Am. J. Hum. Genet. 43:364–373, 1988

Conditioning on Subsets of the Data: Applications to Ascertainment and Other Genetic Problems

Susan E. Hodge

Departments of Biomathematics and Psychiatry, UCLA School of Medicine, Los Angeles

Summary

I here consider the question of when to formulate a likelihood over the whole data set, as opposed to conditioning the likelihood on subsets of the data (i.e., joint vs. conditional likelihoods). I show that when certain conditions are met, these two likelihoods are guaranteed to be equivalent, and thus that it is generally preferable to condition on subsets, since that likelihood is mathematically and computationally simpler. However, I show that when these conditions are not met, conditioning on subsets of the data is equivalent to introducing additional df into our genetic model, df that we may not have been aware of. I discuss the implications of these facts for ascertainment corrections and other genetic problems.

Introduction

A genetic model usually involves a population that is subdivided in several different ways, e.g., family size, mating type, number of affected children, etc. Moreover, not every individual of interest in a population is necessarily *ascertained* (i.e., comes to our attention). In many genetic applications, we may naturally view our data as being first broken down into *subsets* and then further divided into *categories within each subset.* For example, consider a set of nuclear families being prepared for segregation analysis. We may break them down first by sibship size and then consider how many children are affected within families of each sibship size. Here, the sibship sizes represent the subsets, and the number of children affected represent the categories within each subset. Or families might be broken down by parental mating type (subsets) and then by number of affected children (categories within a subset).

In such a situation, we have a choice about how to formulate the likelihood of whatever model we are investigating: We can either formulate the likelihood over the whole data set or condition the likelihood on the subsets of the data. The likelihood conditioned on subsets is usually simpler to formulate and less cumbersome to work with, and it may correspond more closely to how we perceive the problem at hand. But it is not always obvious which formulation is "better" and when.

In this paper I will show that, under certain conditions, the two likelihoods are guaranteed to be equivalent, in the sense of yielding identical maximum likelihood estimates (MLEs) of the parameters of interest. Yet, using the more cumbersome formulation (the likelihood over the whole data set) would require estimating additional parameters, which we are not generally interested in. This clearly represents unnecessary work, so in this situation it is preferable to use the simpler likelihood, the one conditioned on subsets.

However, in other circumstances, the two likelihoods are not equivalent. As a general rule, the likelihood conditioned on subsets will use less of the information available in the data set and will produce estimates with greater variances (and therefore standard errors) than will the likelihood over the whole data set. Which likelihood we choose will then depend on several factors, which will be discussed below.

The point is not to construct two equivalent likelihoods but to determine when two likelihoods are in fact equivalent.

Model and Notation

For the reader's convenience, the notation introduced here is also summarized in Appendix A.

Received February 11, 1988, revision received May 17, 1988.

Address for correspondence and reprints: Dr. Susan E. Hodge, Columbia University, Psychiatric Institute, 722 West 168th Street, Box 14, New York, NY 10032.

0002-9297/88/4304-0002$02.00

Model

We consider a general population which is partitioned into W subsets. We define Q_u as the probability that a member of the population is in subset u: $Q_u = P(\text{subset } u)$, for $u = 1, \ldots, W$. Since this is a partition, every member of the population is a member of one and only one subset, and $\Sigma\, Q_u = 1$.

Probabilities

The subset u contains $I(u)$ categories. We define p_{ui} as the conditional probability for a population member to be in category i, given that member is in subset u: $p_{ui} = P(\text{category } i|\text{subset } u)$. (In what follows, we will write simply I for $I(u)$, unless clarity requires more explicitness.)

Thus, the Q_u are probabilities of being in subset u; the p_{ui} are conditional probabilities of being in category i, given one is in subset u; and the product $Q_u p_{ui}$ is the probability of being in category i of subset u.

Ascertainment

What complicates the model is that not all population *members* end up becoming *observations*, because not all members are ascertained. By definition, we will allot to category $i = 0$ of each subset all those members that are not ascertained. Thus, $p_{u0} = P(\text{not ascertained}|\text{subset } u)$.

Therefore, whereas $\sum\limits_{i=0} p_{ui} = 1$, the sum $\sum\limits_{i=1} p_{ui}$ equals $1 - p_{u0}$, which in turn equals $P(\text{ascertained}|\text{subset } u)$. We call this second sum p_u; thus p_u is summed not over all i but only over $i = 1, \ldots, I(u)$.

To correct the conditional probabilities p_{ui} for ascertainment requires dividing by the probability of being ascertained. Within a subset, the probability of being ascertained is simply p_u, as above, so the corrected probability becomes

$$P(\text{category } i|\text{subset } u, \text{asc'd}) = \frac{p_{ui}}{p_u}. \tag{1}$$

Viewed over the entire data set, however, the probability of being ascertained is $P(\text{asc'd}) = 1 - \sum\limits_u Q_u p_{u0} = \sum\limits_u Q_u p_u$. Now the corrected probability becomes

$$P(\text{subset } u, \text{category } i|\text{asc'd}) = \frac{p_{ui}}{\sum\limits_u Q_u p_u}. \tag{2}$$

We use a parallel notation for observation: n_{ui} is the number of observations in subset u, category i. For convenience, n_u represents the total number of observations in subset u: $n_u = \sum\limits_{i=1} n_{ui}$; and n represents the total number of observations in the entire data set: $n = \sum\limits_u n_u$.

