes involved, u can take on any one of k values. Thus we can associate k values $g_u(z_i)$ with individual i, and k values $g_u(z_j^*)$ with individual j. Then L(pedigree) can be expressed as a function of these $k(m + n)$ values [1], and we can write it as

$$L'(\{g_u(z_1)\},\{g_u(z_2)\}, \ldots ,\{g_u(z_n)\};\{g_u(z_1^*)\},\{g_u(z_2^*)\}, \ldots ,\{g_u(z_m^*)\}),$$

where within each pair of braces there are k values, one for each genotype. Substituting this and equation (4) into equation (5), we find (see also reference [11], where allowance is made for dependence on age and sex),

$$L(\geq 1 \text{ proband}) = 1 - L'(\{\zeta_u\},\{\zeta_u\}, \ldots , \{\zeta_u\}; \{1\},\{1\}, \ldots , \{1\}), \tag{6}$$

where there are n sets $\{\zeta_u\}$, each containing the k values given by

$$\zeta_u = \sum_z g_u (z)\alpha(z,0) = 1 - \sum_z g_u (z)\pi(z), u = 1,2, \ldots , k,$$

and there are m sets $\{1\}$, each containing k unities. Detailed derivations of ζ_u for various functions $g_u(z)$ and $\pi(z)$ are given in the appendix to reference [11].

COMPLETELY KNOWN SAMPLING FRAME, ASCERTAINMENT FUNCTION CONTAINING UNKNOWN PARAMETERS.

If the proband sampling frame and procedure is completely known, we can reasonably assume that the ascertainment function is also completely known. To develop the case where both the sampling frame and the ascertainment function are not completely known, we consider briefly the case in which the sampling frame is completely known but the ascertainment function is not, even though it is unlikely to occur in practice.

We assume the form of the ascertainment function is known, but that it depends upon one or more unknown parameters. These can be considered nuisance parameters, of little or no interest in themselves; they must, however, be jointly estimated when the likelihood is used for either estimating or testing hypotheses about other parameters in the model. The appropriate likelihood is thus the joint likelihood of observing the phenotypes of a pedigree and the proband statuses, conditional on the pedigree containing at least one proband. Using the same symbolism already used above, the required likelihood is thus:

$$L(\text{pedigree, proband statuses}|\geq 1 \text{ proband})$$

$$= \frac{L(\text{pedigree})L(\text{proband statuses}|\text{pedigree})L(\geq 1 \text{ proband}|\text{pedigree, proband statuses})}{L(\geq 1 \text{ proband})}$$

$$= \frac{L(\text{pedigree})L(\text{proband statuses}|\text{pedigree})}{L(\geq 1 \text{ proband})} \tag{7}$$

since, if the pedigree contains at least one proband, $L(\geq 1$ proband|pedigree, proband statuses) = 1. Thus equation (7) is similar to equation (2), the only difference being that expression (3) replaces expression (4). This is the same as the likelihood given in reference [11] if all individuals in the pedigree could be probands.

COMPLETE AND SINGLE ASCERTAINMENT

Complete ascertainment implies $\pi(z) = 1$ for certain values of z, and $\pi(z) = 0$ for all other values of z; and that this holds for all individuals in the original proband sampling frame. Since the pedigree contains at least one proband, expression (4) then becomes unity, thus simplifying the likelihood equation (2); furthermore, the evaluation of ζ_u in equation (6) will usually be simplified as well. Note that if certain individuals are affected with a disease, and because they are outside of the original sampling frame (e.g., due to their place of residence), could not be probands, this does not necessarily imply incomplete ascertainment; the likelihood we have developed allows for this in a natural manner.

Single ascertainment implies that $\pi(z)$ tends to zero, so both numerator and denominator of (2) tend to zero. Thus, when $\pi(z)$ is very small, numerical values of (2) can vary erratically for small changes in the parameter values, because of rounding errors. We therefore derive an expression for the limiting value of (2) as $\pi(z)$ tends to zero, to be used when this occurs or as an approximation, to reduce the computation necessary for some models. We assume that $\pi(z)$ is near zero because of the large pool of possible probands relative to the number of probands actually sampled. We consider an arbitrarily large, but finite, number N of pedigrees of a given size and structure, each of which has n possible probands in the same positions. Then to approximate single ascertainment, we let N tend to infinity.

