Statistica Sinica **19** (2009), 409-425

TESTING FOR FAMILIAL AGGREGATION WHEN THE POPULATION SIZE IS KNOWN

Yixin Fang and Daniel Rabinowitz

Georgia State University and Columbia University

Abstract: We treat the problem of testing for familial aggregation when sampling from a population of a known size. And consider the setting where data from all families in the population containing at least one affected member are obtained. The results are compared to the setting where the population size is unknown and data from a sample of families containing at least one affected member are obtained, and to the setting where the population size is known and data from all families in the population are obtained. Two kinds of local alternatives are considered: one in which a predisposing factor is prevalent but has small penetrance; the other in which the factor is penetrant but has small prevalence. It is found that knowing the population size provides substantial advantage over settings where population size is unknown, but that there is little advantage to settings where data from all families are obtained. The methods are illustrated through an application to data from a child survival study in northeast Brazil reported by Sastry (1997), and reanalyzed by Yu and Zelterman (2002).

Key words and phrases: Asymptotic relative efficiency, locally most powerful unbiased test, random-effect model, sampling design.

1. Introduction

Testing for familial aggregation of common diseases has been the focus of many methodological and statistical efforts. Sometimes, rates of disease in relatives of diseased individuals (case probands) and in relatives of non-diseased individuals (control probands) are compared; see, for example, Khoury, Beaty and Cohen (1993). Other methods take into account the information between relatives, by using a random-effect model with latent family-specific factors (Liang (1987), Commenges, Letenneur, Jacqmin, Moreau and Dartigues (1994), Commenges, Jacqmin, Letenneur and Van Duijn (1995), among others). Analogous methods have not been developed, however, for the setting where family data are obtained from a population of a known size. In cases where comprehensive medical records are available, information on all individuals who might present with a disease is potentially available. Motivated by an ongoing study of epilepsy (Annegers, Hauser, Anderson and Kurland (1982)) we derive methods for these cases in this paper.

The study of epilepsy was carried out in Rochester, Minnesota. The medical records of patients with epilepsy at the Mayo Clinic, the Olmsted Medical Group, the Olmsted County Public Health Department, and the other medical facilities that served the population of southeastern Minnesota were reviewed. The parents of of the patients were identified and, through them, other descendants of the parents of the patients. Patients were children living in Rochester at the time of diagnosis, so everyone who might be diagnosed had a medical record there. Furthermore, in most cases, siblings and parents were also residents of Rochester. In fact, most of the other descendants were also born in Rochester, although there were some cases of moving in or out. In addiction, other information on residency was obtained from city directories and county plat books, as well as medical records.

For study design in such settings, it would be helpful to understand the degree to which obtaining information on the population size and on the composition of families without affected members would increase efficiency.

To investigate this, we compare three designs: one in which all families in a population of a known size are obtained (Design I); one in which all affected families in a population of a known size are obtained (Design II); and one in which a sample of affected families is obtained from a population with an unknown size (Design III).

To model the null hypothesis that diseases occur in individuals uniformly, versus the alternative that there is a family-specific risk factor, a random-effect model with a family-specific latent factor is used; conditionally, given the family-specific latent factor, presence of disease in family members is independent and identically distributed. Under the null hypothesis, the latent factor is a constant and the disease statuses of individuals are independent Bernoulli variables with the same expectation (the population disease rate). We consider settings where reliable information about the population disease rate in the cohort under study is not available externally.

The appropriate model for familial aggregation under the alternative depends on the genetic architecture of the trait under consideration; of course genetic factors are not the only ones that induce familial aggregation. In this paper, we consider two broad classes of local alternatives: one in which the effect of a latent family-specific factor with substantial prevalence tends to zero, and the other in which the prevalence of the latent factor with a substantial effect tends to zero.

Since prior information on the population disease rate is absent, the null hypothesis is composite and no uniformly most powerful tests, even locally most powerful tests, exist. Therefore, we seek locally most powerful unbiased tests. The usual approach of conditioning on the complete sufficient statistics for the

null hypothesis can be applied to Designs I and III, but for Design II, there is no complete sufficient statistic under the null hypothesis. However, in this case there exist asymptotically unbiased tests of almost the same local power as that of the locally most powerful unbiased tests applied to Design I.

In the next section, notation and the models are presented. In the third section, complete sufficient statistics for the null hypothesis in the different designs are presented. In the fourth section, locally most powerful unbiased tests in the setting of substantial prevalence are developed. In the fifth section, locally most powerful tests in the setting of substantial penetrance are developed. In the sixth section, the results are examined in the context of a child survival study in northeast Brazil reported by Sastry (1997), and reanalyzed by Yu and Zelterman (2002).

2. Notation and Models

Let N denote the number of individuals in a population. Suppose that the population is partitioned into families, and let I denote the number of families. Here by a family we mean a group of people related by blood or marriage (see, for example, Pfeiffer, Gail and Pee (2001)). The concept of families is only approximate, and it is a simplifying assumption that families are independent and family-specific latent factors are identically distributed. Let i index families and j index subjects. Let Y_{ij} be the indicator that the jth subject in the ith family is affected. Let d_i denote the number of affected members and n_i the total number of members in the ith family.

Under the null hypothesis of no familial aggregation, the Y_{ij} are independent and identically distributed Bernoulli variables. Here, we consider two broad classes of local alternatives: one corresponding to a common variant with a small effect on the trait; the other corresponding to a rare variant with a substantial impact on the trait.