Let θ, which may be a vector, represent the parameter(s) of the model. We begin quite generally by letting both the population subset probabilities Q_u and the conditional probabilities p_{ui} be functions of θ. We will be explicit about this, writing $Q_u(\theta)$ and $p_{ui}(\theta)$ for these probabilities.

See Appendix A for a concise summary of the notation.

Likelihoods

We now indicate how to formulate the two different likelihoods described in the Introduction. For the likelihood over the whole data set, which we will call L_1, we take each probability $Q_u p_{ui}$ and raise it to the number of families observed in that subset category. To correct for ascertainment we use the same denominator as in equation (2). The resultant likelihood, L_1, is

$$L_1(\theta) = \frac{\prod\limits_u \prod\limits_{i=1} [Q_u(\theta) p_{ui}(\theta)]^{n_{ui}}}{[\sum\limits_u Q_u(\theta) p_u(\theta)]^n}. \tag{3}$$

To form the likelihood conditioned on subsets, which we call L_2, we create a separate likelihood for each subset. Each of these separate terms has a denominator as in equation (1). Then the separate likelihoods are multiplied together to yield L_2:

$$L_2(\theta) = \prod_u \frac{\prod\limits_{i=1} [p_{ui}(\theta)]^{n_{ui}}}{[p_u(\theta)]^{n_u}}. \tag{4}$$

Note that the subset probabilities $Q_u(\theta)$ appear in the likelihood over the whole data set (equation [3]) but not in the likelihood conditioned on subsets (equation [4]).

Equivalence

The equivalence result to be proved here states that although L_1 and L_2 do not yield identical estimates of θ, L_1 can be modified so as to give the same estimates of θ as L_2 does. This modification consists of introducing $W - 1$ separate, unconstrained parameters "t_u" into the model. These parameters t_u appear as terms multiplying the $Q_u(\theta)$ terms. It is easiest to write them as W new parameters with one constraint. Only $W - 1$ parameters are needed because the likelihood is unchanged if all t values are multiplied by a constant. When the t's are included, $L_1(\theta)$ becomes

$$L_1(\theta,t) = \frac{\prod_u \prod_{i=1} [Q_u(\theta)p_{ui}(\theta)t_u]^{n_{ui}}}{[\sum_u Q_u(\theta)p_u(\theta)t_u]^n} . \qquad (5)$$

By definition, two likelihoods are equivalent if they yield identical MLEs. That is, we are saying not only that $L_1(\theta,t)$ and $L_2(\theta)$ yield the same MLEs of θ asymptotically—or for "perfect" data that fit the parameters exactly—but also that no matter what the data are, the MLE of θ found from equation (5) will equal that found from equation (4).

The reader who is not interested in the proof may safely proceed to the next section, Applications, or to the example in Appendix C.

Equivalence Result

$L_1(\theta)$ in equation (3) and $L_2(\theta)$ in equation (4) are not equivalent in the sense just defined, but $L_1(\theta,t)$ in equation (5) and $L_2(\theta)$ are equivalent.

Proof.—We assume that the likelihoods are regular, i.e., that each likelihood is (at least) twice differentiable in the region around its maximum at θ. The argument starts with the likelihood conditioned on subsets, L_2, and then investigates how L_1 (θ) in equation (3) needs to be modified so as to yield the same MLE (θ) as $L_2(\theta)$ does. We proceed in three steps.

Step 1.—We first show that $L_1(\theta)$ can be written as the product of $L_2(\theta)$ and a multinomial likelihood, which we call $f(\theta)$. To do this, rewrite equation (3) as follows:

$$L_1(\theta) = \frac{\prod_u [Q_u(\theta)p_u(\theta)]^{n_u}}{[\sum_u Q_u(\theta)p_u(\theta)]^n} \cdot \frac{\prod_u \prod_i [p_{ui}(\theta)]^{n_{ui}}}{\prod_u [p_u(\theta)]^{n_u}} . \qquad (6)$$

The second term on the right of equation (6) is identical to $L_2(\theta)$; the first term on the right is a new term, which we will call $f(\theta)$. Note that $f(\theta)$ is in turn the product of W terms $\beta_u(\theta)$, raised to powers n_u, as follows:

$$f(\theta) = \prod_u [\beta_u(\theta)]^{n_u}, \text{ where } \beta_u(\theta) = \frac{Q_u(\theta)p_u(\theta)}{\sum_v Q_v(\theta)p_v(\theta)} . \qquad (7)$$

Thus, $f(\theta)$ is a multinomial likelihood. Each $\beta_u(\theta)$ represents the probability that an observation falls into subset u, given that it has been ascertained, i.e., given that it is in the data set. Thus, we have shown that

$$L_1(\theta) = L_2(\theta) \cdot f(\theta) , \qquad (8)$$

where $f(\theta)$ is as defined in equation (7).

