Complex Segregation Analysis

Han Poong Shin*

During the last few years there has been an interest in models for qualitative attributes, where complex signifies that affection may be caused in two or more ways [1-3]. These models have in common the prediction of variable recurrence risks among families with given parental phenotypes. Segregation analysis has covered only a few cases [4, 5]. The present paper extends segregation analysis to three complex models under two mode of ascertainment.

1. Theory

Let ϕ_{hjk} be the probabilily of a mating of genotypes j and k when h parents are affected $(h=0, 1, 2)$. Let m_{jk} be the risk for affection when parents are of genotypes j and k. (In this paper, we suppose that children are enumerated after onset is complete, so that risk of affection may be equated to incidence.) Then, under complete selection, the probability of r affected among s sibs for an affection independent of birth order is

$$P(r;s,h)=\binom{s}{r}\sum_{j,k}\phi_{hjk}\ m_{jk}^{r}(1-m_{jk})^{s-r}, \tag{1}$$

where $0^0=1$. Under incomplete selection, the probability of ascertaining a sibship with r affected is $1-(1-\pi)^r$, where $0<\pi\leq 1$ is the ascertainment probability, assumed constant. Note that

$$\sum_{r=0}^{s}\binom{s}{r}m_{jk}^{r}(1-m_{jk})^{s-r}[1-(1-\pi)^r]=1-(1-m_{jk}\pi)^s$$

* Researcher, Korea Advanced Institute of Science. Earlier version of this paper was presented in the form of Population and Genetic Laboratory Monograph Series 18, Department of Genetics, University of Hawaii, 1971.

(*cf.* [5, eq. 3]). Therefore,

$$P(r;s,h,\pi)=\frac{\binom{s}{r}[1-(1-\pi)^r]\sum_{j,k}\phi_{hjk}\ m_{jk}^r(1-m_{jk})^{s-r}}{1-\sum_{j,k}\phi_{hjk}(1-m_{jk}\pi)^s}. \tag{2}$$

Among families with s children, r of whom are affected, the probability that the $(s+1)$st child will be affected is

$$Q(r;s,h)=\frac{\sum\phi_{hjk}m_{jk}^{r+1}(1-m_{jk})^{s-r}}{\sum_{j,k}\phi_{hjk}m_{jk}^r(1-m_{jk})^{s-r}}. \tag{3}$$

Equations (1∼3) hold for any model, however complicated, which defines ϕ_{hjk} and m_{jk} independently of birth order. Equations (1) and (2) are fundamental to analysis of family data, while equation (3) forms the basis for genetic counseling.

It is not difficult to derive ϕ_{hjk} under general assumptions, but, in practice, the solution for random mating with fertility of affected independent of genotype is most important. Except for parents of unspecified phenotype ($h=?$), it is not necessary to assume that fertility of affected is normal, only that all genotypes of affected are equally fertile. Then, if p_j is the frequency of genotype j in the general population, the probability that an affected parent (or individual) would have genotype j is $a_j=p_jf_j/A$, where f_j is the probability that genotype j be affected and

$$A=\sum_j p_jf_j$$

is the incidence. Similarly, the probablity that a normal parent (or individual) would have genotype j is $c_j=p_j(1-f_j)/(1-A)$. Thus, the two assumptions of random mating and uniform fertility within the phonotypes normal and affected allow us to write

$$\begin{aligned}\phi_{hjk}&=c_jc_k \qquad (h=0)\\ &=c_ja_k \qquad (h=1)\\ &=a_ja_k \qquad (h=2)\\ &=p_jp_k \qquad (h=?),\end{aligned} \tag{4}$$

where only the last result assumes equal fertility of normal and affected. If fertility is severely reduced by affection, the analysis should take $h=0$ rather than $h=?$, although for a rare trait the two results cannot differ by much.

Models which assume that different loci act independently on f_j are called discontinuous. Models which assume that different loci act independently on some other variable, termed liability, which is not linearly proportional to f_j are called quasi-continuous. Sex effects may be incorporated into the p_j and m_{jk}, but will be ignored in this paper.

2. Model 1: The Generalized Two-Allele Single Locus

Consider a gene G with frequency q which determines risks $t+z$, $td+z$ and z in the genotypes GG, GG', and $G'G'$, respectively, where d is the dominance of G, t is the penetrance of GG, and z is the frequency of non-heritable (sporadic) cases (table 1). The 3×3 matrix of risks m_{jk} is

$$M=\begin{pmatrix} t+z & \dfrac{t+td+2z}{2} & td+z \\ \dfrac{t+td+2z}{2} & \dfrac{t+2td+4z}{4} & \dfrac{td+2z}{2} \\ td+z & \dfrac{td+2z}{2} & z \end{pmatrix}. \tag{5}$$

Maximum likelihood analysis conveniently takes the parameters A, d, t, and $x=z/A$, which constrain the gene frequency

$$q=\begin{cases} \dfrac{-td+\sqrt{t^2d^2+A(1-x)(1-2d)t}}{(1-2d)t}, & \text{if } d\neq1/2,\ t>0,\\ A(1-x)/t, & \text{if } d=1/2,\ t>0, \end{cases}$$

$0<A, d, t, x<1$, and $t+z\leq 1$. Under incomplete selection, there is virtually no information about A, which must be estimated from other evidence [6].