Local alternative in which the predisposing factor is prevalent but has small penetrance may be modeled using a latent variable. Let A_i denote independent and identically distributed family-specific latent variables. Let F denote the distribution function of the A_i. Suppose that the conditional probability given A_i, that an individual from the ith family is diseased, is

$$P(Y_{ij} = 1|A_i) = \frac{e^{\alpha+\theta A_i}}{1 + e^{\alpha+\theta A_i}}. \tag{2.1}$$

For convenience, let p_0 denote $e^{\alpha}/(1 + e^{\alpha})$.

Local alternative in which the factor with a substantial effect is rare may be modeled with a latent variable as

$$P(Y_{ij} = 1|A_i = 0) = p_0 \text{ and } P(Y_{ij} = 1|A_i = 1) = p_1, \tag{2.2}$$

where $p_0 < p_1$, $P(A_i = 1) = \theta$ and $P(A_i = 0) = 1 - \theta$. Actually, this model is a special case of (2.1) where the latent variables A_i are Bernoulli.

Let I^a denote the number of families containing at least one affected member in the sample. Let m_s denote the number of families of size s in the population, and m_s^a denote the number of affected families of size s in the sample. Let S denote the largest family size. Let $\mathbf{M}$ and $\mathbf{M}^a$, respectively, denote the sequences $\{m_1, \ldots, m_S\}$ in the population and $\{m_1^a, \ldots, m_S^a\}$ in the sample. Let D denote the total number of affected members in the sample. In Designs I and II, D is also the total number of affected members in the population.

3. Complete Sufficient Statistics

To obtain unbiased tests, it is useful to consider complete sufficient statistics under the null hypothesis. See, for example, Cox and Hinkley (1974, p.146).

In Design I, under the null hypothesis the likelihood is

$$p_0^D (1 - p_0)^{N-D} \prod_{i=1}^{I} \binom{n_i}{d_i}.$$

Since the likelihood belongs to the exponential family of distributions, it is easy to see that the complete sufficient statistic for the null hypothesis is D. In addition, under the null hypothesis, the maximum likelihood estimate (MLE) $\hat{p}$ for p_0 is D/N.

In Design II, under the null hypothesis the likelihood is

$$\prod_{s=1}^{S} \binom{m_s}{m_s^a} [1 - (1 - p_0)^s]^{m_s^a} [(1 - p_0)^s]^{m_s - m_s^a} \prod_{i=1}^{I^a} \frac{\binom{n_i}{d_i}}{1 - (1 - p_0)^{n_i}},$$

where the first product is the probability of sampling families and the second product is the probability of sampling subjects among affected families. The above formula can be expressed as

$$p_0^D (1 - p_0)^{N-D} \prod_{s=1}^{S} \binom{m_s}{m_s^a} \prod_{i=1}^{I^a} \binom{n_i}{d_i},$$

where $\sum s m_s = N$. By the factorization theorem, $\{D, \mathbf{M}^a\}$ is the minimal sufficient statistic for the nuisance parameter $\{p_0, \mathbf{M}\}$. However, it is not complete. See Appendix for a sketch of the proof. Therefore, there is no complete sufficient statistic for the null hypothesis. But under the null hypothesis, the MLE $\hat{p}$ for p_0 is also D/N.

In Design III, under the null hypothesis the likelihood is

$$\prod_{i=1}^{I^a} \binom{n_i}{d_i} p_0^{d_i} \frac{(1-p_0)^{n_i-d_i}}{1-(1-p_0)^{n_i}}.$$

Since the above formula can be expressed as

$$p_0^D (1-p_0)^{\sum sm_s^a - D} \frac{\prod_{i=1}^{I^a} \binom{n_i}{d_i}}{\prod [1-(1-p_0)^s]^{m_s^a}},$$

by the factorization theorem, $\{D, \mathbf{M}^a\}$ is sufficient statistic for the nuisance parameter $\{p_0, \mathbf{M}\}$. We can also show that it is also complete. Therefore, $\{D, \mathbf{M}^a\}$ is the complete sufficient statistic for the null hypothesis. Let p^* denote the MLE for p_0 under the null hypothesis, the solution to $\sum_{i=1}^{I^a} n_i p^*/[1-(1-p^*)^{n_i}] = D$.

4. Tests Against a Small Effect of Common Variants

In this section, we describe the locally most powerful unbiased tests against the local alternative in which the predisposing factor is prevalent but has small penetrance for the different designs.

4.1. The locally most powerful unbiased tests

In Design I, the likelihood is

$$\prod_{i=1}^{I} \int \binom{n_i}{d_i} p_\theta(a)^{d_i} (1-p_\theta(a))^{n_i-d_i} dF(a),$$

where $p_\theta(a) = e^{\alpha+\theta a}/(1+e^{\alpha+\theta a})$. Here θ is the parameter of interest and $\{\alpha, F\}$ is the nuisance parameter. Given the complete sufficient statistic D, the conditional likelihood is

$$\frac{\prod_{i=1}^{I} \int \binom{n_i}{d_i} p_\theta(a)^{d_i} (1-p_\theta(a))^{n_i-d_i} dF(a)}{\sum^* \prod_{i=1}^{I} \int \binom{n_i}{\tilde{d}_i} p_\theta(a)^{\tilde{d}_i} (1-p_\theta(a))^{n_i-\tilde{d}_i} dF(a)},$$

where $\sum^*$ is over all nonnegative integers $\tilde{d}_1, \ldots, \tilde{d}_I$ such that $\sum_{i=1}^{I} \tilde{d}_i = D$.