Step 2.—We next show that $L_1(\theta)$ and $L_2(\theta)$ will not in general yield the same MLEs of θ. As is customary, we work with the log likelihoods. (The only time the log likelihood is not defined is when the likelihood is zero, and we can always find a sufficiently small region around the maximum where the likelihood is not zero.) From equation (8), we have

$$\log L_1(\theta) = \log L_2(\theta) + \log f(\theta) . \qquad (9)$$

The MLE of θ found from L_2 must satisfy $(\delta/\delta\theta_i)$ log $L_2(\hat{\theta}) = 0$, for all i; and, similarly, the MLE of θ from L_1 must satisfy $(\delta/\delta\theta_i)$ log $L_1(\hat{\theta}) = 0$ for all i. Thus, for these two MLES θ to be equal, it follows from equation (9) that $(\delta/\delta\ \theta_i)$ log $f(\hat{\theta})$ must equal 0 for all i as well, at that same value of θ. However, this will not in general be the case (although it may hold for certain well-specified values of the data; see Appendix C for an example).

We have shown that L_1 (θ) and L_2 (θ) do not necessarily yield the same MLE of θ. Thus, they are not in general equivalent.

Step 3.—We then show that when $W - 1$ unconstrained additional "nuisance parameters" t_u are introduced into $L_1(\theta)$, creating L_1 (θ,t) as in equation (5), L_1 (θ,t) and L_2 (θ) yield the same MLE θ. We introduce $W - 1$ new terms t_u into $f(\theta)$ and hence into the likelihood over the whole data set, L_1 (θ). We do this by inserting t_u into the numerator of each β_u and modifying the denominator accordingly:

$$\beta_u(\theta,t) = \frac{Q_u(\theta)p_u(\theta)t_u}{\sum_v Q_v(\theta)p_v(\theta)t_v} . \qquad (10)$$

Recall that there is one constraint on the t terms, so that although W t's appear in equation (10) they represent only $W - 1$ independent parameters. Now the likelihood over the whole data set becomes

$$L_1(\theta,t) = \frac{\prod_u [Q_u(\theta)p_u(\theta)t_u]^{n_u}}{[\sum_u Q_u(\theta)p_u(\theta)t_u]^n} \cdot \frac{\prod_u \prod_i [p_{ui}(\theta)]^{n_{ui}}}{\prod_u [p_u(\theta)]^{n_u}} . \qquad (11)$$

As in equation (6), the second term on the right of equation (11) is identical to $L_2\ (\theta)$, whereas the first term on the right, $f(\theta,\ t)$, is a function of both θ and t. Now equation (9) becomes

$$\log L_1\ (\theta,t) = \log L_2\ (\theta) + \log f(\theta,t)\ . \quad (12)$$

As discussed earlier, there must be exactly one constraint on the t_u's; for example, we could arbitrarily set t_1 to unity. There must be $W - 1$ unconstrained t's so as to match the $W - 1$ β values. (In cases where other constraints on θ allow it, since the β_u's represent the ratios of the $Q_u(\theta)\ p_u(\theta)$ terms to each other, each t_u will simply be estimated at whatever value is needed to make β_u equal the observed n_u/n. Also see Appendix C.)

We then estimate the t_u's along with θ, even though they are not themselves of any meaning or interest to our model. When we do this (see Appendix B), we find that $L_1\ (\theta,t)$ and $L_2\ (\theta)$ yield the same $\hat{\theta}$, for all values of the observed data.

Intuitively, this works because the t's introduce sufficient df so that the β's can be made to match the observed proportions n_u/n. See the example and applications below for further explanation. For a formal proof of step 3, see Appendix B.

Applications

The equivalence result proved above may seem remote when viewed in the abstract. We present here two applications in human genetics. These should help to clarify the implications of the equivalence result. In addition, Appendix C presents a simple numerical example involving tossed coins and balls in urns.

1. Application: Family-Size Distribution

In an earlier work, I considered the effects of the population family-size distribution (FSD) on a segregation analysis (Hodge 1985). There I defined the "separate-multinomials" likelihood as that formulation of the likelihood which treats each family size as representing a separate multinomial distribution and the "grand-multinomial" likelihood as that which treats the entire data set as representing one distribution. I showed that these two likelihood formulations are equivalent, in the same sense as used here, if and only if the population FSD is completely unknown.

I will show now that this earlier work represents a special case of the equivalence result presented here. The separate-multinomials likelihood is the same as the likelihood conditioned on subsets (L_2), and the grand-multinomial likelihood is the likelihood over the whole dataset (L_1). The subsets represent different family sizes, with categories representing the configurations of affected and unaffected individuals within each given family size. The subset probabilities $Q_u(\theta)$ represent the FSD in the population. Moreover, we assumed that the FSD is independent of the genetic parameters θ. That is, each $Q_u\ (\theta)$ is really just Q_u. Therefore, we can simply view each Q_u as *being* one of the nuisance parameters t_u. The single constraint is the fact that the Q_u terms must sum to unity. The conditional probabilities $p_{ui}(\theta)$ represent, for example, the probability of i affected children, given a u-child family. (More generally, $p_{ui}\ (\theta)$ is the probability of the ith configuration of affected and unaffected individuals, given that the family is in size class u).