Assume that $\pi(z)$ is of the form

$$\pi(z) = e^{-\delta(N)}h(z), \tag{8}$$

where $h(z)$ is independent of N, and $\delta(N) \to \infty$ as $N \to \infty$. The ascertainment functions discussed so far are of this form. Then, as $N \to \infty$,

$$\frac{L(\geq 1 \text{ proband}|\text{pedigree})}{\sum_{i=1}^{n} \pi(z_i)} \to 1. \tag{9}$$

Now the total pool of possible probands contains nN individuals. Suppose t of these are destined to become probands in the sample of N pedigrees, so that as $N \to \infty$ we can equate t/nN to the probability that an arbitrary individual from the proband pool is selected. Then, letting $G(z)$ be the cumulative distribution function of z in the population, as $N \to \infty$ we have

$$t/nN \to \int_{-\infty}^{\infty} \pi(z)\, dG(z) = E\,[\pi(z)] = e^{-\delta(N)}\, E\,[h(z)]. \tag{10}$$

But $L(\geq 1$ proband), the probability that a given pedigree is in fact sampled, is asymptotically equal to t/N. Thus, using equations (8), (9), and (10), we have the asymptotic result

$$\frac{L(\geq 1 \text{ proband}|\text{pedigree})}{L(\geq 1 \text{ proband})} \to \frac{\sum_i \pi(z_i)}{n\,E[\pi(z)]} = \frac{\sum_i h(z_i)}{n\,E[h(z)]}\,, \tag{11}$$

the summation being over all individuals who could be probands. Thus under single ascertainment, the desired likelihood (2) is L(pedigree) multiplied by the right side of equation (11), which we now investigate for special cases.

Case 1 is the traditional dichotomous model: $\pi(z) = \pi$ if $z = 1$, 0 if $z = 0$. Here we can put $h(z) = 1$ if $z = 1$, 0 if $z = 0$, and $\delta(N) = -1n\pi$, where $\pi = t/nN$. Thus in this case the numerator of equation (11) is the number of individuals in the sample who could be probands and who are affected, and the denominator is $nL(z = 1)$, where $L(z = 1)$ is the probability that a random individual is affected, expressed as a function of the parameters in the model.

Case 2 is the threshold ascertainment for a quantitative variable: $\pi(z) = \pi$ if $z \geq T$, 0 if $z < T$. This is similar to case 1, the denominator of equation (11) now being $nL(z > T)$.

Case 3 is the quadratic exponential ascertainment for a quantitative variable: $\pi(z) = e^{K_0 + K_1 z + K_2 z^2}$, where K_1 and K_2 must follow the constraints specified in reference [11]. Here we put $h(z) = e^{K_1 z + K_2 z}$ and $\delta(N) = -1n\, K_0$. The numerator of equation (11) is then easily expressed as a function of the sample values of z and the unknown parameters K_1 and K_2; and $E\,[h(z)]$ is expressed as a function of K_1,K_2 and the parameters in the model. (The details are similar to certain results given in reference [11].)

ILL-DEFINED SAMPLING FRAME

Although we may not have phenotypic measures of z on every individual in a well-defined sampling frame from which to choose probands, the likelihoods derived above provide a rigorous way of allowing for ascertainment when analyzing pedigrees of any structure, *provided each individual can be unequivocally classified as to whether he could or could not be a proband.* In certain special cases, this classification is unnecessary. The traditional sampling of sibships with single ascertainment is such a case, and comes about in the following way. Assume a dichotomous phenotype, as in case 1. Suppose we have a sibship containing n individuals who could be probands, s of whom are affected, and m individuals who could not be probands, r of whom are affected. Assume for simplicity there is only one parental mating type, with segregation probability p; and let π be the probability that an affected sib is a proband. (The probability that an unaffected sib is a proband is zero.) The likelihood (2) then becomes, apart from a constant multiplier,

$$\frac{p^{s+r}(1-p)^{n+m-s-r}[1-(1-\pi)^{s}]}{1-(1-\pi p)^{n}}$$

which $\rightarrow p^{s+r}(1-p)^{n+m-s-r}. \, s/np$ as $\pi \rightarrow 0$.

In this expression, s/np corresponds to the right side of equation (11), s being the number affected (out of n) and p the likelihood that $z = 1$ in the sibship. If we ignore the fact that m of the children could not be probands, we have for the corresponding expression (using p' and π' to indicate the parameters):

$$\frac{p'^{s+r}(1-p')^{n+m-s-r}[1-(1-\pi')^{s+r}]}{1-(1-\pi'p')^{n+m}}$$

and this $\rightarrow p'^{s+r}(1-p')^{n+m-s-r}. \, (s+r)/(n+m)p'$ as $\pi' \rightarrow 0$. Maximizing these limiting expressions for p and p', respectively, yields the same result. Thus in this case, there is no need to distinguish between those who could or could not be probands.