Usually, the derivative of the conditional log-likelihood evaluated at $\theta = 0$ would be a locally most powerful unbiased test statistic. However, in this case, the derivative at $\theta = 0$ is

$$\frac{E(\frac{\partial p_\theta(A_1)}{\partial \theta}\Big|_{\theta=0})}{p_0(1-p_0)} \Big[\sum_{i=1}^{I} (d_i - n_i p_0) - E_0 \Big\{ \sum_{i=1}^{I} (d_i - n_i p_0) | D \Big\} \Big],$$

identically zero. Liang (1987), Commenges et al. (1994) and Commenges et al. (1995) dealt with this issue by a transformation of the parameter: $\text{logit}P(Y_{ij} = 1|A_i) = \alpha + \sqrt{\theta}A_i$. In this case, the second order derivative of the conditional log-likelihood evaluated at $\theta = 0$ is the locally most powerful unbiased test statistic, and is proportional to

$$\sum_{i=1}^{I}(d_i - n_i p_0)^2 - E_0\Big\{\sum_{i=1}^{I}(d_i - n_i p_0)^2|D\Big\}. \tag{4.1}$$

The nuisance parameter p_0 in the above formula can be replaced by its MLE $\hat{p}$; this does not influence the power of the test statistic asymptotically. The exact p-value can be obtained through a permutation procedure where D is fixed and the disease statuses are permuted among subjects. An approximation based on the z-score is described in the next section.

Similarly, for Design III , because the complete sufficient statistic for the null hypothesis is $\{D, \mathbf{M}^a\}$, the locally most powerful unbiased test statistic is

$$\sum_{i=1}^{I^a}(d_i - n_i p_0)^2 - E_0\Big\{\sum_{i=1}^{I^a}(d_i - n_i p_0)^2|D, \mathbf{M}^a\Big\}. \tag{4.2}$$

The nuisance parameter p_0 can be replaced by its MLE p^*; this does not influence the power of the test statistic asymptotically. The exact p-value can be obtained through a permutation procedure. An algorithm is described briefly in the next section, along with an approximation based on the z-score.

For Design II, since there is no complete sufficient statistic, we cannot follow the above steps to get the locally most powerful unbiased test. However, it might be important to note that for Design I, the locally most powerful unbiased test statistic is the projection of $\sum_{i=1}^{I}(d_i - n_i p_0)^2$ on the orthocomplement of the space spanned by $\{D\}$, while for Design III, the locally most powerful unbiased test statistic is the projection on the orthocomplement of the space spanned by $\{D, \mathbf{M}^a\}$. Therefore, it is natural to consider the projection of $\sum_{i=1}^{I}(d_i - n_i p_0)^2$ on the orthocomplement of the space spanned by $\{\mathbf{M}^a\}$. This leads to the test statistic

$$\sum_{i=1}^{I^a}(d_i - n_i \hat{p})^2 - E_0\Big\{\sum_{i=1}^{I^a}(d_i - n_i p_0)^2|\mathbf{M}^a\Big\}\Big|_{p_0=\hat{p}}. \tag{4.3}$$

The p-value can be calculated based on the z-score, the detail is described in the next section. First of all, the above test statistic is asymptotically unbiased because it does not depend on the composition of those unaffected families which

are unavailable. Moreover, in Subsection 4.3, it is shown that the asymptotic efficiency of this test is almost the same as that of (4.1) under Design I.

4.2. Computation of test statistics

For Design I, the test statistic (4.1) is denoted by S_D, with the subscript standing for "projection on $\{D\}$", and is

$$S_D = \sum_{i=1}^{I} \Big\{ d_i^2 - 2n_i\hat{p}d_i - \Big[n_i \frac{D}{N} - n_i \frac{D(D-1)}{N(N-1)} + n_i^2 \frac{D(D-1)}{N(N-1)} - 2n_i^2 \hat{p} \frac{D}{N} \Big] \Big\}.$$

If we replace $(D-1)/(N-1)$ by D/N, this becomes

$$S_D = \sum_{i=1}^{I} \Big[(d_i - n_i\hat{p})^2 - n_i\hat{p}(1-\hat{p}) \Big]. \tag{4.4}$$

Obviously, the difference between these two formulae is ignorable. The variance of S_D is described in the Appendix.

For Design II, the test statistic (4.3) is written as

$$S_{\mathbf{M}^a} = \sum_{i=1}^{I^a} \Big[d_i^2 - 2n_i d_i \hat{p} - \frac{n_i\hat{p}(1-\hat{p}) - n_i^2\hat{p}^2}{1-(1-\hat{p})^{n_i}} \Big]. \tag{4.5}$$

The variance of $S_{\mathbf{M}^a}$ is described in the Appendix.

For Design III, denote the test statistic (4.2) as $S_{D,\mathbf{M}^a}$. However, we do not have an explicit formula for $S_{D,\mathbf{M}^a}$ and its conditional variance. Here we outline an algorithm (like the one in Yu and Zelterman (2002)) that facilitates the computation of $S_{D,\mathbf{M}^a}$ and its conditional variance. Let X_{sj} denote the number of families of size s having j $(s = 1, \ldots, S; j = 1, \ldots, s)$ diseased members. These numbers are constrained by

$$\sum_{s=1}^{S} \sum_{j=1}^{s} jX_{sj} = D \text{ and } \sum_{j=1}^{s} X_{sj} = m_s^d, \text{ for } s = 1, \ldots, S.$$

First we generate all possible sets of nonnegative integers x_{sj} that satisfy the conditions. Then, for each set of integers, we can calculate the corresponding probability

$$P(\{X_{sj}\} = \{x_{sj}\} | D, \mathbf{M}^a) = \frac{\prod_{s=1}^{S} \prod_{j=1}^{s} \binom{s}{j}^{x_{sj}} / x_{sj}!}{\sum^* \prod_{s=1}^{S} \prod_{j=1}^{s} \binom{s}{j}^{\tilde{x}_{sj}} / \tilde{x}_{sj}!},$$

where $\sum^*$ is summation over all possible sets $\{\tilde{x}_{sj}\}$ that satisfy the conditions. Based on these probabilities, computation of $S_{D,\mathbf{M}^a}$ and its conditional variance is straightforward.