The equivalence result presented here tells us that when the FSD is unknown, the likelihood conditioned on subsets (i.e., on family sizes) will yield the same estimate of θ as the more cumbersome likelihood does over the whole data set. This is because the $W - 1$ nuisance parameters are already present in the model. Thus, we may as well use the more tractable likelihood conditioned on subsets, as it makes no difference in the estimation of θ. This result is useful in human genetics, where we generally do not know the FSD with certainty, if at all.

However, it also follows that if we do know something about the distribution of family sizes, i.e., the Q_u's, then this additional knowledge will represent additional constraints on the t_u's, and the two formulations of the likelihood will not in general yield the same estimates of θ. Ewens et al. (1986) have shown that the resultant differences in estimates of θ are generally not large, but the point is that the MLEs of θ are not mathematically identical when anything is known about the FSD. Clearly, if we were in fact *interested* in the FSD, then the FSD parameters would be incorporated in θ, and the above analysis would not be relevant, since the Q_u's would no longer be independent of θ.

2. Application: Conditioning on "That Part of the Data Relevant to Ascertainment"

Background. — Ewens and Shute (1986) have proposed a method of ascertainment correction which is extremely nonparametric or assumption free. Unlike the classical Weinberg model of ascertainment (Weinberg 1928; Morton 1959), this approach makes no assumptions about equality of ascertainment probabilities or independence of ascertainment events. In fact, it circumvents the whole concept of probands and simply defines the probability that a *family* of a certain type is ascertained (since the family is the basic unit of genetic study). For the moment, we limit this discussion to families of the same size. However, the Ewens and Shute method is not limited to single family sizes. We will deal with the issue of different family sizes below, in connection with equation (16). Ewens and Shute partition the data conceptually into the part "relevant to ascertainment," which they denote d_1, and the part not relevant to ascertainment, denoted d_2. Thus, by definition, all families with the same value of d_1 have the same probability of being ascertained, which we denote $a(d_1)$ for the moment. For example, if we are studying nuclear families, we might determine that the "number of affected children" represents that part of the data relevant to ascertainment; then d_1 would equal the number of affected children.

The nonparametric nature of this approach arises from the fact that we do not need to know *what* the values of the $a(d_1)$'s are, because Ewens and Shute show that if the $a(d_1)$'s are unknown, then the likelihood of the genetic parameters is the same as if one simply conditioned on the value of d_1.

Rather than reiterate all their work, we will now express it in terms of "subsets" and "categories." We will show that Ewens and Shute's approach utilizes a special case of our equivalence result and then point out some limitations of the assumption-free approach.

Each value of d_1 represents a subset u, whereas further breakdowns of the data by d_2 are to be viewed as categories within a subset, i. To continue the example with nuclear families, if u represents the number of affected children, then i represents whatever else is contained in the data: affectedness status of the parents, linkage information if there is any, etc. The parameters of interest, θ, are still genetic (gene frequency, penetrance, recombination fraction, etc.) but explicitly exclude ascertainment. The population subset probabilities Q_u (θ) of subset u do in general include θ and do not correspond to the nuisance parameters t_u as they did in the FSD application.

Instead of $a(d_1)$, we write a_u to denote the probability that a family is ascertained, given it is in subset u. Thus, within each subset u, the probability of not being ascertained is $p_{u0}(\theta) = 1 - a_u$. The likelihood over the whole data set is

$$L_1(\theta) = \frac{\prod_u \prod_i [Q_u(\theta) p_{ui}(\theta) a_u]^{n_{ui}}}{[\sum_u Q_u(\theta) p_u(\theta)] a_u]^{n_u}} \quad \begin{array}{l} u = 1, \ldots, W; \\ i = 1, \ldots, I(u) \end{array} \tag{13}$$

The likelihood conditioned on subsets is

$$L_2(\theta) = \frac{\prod_u \prod_i [p_{ui})^{n_{ui}}}{\prod_u (p_u)^{n_u}} = \prod_u \prod_i \frac{p_{ui}}{p_u}^{\;\;n_{ui}} . \tag{14}$$

Ewens and Shute prove that equations (13) and (14) yield the same estimate of θ. We see that this is a special case of our equivalence result, with the a_u in equation (13) taking the place of the parameters t_u. However, we also see that the equivalence of equations (13) and (14) holds only when the ascertainment probabilities are *completely* unknown, without any constraints.

Limitations to the Ewens and Shute approach to correcting for ascertainment.—As Ewens and Shute point out, using equation (14) is the same as *conditioning* on "that part of the data relevant to ascertainment," i.e., conditioning on the subsets. Since this part of the data may contain a good deal of genetic information, (e.g., numbers of affected children), an obvious criticism of the method is that one is "conditioning out" much of one's data. However, the point of the equivalence between equations (13) and (14) is that if we do not know the a_u values, then we *may as well* condition on u; we will get precisely the same likelihood anyway. Thus, if we really do not know anything about ascertainment beyond the fact that it depends in some way on the value of u, this criticism is not valid.