In the general case there are several problems in determining which individuals in a sample could be probands. Probands are often obtained from hospital records, and there is no difficulty in determining that certain individuals, because of their place of residence, could not be probands. But if more than one hospital serves the same area, and the records of one or more of these hospitals are not used for obtaining the probands, it may be very difficult to determine whether a given individual occupying a particular position in the pedigree, if he were to go to a hospital at all, would go to one of those used for selecting probands. One way to obviate this difficulty is to define a geographic area and use the records of *all* the hospitals serving that area to select probands and to restrict the selection of probands to individuals residing in the defined area. This method of sampling would, in theory, allow for a rigorous analysis without the need to assume complete ascertainment. It would, however, be necessary to define a relatively short time period over which admissions to the hospitals might make a person eligible to be a proband, and to determine for each individual in the sample his place of residence at that time. If the chosen time period is too long, it might be difficult to allow for those who had moved.

In addition to time and place, position in the pedigree (especially its effects on estimates of segregation ratios) is an important determinant of who could be probands. It is well known that ascertainment of a nuclear family through the parents leads to segregation ratios in the sibship different from those expected when a member of the sibship is a proband. Traditionally, inferences have been based on the likelihood of the sibs' phenotypes conditional on the parents' phenotypes. It is tacitly assumed that nuclear families are the sampling unit, and these are ascertained through *either* children *or* parents. But if we decide to sample nuclear families by taking probands, their sibs, and their parents, one of the parents could also be a proband, in which case there would be a three generational family in the data. If we assume a model in which the likelihoods of sibships conditional on the parental phenotypes are independent, then the likelihood of the three generational family can be factored into two such likelihoods (i.e., we can consider such a case as two independent sibships). But for this to be legitimate it is necessary, under the usual genetic models, for the actual genotypes of all the parents to be inferable from their phenotypes.

To avoid this difficulty, we can restrict probands to ages within a range shorter than the generation time. We could sample probands in the age range 20–30 years, for example, together with their sibs and parents; or within the same age range, generational families. (We can avoid one generational families by insisting on the availability of two generations as a further criterion for being a proband.)

With this method of sampling nuclear families, only individuals in the proband's generation could be probands. This suggests that perhaps possible probands should be restricted to the proband's generation for general pedigrees, a strategy that has been independently considered by T. E. Reich (personal communication, 1978). If we sample probands 20–30-years-old with their parents, sibs, spouses, and children, then only individuals in the middle (proband's) generation could be probands. But if we include uncles, aunts, nieces, or nephews of the proband, some of these could also be probands. It seems preferable to define those who could be probands by using a restricted age range (as well as time and place of residence), including the appropriate individuals from other generations, and excluding those individuals from the proband's generation not in the chosen age range. The probability of being a proband may depend upon the age of an individual at the onset of an illness, and this can be allowed for if necessary [10], but if probands are selected from a very restricted age range (e.g., 1 or 2 years), this will not be necessary. In any case, we have seen that when we can specify who could be probands, if the appropriate assumptions are valid and we use the appropriate likelihood (2) or (7), there is no need to allow any further for the fact that segregation ratios depend upon how the pedigree has been ascertained.

REFERENCES

1. Elston RC, Stewart J: A general model for the genetic analysis of pedigree data. *Hum Hered* 21:523–542, 1971
2. Lange K, Elston RC: Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Hum Hered* 25:95–105, 1975
3. Elston RC, Namboodiri KK, Glueck CJ, Fallat R, Tsang R, Leuba V: Study of the genetic transmission of hypercholesterolemia and hypertriglyceridemia in a 195 member kindred. *Ann Hum Genet* 39:67–87, 1975
4. Elandt-Johnson RC: Segregation analysis for complex modes of inheritance. *Am J Hum Genet* 22:129–144, 1970
5. Fisher RA: The affects of methods of ascertainment upon the estimation of frequencies. *Ann Eugen (Lond)* 6:13–25, 1934
6. Morton NE: Genetic tests under incomplete ascertainment. *Am J Hum Genet* 11:1–16, 1959
7. Morton NE: Segregation analysis, in *Computer Applications in Genetics,* edited by Morton NE, Honolulu, Univ. of Hawaii Press, 1969, pp 129–139
8. Stene J: Assumptions for different ascertainment models in human genetics. *Biometrics* 33:523–527, 1977
9. Go RCP, Elston RC, Kaplan EB: Efficiency and robustness of pedigree segregation analysis. *Am J Hum Genet* 30:28–37, 1978
10. Elston RC: Ascertainment and age of onset in pedigree analysis. *Hum Hered* 23:105–112, 1973
11. Elston RC, Yelverton KC: General models for segregation analysis. *Am J Hum Genet* 27:31–45, 1975