However, the above algorithm is time-consuming if we consider a population of a moderate or large size. To overcome this difficulty, it is reasonable to propose an approximation for $S_{D,\mathbf{M}^a}$. As a matter of fact, the hypergeometric distribution can be approximated very accurately by a binomial distribution (Feller (1950), Bennett and Franklin (1954), among others). Therefore, the distribution of d_i given $\mathbf{M}^a$ and D can be approximated very well by the conditional distribution of $\tilde{d}_i$ given $\tilde{d}_i > 0$, where $\tilde{d}_i$ is a binomial variable with parameters n_i and p^*. Since

$$E(\tilde{d}_i|\tilde{d}_i > 0) = \frac{n_i p^*}{1-(1-p^*)^{n_i}} \text{ and } E(\tilde{d}_i^2|\tilde{d}_i > 0) = \frac{n_i p^*(1-p^*) + n_i^2 p^{*2}}{1-(1-p^*)^{n_i}},$$

an approximation for $S_{D,\mathbf{M}^a}$ is

$$\hat{S}_{D,\mathbf{M}^a} = \sum_{i=1}^{I^a} \Big[d_i^2 - 2n_i d_i p^* - \frac{n_i p^*(1-p^*) - n_i^2 p^{*2}}{1-(1-p^*)^{n_i}} \Big]. \tag{4.6}$$

The variance of $\hat{S}_{D,\mathbf{M}^a}$ is described in the Appendix.

4.3. Asymptotic relative efficiency

In this subsection, in order to compare Design II with the others, we focus on a special case of Design III, one in which all affected families are obtained from a population with an unknown size.

Recall that the derivative of the conditional log-likelihood under the null hypothesis for any of the three designs is zero. This implies (see the Appendix)

$$\frac{\partial E_\theta(S_D)}{\partial\theta}\Big|_{\theta=0} = \frac{\partial E_\theta(S_{\mathbf{M}^a})}{\partial\theta}\Big|_{\theta=0} = \frac{\partial E_\theta(S_{D,\mathbf{M}^a})}{\partial\theta}\Big|_{\theta=0} = 0.$$

Therefore, we cannot use Pitman asymptotic efficiency (PAE, see Zacks (1985)) as a criterion. In order to compare relative efficiencies, for any test statistic T such that $E_0(T) = 0$ and $\frac{\partial E_\theta(T)}{\partial\theta}|_{\theta=0} = 0$, we define

$$\frac{(\partial^2 E_\theta(T)/\partial\theta^2|_{\theta=0})^2}{N\mathrm{Var}_0(T)}$$

as the asymptotic efficiency of T, where the variance is taken at $\theta = 0$. The rationale is that the locally most powerful test is the one maximizing $E_\theta(T)/\sqrt{\mathrm{Var}_0(T)}$, as $\theta \to 0$, among all tests T such that $E_0(T) = 0$.

First, by arguments in the Appendix, the asymptotic efficiencies of S_D and $S_{D,\mathbf{M}^a}$ are, respectively,

$$AE(S_D) = \frac{\mathrm{Var}_0(S_D)}{N} \text{ and } AE(S_{D,\mathbf{M}^a}) = \frac{\mathrm{Var}_0(S_{D,\mathbf{M}^a})}{N}.$$

The Pythagorean decomposition implies that $\mathrm{Var}_0(E_0\{\sum(d_i - n_i p_0)^2|D, \mathbf{M}^a\} - E_0\{\sum(d_i - n_i p_0)^2|D\})/N$ measures the loss of efficiency when those unaffected families and the population size are unavailable.

Furthermore, through some tedious calculations, it is shown that if the disease rate under the consideration is rare, $AE(\hat{S}_{D,\mathbf{M}^a})/AE(S_D) = o(1)$, where $o(\cdot)$ is in the sense that p_0 is tending to zero. In addition, if p_0 is small,

$$AE(S_{\mathbf{M}^a}) \approx AE(S_D) = \frac{2}{N}\sum_{i=1}^{I} n_i(n_i - 1)[p_0^2 + o(p_0^2)].$$

In particular, in the case where $n_i \equiv n$,

$$AE(S_{\mathbf{M}^a}) \approx AE(S_D) = 2(n-1)(p_0^2 + o(p_0^2)).$$

To sum up, if the population size is known, obtaining information on the composition of those unaffected families provides little advantage for the test for familial aggregation. Still, knowing the population size provides substantial advantage over the settings where population size is unknown.

This section concludes with a numeric example. In Table 1, assuming $n_i \equiv n$, asymptotic relative efficiencies (ARE) $ARE(S_{\mathbf{M}^a}, S_D)$ and $ARE(\hat{S}_{D,\mathbf{M}^a}, S_D)$ are presented for different settings. From this example, it seems that to test against small effect of family factor, the local power of $S_{\mathbf{M}^a}$ is almost the same as that of S_D in all cases. This means that if the population size is known, then having unaffected families is not important. But, compared to S_D, the local power of $\hat{S}_{D,\mathbf{M}^a}$ is very small. Particularly, for a rare disease, the ratio of local power of two tests is extremely small. Although it increases with the population disease rate, the local power of $\hat{S}_{D,\mathbf{M}^a}$ is still small for large population disease rate in comparison to S_D. This means that knowing the population size is very important.

5. Tests Against a Rare Variant with a Substantial Effect

In this section, we describe the locally most powerful unbiased tests against a rare variant with a substantial effect, under model (2.2). To be concordant with model (2.1), take $\beta = \log[p_1/(1-p_1)]/[p_0/(1-p_0)]$ to measure the effect of the rare variant. In the first three subsections, we temporarily assume that β is known. In Subsection 5.4, we provide an estimate of β and analyze the sensitivity of the results to the choice of β.