However, the other side of the coin is that if we do in fact know something more about ascertainment than that, then conditioning on subsets u will *lose* that additional information. We consider two examples of this situation. The first is straightforward: Assume again that the ascertainment probability for a family depends only on the number of affected children, so that we define our subsets by "number of affected children." In this situation, although we do not know the actual value of the ascertainment probabilities a_u, we might well still believe that this probability should *increase* as the number of affected children increases, i.e.,

$$a_u < a_v \text{ for } u < v . \tag{15}$$

However, incorporating an assumption such as formula (15) into the likelihood would represent additional constraints on the a_u and would therefore invalidate the equivalence of equations (13) and (14). In other words, by using equation (14) and conditioning on subsets, we are unable to incorporate additional information or assumptions such as formula (15). The only way to incorporate formula (15) is to go back to the likelihood over the whole data set (equation [13]) and estimate the a_u along with the genetic parameters, θ. This same caveat would apply for any other assumptions we might want to include about relationships among the probabilities a_u.

The second example is more subtle. In our recapitulation of Ewens and Shute's results above, we ignored the issue of family size. However, most data sets contain more than one family size. Here is how the family-size issue enters the discussion. (We continue to illustrate the situation where d_1, that part of the data relevant to ascertainment, consists of the number of affected children.) Recall that the probabilities p_{ui} represent P(configuration u of parental phenotypes, marker data, etc. | i affected children). These probabilities in fact depend on the total sibship size, not just on the number of affected children. (For example, for a recessive disease P(1 affected parent | 1 affected child) is lower in a five-child family than in a one-child family.)

One solution—i.e., one way to allow for the FSD—would be to incorporate the population FSD into the p_{ui}, as follows:

$$P(\text{configuration } u \mid i \text{ affected children}) = \frac{\sum_s P(u \mid i \text{ aff. children}, s)\, P(i \text{ aff. children} \mid s)\, P(s)}{\sum_s P(i \mid s)\, P(s)} . \qquad (16)$$

Here, s represents sibship size and $P(s)$ is the population probability of an s-child family, i.e., the FSD. However, this solution is impractical. We do not generally know the FSD accurately, and Ewens and Asaba (1984) have shown that the estimation of genetic parameters is quite sensitive to the FSD when the latter is used in the likelihood; that is, parameters may be asymptotically biased if an incorrect FSD is used.

The simpler, more obvious solution to the FSD problem is to redefine our subsets so as to break them down by sibship size as well as by number of affected children. To be explicit, rather than defining the subsets as $u = 1$ denotes 1 affected child, $u = 2$ denotes 2 affected children, etc., we use the following:

$u = 1$ denotes two-child families with one affected child;
$u = 2$ denotes two-child families with two affected children;
$u = 3$ denotes three-child families with one affected child; and, in general,
$u = v$ denotes s-child families with a affected children, where $v = s(s - 1)/2 + a - 1$.

(17)

This solution is in fact what Ewens and Shute do, since they condition on sibship size as well as on number of affected children; see, for example, their equation (6). The disadvantage is that this solution results in conditioning on a much larger number of subsets than we may have anticipated—and hence in introducing that many more nuisance parameters t_u or df. For example, say our maximum sibship size is five. Then by the scheme in equation (17), the total number of subsets is 14, not five.

To put this another way: Recall that every subset has its own probability of ascertainment, a_u or t_u, and that we cannot constrain these probabilities in any way. Now, a reasonable constraint which even a fairly assumption-free ascertainment theory might want to encompass is that the probability of ascertainment for a family should depend on the number of affected children but not additionally on the number of *un*affected children. In other words, a reasonable constraint is that P(family ascertained | i affected children, s) should equal P(family ascertained | i affected children), without the s. However, we cannot incorporate even this constraint into the Ewens and Shute method. P(family ascertained | 1 affected child, $s = 2$) and P(family ascertained | 1 affected child, $s = 3$) represent ascertainment probabilities for two different subsets and cannot be constrained to be equal.

Discussion

In this paper we have examined two different formulations of the likelihood and have shown that they are not in general equivalent, in the sense of yielding the same estimates of θ. The likelihood over the whole data set, L_1 (θ) in equation (3), tends to be more cumbersome, and the likelihood conditioned on subsets, L_2 (θ) in equation (4), tends to be simpler to work with.

We have also shown that if a sufficient number of "nuisance parameters," t_u, are present in or are introduced into the model—so that instead of L_1 (θ) in equation (3) we have L_1 (θ,t) in equation (5)—then L_1 (θ, t) and L_2 (θ) *are* equivalent. In these situations, there is no reason not to use the simpler formulation L_2 (θ). However, in situations where the t_u terms are not present, the more cumbersome L_1 (θ) will, in general, yield more information (in the sense of lower variances of the estimates) than will L_2 (θ). The user will need to decide, in each particular case, whether the gain in information is worth using the more awkward formulation.

I have shown that L_1 (θ,t) in equation (5) is equivalent to L_2 (θ) in equation (4) if and only if there are $W - 1$ independent t's; that is, the presence of $W - 1$ t's represents a necessary and sufficient condition for equivalence between equations (5) and (4). However, we have not ruled out the possibility that some *other* modification of L_1 (θ) in equation (3) might also lead to a likelihood that is equivalent to L_2 (θ). For example, we could modify L_1 (θ) with *linear combinations* of the t's; this would also lead to equivalence as long as the matrix of linear combinations had dimension $W - 1$. However, it is difficult to imagine a situation in which we would choose to do this.

The t_u's may seem artificial. The point, however, is not that we deliberately choose to introduce $W - 1$ t terms into a likelihood but that when we condition on subsets, this is in fact equivalent to introducing these t's, and we need to be aware of this fact.

It can be shown that in large samples the MLEs of θ found from L_1 and L_2 will approach each other asymptotically. However, this paper is concerned with estimates in finite samples.

An interesting special case occurs when there is no ascertainment aspect to the model, i.e., when there is no distortion due to ascertainment. For example, consider the situation in which *all* observations in each subset are ascertained, i.e., where $p_{u0} = 0$ and therefore $p_u = 1$ for all u. The likelihood over the whole data set simplifies considerably, since the denominator becomes unity. That is, L_1 (θ) in equation (3) becomes

$$\begin{aligned} L(\theta) &= \prod_u \prod_{i=1} [Q_u(\theta) p_{ui}(\theta)]^{n_{ui}} \\ &= \prod_u \beta_u(\theta)^{n_u} \cdot \prod_u \prod_i p_{ui}(\theta)^{n_{ui}} . \end{aligned} \tag{18}$$

The first term on the right in equation (18) is the multinomial term $f(\theta)$, and the second term equals the likelihood conditioned on subsets, L_2 (θ). (See equations [6] and [7].) In equation (18), the β_u (θ) terms *equal* the $Q_u(\theta)$'s. However, despite this simplification, the general results still hold; that is, the two likelihoods L_1 and L_2 are not in general equivalent. Again, $W - 1$ t terms must be introduced into each β_u in order to ensure equivalence (see equation [10]):

$$\beta_u(\theta,t) = \frac{Q_u(\theta) t_u}{\sum_v Q_v(\theta) t_v} .$$

The other special case worth mentioning occurs when the subset probabilities Q_u do not depend on θ. This condition *alone* does not affect the equivalence result. In fact, as we have already seen, the FSD example discussed above represents an example in which the subset probabilities are independent of θ. However, when the Q_u's are constant with respect to θ *and* there is no ascertainment aspect to the model (as in the above paragraph), then the two likelihoods, L_1 (θ) and L_2 (θ), are always equivalent. This is because $f(\theta)$ in equation (7) is now not a function of θ at all. Therefore, from equation (8), L_1 (θ) and L_2 (θ) differ only by a constant (with respect to θ), and so they will yield the same MLE of θ.

Ewens and Shute (1986), Greenberg (1986), and others have shown that when we assume an incorrect mode of ascertainment, we may introduce substantial biases (also asymptotically) into our estimates of genetic parameters. In contrast, Ewens and Shute's nonparametric method of conditioning on that part of the data relevant to ascertainment yields asymptotically unbiased estimators. On the other hand, of course, when the classical ascertainment model *is* correct, it is preferable, because it yields smaller standard errors of the estimates than the Ewens and Shute method does. Note, too, that the Ewens and Shute method is more appropriate (i.e., results in less loss of information) in some situations than in others; see Ewens and Shute (1986) and Shute and Ewens (1988) for a more detailed discussion.

In situations in which we do not know the ascertainment model but in which the Ewens and Shute method yields unacceptably large standard errors, it may be worthwhile to develop methods of ascertainment correction that fall between the classical one and the nonparametric approach of Ewens and Shute. Such methods would require using the more cumbersome likelihood over the whole data set, L_1, but would permit us to incorporate some constraints into the likelihood. Further research will be needed to determine how feasible this approach may be.

In summary, then, if we condition on W subsets of the data, we need to know that this is equivalent to having $W - 1$ nuisance parameters or additional df in the model. Whether this fact is desirable depends on other factors. On the one hand, if those df correspond to information that we really do not have, then there is no disadvantage to conditioning on subsets. On the other hand, if they correspond to information that we do in fact possess or to assumptions that we would be willing to make, then we need to evaluate for each particular situation how severe the subsequent loss of information is (as measured by increased standard errors). This is what Ewens et al. (1986) did for the FSD problem. If the loss is too severe, then we may wish to consider using the more cumbersome but more informative likelihood over the whole data set, L_1.

Acknowledgments

This work was supported in part by National Institutes of Health (NIH) grant DK-31813 and NIH Research Career Development Award DK-01145. I wish to thank Drs. Warren Ewens, David Greenberg, Ken Lange, and Anne Spence for critical reading and helpful discussions. Gaylen Ducker typed the manuscript with her customary speed, accuracy, good humor, and patience.