Table 1. $ARE(S_{\mathbf{M}^a}, S_D)$ and $ARE(\hat{S}_{D,\mathbf{M}^a}, S_D)$.

		$p_0 = 0.001$	$p_0 = 0.01$	$p_0 = 0.05$	$p_0 = 0.1$
$ARE(S_{\mathbf{M}^a}, S_D)$	$n = 2$	1.0000	1.0000	1.0000	1.0000
	$n = 3$	1.0000	0.9996	0.9986	0.9980
	$n = 4$	0.9999	0.9993	0.9975	0.9972
	$n = 5$	0.9999	0.9990	0.9967	0.9973
	$n = 6$	0.9999	0.9987	0.9961	0.9980
$ARE(\hat{S}_{D,\mathbf{M}^a}, S_D)$	$n = 2$	0.0000	0.0000	0.0000	0.0000
	$n = 3$	0.0003	0.0034	0.0172	0.0357
	$n = 4$	0.0007	0.0067	0.0343	0.0707
	$n = 5$	0.0010	0.0101	0.0513	0.1051
	$n = 6$	0.0013	0.0134	0.0680	0.1387

5.1. The locally most powerful unbiased tests

This subsection is parallel to Subsection 4.1. In Design I, the likelihood is

$$\prod_{i=1}^{I}[\theta p_1^{d_i}(1-p_1)^{n_i-d_i} + (1-\theta)p_0^{d_i}(1-p_0)^{n_i-d_i}].$$

The derivative of the conditional log-likelihood given the complete sufficient statistic D at $\theta = 0$, is

$$\sum_{i=1}^{I} \frac{p_1^{d_i}(1-p_1)^{n_i-d_i}}{p_0^{d_i}(1-p_0)^{n_i-d_i}} - E_0\Big\{ \sum_{i=1}^{I} \frac{p_1^{d_i}(1-p_1)^{n_i-d_i}}{p_0^{d_i}(1-p_0)^{n_i-d_i}} \Big| D \Big\}.$$

Unlike Subsection 4.1, where the derivative of the conditional log-likelihood given the complete sufficient statistic is zero, here it is non-zero and is the locally most powerful unbiased test statistic. The above formula can be rewritten as

$$\sum_{i=1}^{I} \frac{e^{\beta d_i}}{[1+(e^\beta-1)p_0]^{n_i}} - E_0\Big\{ \sum_{i=1}^{I} \frac{e^{\beta d_i}}{[1+(e^\beta-1)p_0]^{n_i}} \Big| D \Big\}, \tag{5.1}$$

where p_0 can be estimated by its MLE $\hat{p} = D/N$. The exact p-value can be obtained through a permutation procedure, and an approximation based the z-score is described in the next section.

Similarly, in Design III, the locally most powerful unbiased test is

$$\sum_{i=1}^{I^a} \frac{e^{\beta d_i}}{[1+(e^\beta-1)p_0]^{n_i}} - E_0\Big\{ \sum_{i=1}^{I^a} \frac{e^{\beta d_i}}{[1+(e^\beta-1)p_0]^{n_i}} \Big| D, \mathbf{M}^a \Big\}, \tag{5.2}$$

where p_0 can be estimated by its MLE p^*. Again, the exact p-value can be obtained through a permutation procedure, and an approximation based the z-score is described in the next section.

In Design II, the same arguments in Subsection 4.1 suggest the test statistic

$$\sum_{i=1}^{I^a} \frac{e^{\beta d_i}}{[1+(e^\beta-1)\hat{p}_0]^{n_i}} - E_0\Big\{\sum_{i=1}^{I^a} \frac{e^{\beta d_i}}{[1+(e^\beta-1)p_0]^{n_i}}\Big|\mathbf{M}^a\Big\}\Big|_{p_0=\hat{p}}. \tag{5.3}$$

The p-value can be calculated based on the z-score, the detail is described in the next section. First of all, the above test is asymptotically unbiased because it does not depend on the composition of those unaffected families which are unavailable. Moreover, in Subsection 5.3, we show that the asymptotic efficiency of this test is almost the same as that of (5.1) for Design I.

5.2. Computation of test statistics

For Design I, denote the test statistic (5.1) as T_D. It can be calculated exactly by using the hypergeometric distribution, but because of the intractable nature of this distribution, an approximation is necessary. Note that the conditional distribution of d_i given D can be approximated accurately by the distribution of $\tilde{d}_i$, where $\tilde{d}_i$ is a binomial random variable with parameters n_i and $\hat{p}$. Since $E(e^{\beta\tilde{d}_i}) = [1+(e^\beta-1)\hat{p}]^{n_i}$, an approximation to T_D is

$$\hat{T}_D = \sum_{i=1}^{I}\Big\{\frac{e^{\beta d_i}}{[1+(e^\beta-1)\hat{p}]^{n_i}} - 1\Big\}. \tag{5.4}$$

The variance of $\hat{T}_D$ is described in the Appendix.

For Design II, the test statistic (5.3) is

$$T_{\mathbf{M}^a} = \sum_{i=1}^{I^a} \frac{1}{[1+(e^\beta-1)\hat{p}]^{n_i}}\Big\{e^{\beta d_i} - \frac{[1+(e^\beta-1)\hat{p}]^{n_i} - (1-\hat{p})^{n_i}}{1-(1-\hat{p})^{n_i}}\Big\}. \tag{5.5}$$

The variance of $T_{\mathbf{M}^a}$ is described in the Appendix.