Appendix A

Summary of Notation and Relationships

$Q_u = P(\text{subset } u),\ u = 1, \ldots, W$

$$\sum_{u=1}^{W} Q_u = 1$$

$p_{ui} = P(\text{category } i \mid \text{subset } u),\ i = 0, 1 \ldots, I(u)$
$p_{u0} = P(\text{not ascertained} \mid \text{subset } u)$

$$\sum_{i=0}^{I(u)} p_{ui} = 1$$

$$p_u = \sum_{i=1}^{I(u)} p_{ui} = 1 - p_{u0} = P(\text{ascertained} \mid \text{subset } u)$$

n_{ui} = number of observations in subset u, category i

$$n_u = \sum_{i=1}^{I(u)} n_{ui} = \text{number of observations in subset } u$$

$n = \sum n_u$ = total number of observations in data set
θ = parameter(s) of the model

Appendix B

Proof That When $W - 1$ Unconstrained Terms t_u Are Introduced into L_1, as in equations (5) and (10), the Two Likelihoods L_1 and L_2 Are Equivalent

Introduce a parameterization from (θ,t) to (θ,β), based on the relationship in equation (10). Thus, $f(\theta,t)$ becomes $f(\beta)$—see equation (7)—so equation (12) can be written as

$$\log L_1\ (\theta,\beta) = \log L_2\ (\theta) + \log f(\beta)\ . \tag{B1}$$

The parameter θ no longer appears in f; therefore, L_1 and L_2 in (B1) must yield the same MLE of θ. QED.

Note, however, that this technique works only if the reparameterization (transformation) is well defined and one to one, i.e., if the Jacobian of the transformation is nonzero. These conditions are met if and only if there are as many independent t's as there are β's, i.e., $W - 1$ independent t's.

To see that the existence of $W - 1$ t's corresponds to the transformation being welll defined and one to one, note from equation (10) that the t's must satisfy the following relationship:

$$\frac{t_u}{\sum_v t_v} = \frac{\beta_u/(Q_u p_u)}{\sum_v [\beta_v/(Q_v p_v)]}\ ;\ u = 1, \ldots, W\ . \tag{B2}$$

As written, equation (B2) does not have a unique solution. This is because there are W t's, but equation (B2) actually represents only $W - 1$ equations. (Subtracting the sum of the first $W - 1$ equations in equation [B2] from unity yields the Wth equation.) However, putting one additional linear constraint on the t's does provide a unique solution. For example, that constraint could be $\Sigma t_v = 1$; then equation (B2) would yield

$$t_u = \frac{\beta_u/(Q_u p_u)}{\sum_v [\beta_v/(Q_v p_v)]}\ .$$

For another example, the constraint could be $t_1 = 1$; then equation (B2) would yield

$$t_u = \frac{\beta_u/(Q_u p_u)}{\beta_1/(Q_1 p_1)}\ .$$

Clearly, however, putting more than one constraint on the t's would overdetermine them.

Appendix C

Simple Numerical Example

The Model

This example involves a coin and two urns. The coin lands heads with probability θ and tails with probability $1 - \theta$: The urns contain red and green balls in different proportions: θ red and $1 - \theta$ green in the first urn, 2θ red and $1 - 2\theta$ green in the second urn, where $0 < \theta < .5$. All sampling is with replacement. Define a trial as follows: Toss the coin. If it comes up heads, draw two balls from the first urn; otherwise, draw two balls from the second urn. Each set of two balls becomes part of our sample only if it contains at least one red ball; this aspect of the example represents ascertainment. We do not know how many sets were discarded because of failing to be ascertained. However, we do know which urn each set of two balls was drawn from. The random variable we observe is the number of red and green balls in each ascertained set of two balls. The parameter we wish to estimate is θ.

Notation

The model contains two subsets ($W = 2$), corresponding to whether the coin comes up heads or tails and hence to whether the two balls are drawn from the first or second urn. Thus, $Q_1(\theta) = \theta$ and $Q_2(\theta) = 1 - \theta$. Any given set of two balls is unascertained ($i = 0$) if it contains no red balls; otherwise, it is in category 1 or 2 if it contains one or two red balls, respectively. Thus, the p_{ui} represent the probabilities that a set of two balls will contain i red balls, conditioned on which urn the set was drawn from. For $u = 1$, $p_{1i} = \binom{2}{i}\theta^i(1 - \theta)^{2-i}$, whereas for $u = 2$, $p_{2i} = \binom{2}{i}(2\theta)^i(1 - 2\theta)^{2-i}$. The probability p_u that a set of two balls will be ascertained, given it is in subset u, is $1 - (1 - \theta)^2$ for $u = 1$ and $1 - (1 - 2\theta)^2$ for $u = 2$. The likelihood over the whole data set (equation [3]) is

$$L_1(\theta) = \frac{\theta^{n1}(1 - \theta)^{n2}}{\{\theta[1 - (1 - \theta)^2] + (1 - \theta)[1 - (1 - 2\theta)^2]\}^n} \cdot A\ , \tag{C1}$$

whereas the likelihood conditioned on subsets (equation [4]), is

$$L_2(\theta) = \frac{A}{[1 - (1 - \theta)^2]^{n1}\ [1 - (1 - 2\theta)^2]^{n2}}\ , \tag{C2}$$

where $A = [2\theta(1 - \theta)]^{n11}\ [\theta^2]^{n12}\ [2(2\theta)(1 - 2\theta)]^{n21}\ [4\theta^2]^{n22}$ in both equation (C1) and equation (C2).