For Design III, denote the test statistic (5.2) as $T_{D,\mathbf{M}^a}$. The same arguments in Subsection 5.2 suggests an approximation for $T_{D,\mathbf{M}^a}$,

$$\hat{T}_{D,\mathbf{M}^a} = \sum_{i=1}^{I^a} \frac{1}{[1+(e^\beta-1)p^*]^{n_i}}\Big\{e^{\beta d_i} - \frac{[1+(e^\beta-1)p^*]^{n_i} - (1-p^*)^{n_i}}{1-(1-p^*)^{n_i}}\Big\}. \tag{5.6}$$

The variance of $\hat{T}_{D,\mathbf{M}^a}$ is described in the Appendix.

5.3. Asymptotic relative efficiency

In this subsection Pitman asymptotic efficiency (PAE, see Zacks (1985)) is used as a criterion to compare the asymptotic efficiencies of different test

statistics. For any test statistic T with $E_0(T) = 0$, Pitman asymptotic efficiency is $(\partial E_\theta(T)/\partial\theta|_{\theta=0})^2/[N\mathrm{Var}_0(T)]$.

In order to compare Design II with the two others, we focus on a special case of Design III, one in which all affected families are obtained from a population with an unknown size. First, following the steps in Subsection 4.3, we have $PAE(T_D) = \mathrm{Var}_0(T_D)/N$ and $PAE(T_{D,\mathbf{M}^a}) = \mathrm{Var}_0(T_{D,\mathbf{M}^a})/N$. Through arguments in the Appendix and a Taylor expansion, it is shown that when β is small, $PAE(\hat{T}_{D,\mathbf{M}^a})$ is much less than $PAE(\hat{T}_D)$; but if β is big, $PAE(\hat{T}_{D,\mathbf{M}^a})$ is not trivial in comparison to $PAE(\hat{T}_D)$ and it may approximate $PAE(\hat{T}_D)$ well when β is large enough. Actually, the result in the latter situation is not surpring. Imagine that β is big. In the sample obtained by Design III, there would likely appear two clusters of families: one cluster with only one affected member and the other with many affected members. This clear difference between two clusters of families makes it easy to test against the local alternative. On the other hand, although $PAE(\hat{T}_{D,\mathbf{M}^a})$ may not be trivial, we have to keep in mind that here we assume β is known. In the setting where N is unknown, it is difficult to choose an appropriate β in the construction of $\hat{T}_{D,\mathbf{M}^a}$.

Second, by arguments in the Appendix, we find that when the disease is rare and β is small, $PAE(T_{\mathbf{M}^a}) \approx PAE(\hat{T}_D) \approx (1/(2N))\sum_{i=1}^{I} n_i(n_i-1)p_0^2(e^\beta-1)^4$. If β is moderate or big, we have $PAE(T_{\mathbf{M}^a}) \approx PAE(\hat{T}_D) \approx (1/(2N))\sum_{i=1}^{I}\{[1+e^\alpha\big((e^\beta-1)^2/(1+e^{\alpha+\beta})^2\big]^{n_i}-1\}$, where $\alpha = \log[p_0/(1-p_0)]$.

We can make similar conclusions in testing against the second alternative: if the population size is known, obtaining information on the composition of those unaffected families provides little advantage, but knowing the population size provides substantial advantage over the settings where population size is unknown.

This subsection concludes with an example. In Table 2, assuming that $p_0 = 0.01$ and $n_i \equiv n$, $ARE(T_{\mathbf{M}^a}, \hat{T}_D)$ and $ARE(\hat{T}_{D,\mathbf{M}^a}, \hat{T}_D)$ are presented. The settings $p_1 = 0.02, 0.05, 0.10, 0.20$ correspond to $\beta = 0.70, 1.65, 2.39, 3.21$, respectively. From this example, it seems that to test against a rare variant with a substantial effect, the local power of $T_{\mathbf{M}^a}$ is almost the same as that of $\hat{T}_D$ in all cases. This means that if the population size is known, having unaffected families is not important. On the other hand, if the effect of the rare variant is small or moderate, compared to $\hat{T}_D$, the local power of $\hat{T}_{D,\mathbf{M}^a}$ is very small - knowing N is important. However, when the effect of the latent factor is big, the local power $\hat{T}_{D,\mathbf{M}^a}$ could be close to that of $\hat{T}_D$. This implies that in this case, those affected families contain almost all the information for the testing for familial aggregation. Here we assume again that the effect the family risk factor β is known, and knowing N is very helpful in finding an accurate estimate for β.

Table 2. $ARE(T_{\mathbf{M}^a}, \hat{T}_D)$ and $ARE(\hat{T}_{D,\mathbf{M}^a}, \hat{T}_D)$.

		$p_1 = 0.02$	$p_1 = 0.05$	$p_1 = 0.10$	$p_1 = 0.20$
$ARE(T_{\mathbf{M}^a}, \hat{T}_D)$	$n = 2$	1.0000	1.0000	1.0000	1.0000
	$n = 3$	0.9967	0.9971	0.9982	0.9995
	$n = 4$	0.9935	0.9949	0.9977	0.9998
	$n = 5$	0.9904	0.9931	0.9978	0.9999
	$n = 6$	0.9874	0.9918	0.9982	1.0000
$ARE(\hat{T}_{D,\mathbf{M}^a}, \hat{T}_D)$	$n = 2$	0.0000	0.0000	0.0000	0.0000
	$n = 3$	0.0134	0.0796	0.2637	0.6059
	$n = 4$	0.0266	0.1538	0.4695	0.8643
	$n = 5$	0.0396	0.2230	0.6261	0.9582
	$n = 6$	0.0524	0.2873	0.7422	0.9881

5.4. Sensitivity to the choice of β

In the preceding subsections, we temporarily assumed that β was known. In the Appendix, we provide moment estimates of β for different settings. In the following, we take the true value of β to be β_0, but we use β rather than β_0 in the construction of $\hat{T}_D$ and $T_{\mathbf{M}^a}$. Therefore, the expressions of $\hat{T}_D(\beta)$ and $T_{\mathbf{M}^a}(\beta)$ emphasize their dependence on β, and we analyze the sensitivity of the results to the choice of β. Here we only consider $\hat{T}_D(\beta)$, consideration of $T_{\mathbf{M}^a}(\beta)$ is almost the same.

Noting that $\partial E_\theta(\hat{T}_D(\beta))/\partial\theta|_{\theta=0}$ is

$$\sum_{i=1}^{I}\Big\{\Big[1+e^{\alpha}\frac{(e^{\beta}-1)(e^{\beta_0}-1)}{(1+e^{\alpha+\beta})(1+e^{\alpha+\beta_0})}\Big]^{n_i}-1-n_ie^{\alpha}\frac{(e^{\beta}-1)(e^{\beta_0}-1)}{(1+e^{\alpha+\beta})(1+e^{\alpha+\beta_0})}\Big\},$$

and $PAE(\hat{T}_D(\beta)) = (\partial E_\theta(\hat{T}_D(\beta))/\partial\theta|_{\theta=0})^2/N\mathrm{Var}_0(\hat{T}_D(\beta))$, if the disease is rare and β and β_0 are small, $PAE(\hat{T}_D(\beta))$ approximates $\sum n_i(n_i-1)p_0^2(e^{\beta_0}-1)^4/2N$, which is also an approximation to $PAE(\hat{T}_D(\beta_0))$. Here we use two examples to demonstrate the sensitivity to the choice of β. In Figure 1, when $p_0 = 0.01$, ratios of $PAE(\hat{T}_D(\beta))$ and $PAE(\hat{T}_D(\beta_0))$ are presented for four different settings, where the circles represent the points at β_0 and $\beta = 1-6$ correspond to $p_1 = 0.027, 0.069, 0.169, 0.355, 0.600$ and 0.803, respectively. From Figures 1 and 2, we see that ratios are close to one if β is in $\beta_0 \pm 0.5$. The results for different values of p_0 are similar, so they are not reported here.

6. An Example

A study of child survival in northeast Brazil, summarized in Table 3, was reported by Sastry (1997), and also analyzed by Yu and Zelterman (2002). This was a household survey conducted as part of a Demographic and Health Survey

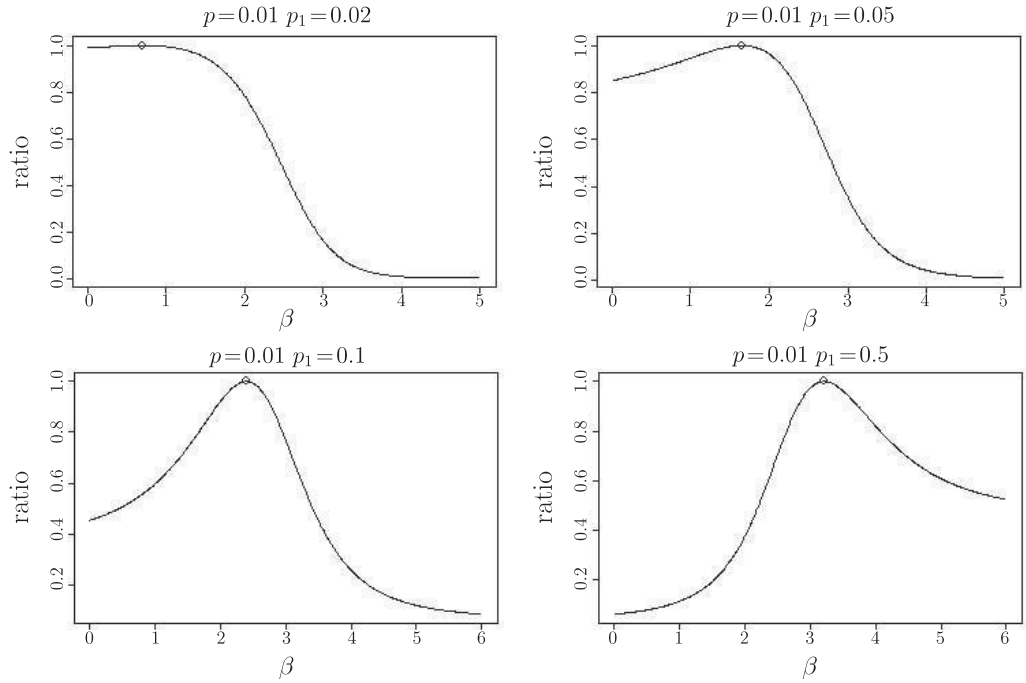

Figure 1. $PAE(\hat{T}_D(\beta))/PAE(\hat{T}_D(\beta_0))$ when $p_0 = 0.01$.

program. This was a random sampling design and the data set included those unaffected families. A reason to use this data-set as an example is that we want to illustrate and compare the six tests developed in Sections 4 and 5.

In this example, $N = 2,946$ and $D = 430$. The MLE of p_0 under the null hypothesis is $\hat{p} = D/N = 0.146$. For the design where a sample of affected families is obtained but N is unknown, the MLE of p_0 is $p^* = 0.259$. To test against a rare variant with a substantial effect under (2.2), we need to estimate the log-odds ratio β in the construction of the tests. For the random sampling setting in which unaffected families are also available, following steps in the Appendix, moment estimates of p_0 and p_1 are 0.071 and 0.591, respectively. This implies that an estimate for β is 2.937. For the design in which the affected families are obtained and N is know, following steps in the Appendix and using $p_0 = p_1 = 0.146$ and $\theta = 0$ as initial values, moment estimates of p_0 and p_1 are 0.095 and 0.490, respectively. This implies that an estimate of β is be 2.207 in this case. This example shows that these two estimates are very close.