Numerical Example

We have 101 ascertained sets of two balls, 20 from the first urn and 81 from the second. Among the 20 sets from the first urn, 15 contain one red ball and five contain two red balls; and among the 81 sets from the second urn there are 27 sets with one red ball and 54 with two. Thus, $n_{11} = 15$, $n_{12} = 5$, $n_{21} = 27$, $n_{22} = 54$; $n_1 = 20$, $n_2 = 81$; and $n = 101$. These are the numbers that go into equations (C1) and (C2).

Illustration of Equivalence Result

Let $\hat{\theta}_2$, $\hat{\theta}_f$, and $\hat{\theta}_1$ denote the MLEs of θ found by maximizing, respectively, $L_2(\theta)$ alone, $f(\theta)$ alone, and $L_1(\theta)$. By maximizing $L_2(\theta)$ in equation (C2) using these numerical data, we obtain $\theta_2 = .4$ exactly: Thus, the two β terms, as in equation (7), become

$$\beta_1(\theta_2) = \frac{\theta[1 - (1 - \theta)^2]}{\theta[1 - (1 - \theta)^2] + (1 - \theta)[1 - (1 - 2\theta)^2]} = .308\ ; \tag{C3}$$

$$\beta_2(\theta_2) = \frac{(1 - \theta)[1 - (1 - 2\theta)^2]}{\theta[1-(1-\theta)^2]+(1-\theta)[1-(1-2\theta)^2]} = 1-\beta_1(\theta) = .692.$$

However, we can show that this solution in equation (C3) does not maximize $L_1(\theta)$, by the following reasoning. Recall from equation (9) that if equation (C3), which maximizes L_2, is to maximize L_1 as well, it must also maximize $f(\theta)$. By equation (7), $f(\theta)$ is a multinomial likelihood. Since there are no other constraints on θ, this multinomial likelihood is maximized when $\beta_u(\theta) = n_u/n$, for $u = 1,2$; that is, $\beta_1(\hat{\theta}_f)$ must equal $20/101 = .198$, not .308 as in equation (C3); similarly, $\beta_2(\hat{\theta}_f)$ must equal .802, not .692. Therefore, for these data, as in general, the two likelihoods (C1) and (C2) will not yield identical estimates of θ. In fact, maximizing $L_1(\theta)$ in (C1) yields $\hat{\theta}_1$ of approximately .381, as opposed to .40.

Observe how introducing the t's changes the situation. We insert t_1 and t_2 into the β's, so that equation (C4) is replaced by $\beta_1(\theta,t) = \theta[1 - (1 - \theta)^2]t_1/\{\theta[1 - (1 - \theta)^2]t_1 + (1 - \theta)[1 - (1 - 2\theta)^2]t_2\}$, and a similar modification of $\beta_2(\theta)$ results in $\beta_2(\theta,t)$. The likelihood over the whole data set (equation [C1]) becomes

$$L_1(\theta,t) = [\beta_1(\theta,t)]^{n1}\ [\beta_2(\theta,t)]^{n2} \cdot L_2(\theta)\ , \tag{C4}$$

with β_1 (θ,t) and $\beta_2(\theta,t)$ as in the previous sentence and $L_2(\theta)$ as in equation (C2). Maximizing equation (C4) with respect to θ and t now yields θ = .40; the MLEs of the t's depend on which constraint we choose. For example, if we set t_1 = 1, then $\hat{t}_2$ = 1.8; if we constrain $t_1 + t_2 = 1$, then $\hat{t}_1$ = .3571 and $\hat{t}_2$ = .6429. The reader can confirm that these MLEs of θ and t do yield $\beta_1 = n_1/n$ = 20/101 and $\beta_2 = n_2/n$ = 81/101, as required if L_1 and L_2 are to be equivalent.

In summary, then, for the numerical data as given, the MLE of θ is approximately .38 when the likelihood over the whole data set (equation [Cl]) is used but equals .40 when the likelihood conditioned on subsets (equation [C2]) is used. Both estimators are asymptotically unbiased, but the former has the smaller asymptotic variance (Ewens et al. 1986). On the other hand, if we change the model by introducing the t_u terms, as in equation (C4), then both estimates of θ equal .40.

References

Ewens, W. J., and B. Asaba. 1984. Estimating parameters of the family-size distribution in ascertainment sampling schemes: numerical results. Biometrics **40**:367–374.

Ewens, W. J., S. E. Hodge, and H. P. Foo. 1986. The effects of a known family-size distribution on the estimation of genetic parameters. Am. J. Hum. Genet. **38**:555–566.

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

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

Hodge, S. E. 1985. Family-size distribution and Ewens' equivalence theorem. Am. J. Hum. Genet. **37**:166–177.

Morton, N. E. 1959. Genetic tests under incomplete ascertainment. Am. J. Hum. Genet. **11**:1–16.

Shute, N. C. E., and W. J. Ewens. 1988. Resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am. J. Hum. Genet. **43**:374–386.

Weinberg, W. 1928. Mathematische Grundlagen der Probandenmethode. Z. Induktive Abstammungs- und Vererbungslehre **48**:179–228.