In Table 4, we present the values of six different test statistics. Among them, $S_D, S_{\mathbf{M}^a}$ and $\hat{S}_{D,\mathbf{M}^d}$ are from Section 4 for the first kind of local alternative, while $\hat{T}_D, T_{\mathbf{M}^a}$ and $\hat{T}_{D,\mathbf{M}^a}$ are from Section 5 for the second kind of local alternative. This example shows that, for both local alternatives, knowing N can boost the

Table 3. Child survival data reported by Sastry (1997).

Number of siblings	Number of families	Number of affected siblings							
		0	1	2	3	4	5	6	7+
1	267	255	12						
2	285	239	44	2					
3	202	143	41	15	3				
4	110	69	30	9	2	0			
5	104	43	34	15	9	3	0		
6	50	15	18	8	5	3	0	1	
7	21	4	4	7	4	2	0	0	0
8	12	1	2	4	3	1	1	0	0

Table 4. Values and Z-scores of six test statistics.

	Test	Value	Variance	Z-score
H_{01}	S_D	159.50	256.08	9.98
	$S_{\mathbf{M}^a}$	210.75	317.09	11.95
	$\hat{S}_{D,\mathbf{M}^a}$	26.55	143.04	2.23
H_{02} ($\hat{\beta} = 2.207$)	$\hat{T}_D$	6064.65	100605.10	19.12
	$T_{\mathbf{M}^a}$	6295.61	120032.80	18.17
	$\hat{T}_{D,\mathbf{M}^a}$	637.98	29266.46	3.73

power substantially but, if N is known, having unaffected families provides little advantage.

7. Discussion

This paper is motivated by a study of epilepsy (Annegers et al. (1982)). The sampling design is one in which all the affected families in a particular population of a known size is obtained. If the population size were unknown, the power of the tests based on the affected families only would be very limited, because one would be dealing only with the testing for homogeneity of the diseased cases among affected families. Knowing the population size boosts the power of the test for familial aggregation considerably. Since the families without any affected member are not available, we look at how much efficiency is lost caused by not having them. We find that even though all unaffected families are available in that population, the gain of relative efficiency is trivial.

A contribution of the paper is a procedure through which we can easily develop the locally most powerful unbiased tests for testing for familial aggregation. That is, starting with the complete sufficient statistic for the null hypothesis, the locally most powerful unbiased test can be constructed on the conditional log-likelihood. Here the procedure is applied repeatedly to a number of sampling

designs. In addition, it is also applied to the case of proband studies in Fang (2006), in which two types of control groups are examined and compared. Moreover, Fang (2006) also adjusts the methods in this paper to account for covariates, such as age.

It might be of interest to note that the methods developed in the Section 4 are robust in the sense that they do not depend on the distribution of the latent variable. One might specify some parametric distribution, say normal, for the latent variable to gain some power. This is practical despite possible misspecification.

Future research works will focus on developing approaches to the testing for familial co-aggregation of two types of diseases when the sample size is known. However, the procedure developed here cannot be generalized to testing for familial co-aggregation since the complete sufficient statistic does not exist even for the random sampling design. Additionally, much more effort must be put into this problem because the study of two diseases is far more complicated than that of a single disease. For instance, we need to consider whether two diseases share the same mechanism, whether they are caused by two different mechanisms, or whether one disease is the cause of the other.

Appendix

Available at `http://www.stat.sinica.edu.tw/statistica/` as an online supplement.

Acknowledgement

The authors thank Dr. Ottman for describing the ongoing study of epilepsy to them. They also thank the Co-Editors and two referees for their constructive comments and suggestions.

References

Annegers, J. F., Hauser, W. A., Anderson, V. E. and Kurland, L. T. (1982). The risks of seizure disorders among relatives of patients with childhood onset epilepsy. *Neurology* **32**, 174-179.

Bennett, C. A. and Franklin, N. L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*. Wiley, New York.

Commenges, D., Jacqmin, H., Letenneur, L. and Van Duijn, C. M. (1995). Score test for familial aggregation in Probands studies: application to Alzheimer's disease. *Biometrics* **51**, 542-551.

Commenges, D., Letenneur, L., Jacqmin, H., Moreau, T. and Dartigues, J. F. (1994). Test of homogeneity of binary data with explanatory variables. *Biometrics* **50**, 613-620.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Fang, Y. (2006). Testing for familial aggregation when the population size is known. Ph.D. Thesis, Department of Statistics, Columbia University, New York.

Feller, W. (1950). *Introduction to Probability and Its Application*. Vol. 1. Wiley, New York.

Khoury, M. J., Beaty, T. H. and Cohen, B. H. (1993). *AnnegersFundamentals of Genetic Epidemiology*. Oxford University Press, Oxford.

Liang, K. Y. (1987). A locally most powerful test for homogeneity with many strara. *Biometrika* **74**, 259-264.

Pfeiffer, R. M., Gail, M. H. and Pee, D. (2001). Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika* **88**, 933-948.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. *J. Amer. Statist. Assoc.* **92**, 426-434.

Yu, C. and Zelterman, D. (2002). Statistics inference for familial disease clusters. *Biometrics* **58**, 481-491.

Zacks, S. (1985). Pitman efficiency. In *Encyclopedia of Statistical Sciences* (Edited by S. Kotz and N. L. Johnson). Wiley, New York.

Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, U.S.A.

E-mail: matyxf@langate.gsu.edu

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.

E-mail: dan@stat.columbia.edu

(Received February 2007; accepted January 2008)