We define the rank of a hypothesis as the number of parameters to be estimated. The most important special cases of rank 2 are:

No phenocopies	$(x=0)$
G dominant	$(d=1)$
G recessive	$(d=0)$
G additive	$(d=1/2)$
GG completely penetrant	$(t=1-z)$

The most important special cases of rank 1 are:

No phenocopies, GG completely penetrant	$(x=0,\ t=1)$
G dominant, completely penetrant	$(d=1,\ t=1-z)$
G additive, completely penetrant	$(d=1/2, t=1-z)$
No phenocopies, G additive	$(d=1/2,\ x=0)$
G recessive, completely penetrant	$(d=0,\ t=1-z)$
No phenocopies, G recessive	$(d=0,\ x=0)$
No phenocopies, G dominant	$(d=1,\ x=0)$

A hypothesis may be said to be better than an alternative of the same rank if it has a smaller likelihood-ratio criterion,

$$\chi^2(L)=2\sum_{s,r,h} n_{srh}\ \ln(n_{srh}/e_{srh}) \qquad (6)$$

(Barrai et al. [6]) where n_{srh}, and e_{srh} are the observed and expected numbers of sibships of size s with r affected from h affected parents, and the sum is over all nonzero values of n_{srh} and e_{srh}. By definition, $e_{srh}=n_{sh}P(r;s,h)$, where n_{sh} is the observed number of families of size s with h affected parents and $P(r;s,h)$ is given by equation (1).

The quantity

$$L = \prod_{s,r,h} P(r;s,h)^{rsrh}$$

is called the likelihood and in the limit for large samples, approaches the multinormal form $ce^{-x^2/2}$. The hypothesis which maximizes L in a large body of good data gives the best basis for genetic interpretation and counseling.

A hypothesis may be said to be appreciably better than an alternative of lower rank if its likelihood-ratio criterion is smaller by at least 4. However, only if the one hypothesis is a special case of the other can statistical significance be asserted (for example, if the hypothesis that $d=0$ with A known and t, x iterated simultaneously gives $\chi^2=20$, and the sub-hypothesis that $d=0$, $t=1$ with A known and x iterated gives $\chi^2=25$, then the difference of 5 is distributed in large sample theory as χ^2_1, with $P=.025$).

Discontinuous models may be generalized further in two directions: we may assume that any one of n loci can independently produce affection (the genetic-load model) or that two or more loci interact (epistasis). The first situation can be resolved for rare recessive genes by consanguinity analysis, and more generally by finer phenotypic discrimination-ideally at the level of protein structure. The possibility of analyzing epistasis is remote in man unless the effect of each locus can be recognized separately (as for the Lewis-secretor interaction).

3. Model 2: Beta Distribution of Risk

A reasonable and convenient generalization of complex models is the distribution introduced by Gini [7] and Skellam [8], which assumes that the recurrence risk m varies among families with h affected parents according to the beta density,

$$(fm) = \frac{(\xi-1)!}{(\zeta-1)!(\xi-\zeta-1)!}\, m^{\zeta-1}(1-m)^{\xi-\zeta-1}, \qquad (7)$$

$0<\zeta<\xi$ and $0<m<1$, where the symbols have the same meaning as in

Morton et al. [3] but different from Morton [4]. Then the Gini-Skellam distribution of r affected in families of size s is

$$P(r;s,\zeta,\xi)=\int_0^1 f(m)\binom{s}{r}m^r(1-m)^{s-r}dm$$
$$=\frac{\binom{r+\zeta-1}{r}\binom{s-r+\xi-\zeta-1}{s-r}}{\binom{s+\xi-1}{s}}. \tag{8}$$

The population incidence (the probability that the first child be affected; the mean affection risk) for families with a given value of h is

$$A=\int_0^1 m\ f(m)dm=\frac{\zeta}{\xi}.$$

Also, given h, the segregation frequency (the probability that the sib of an affected singleton be affected: the mean recurrence risk) is

$$T=\frac{\int_0^1 m^2 f(m)dm}{\int_0^1 m\ f(m)dm}=\frac{\zeta+1}{\xi+1}.$$

These are the special cases $s=r=0$ and $s=r=1$ of the probability that the next child is affected after s sibs have been born, r of whom were affected:

$$Q(r;s)=\frac{\int_0^1 f(m)m^{r+1}(1-m)^{s-r}dm}{\int_0^1 f(m)m^r(1-m)^{s-r}dm}=\frac{\zeta+r}{\xi+s}. \tag{9}$$

Thus, we may substitute A and T for the parameters,

$$\xi=\frac{1-T}{T-A},\quad \zeta=A\xi$$

Under incomplete selection, the distribution of r affected becomes

$$P(r;s,\zeta,\xi,\pi)=\frac{[1-(1-\pi)^r]\binom{r+\zeta-1}{r}\binom{s-r+\xi-\zeta-1}{s-r}}{\sum_{r=1}^{s}[1-(1-\pi)^r]\binom{r+\zeta-1}{r}\binom{s-r+\xi-\zeta-1}{s-r}}. \tag{10}$$

With incomplete selection, the estimation of A for given h is more difficult under model 2 than for a specific discontinuous or quasi-continuous hypothesis. Therefore, model 2 is most useful for parents of unspecified phenotype or under complete selection.

4. Model 3: The Polychotomized Normal Distribution of Liability

Falconer [1] considered an additive liability scale, normally distributed, one tail of which determines affection. This type of variation has been called quasi-continuous [9]. In the original derivation of Falconer's model, the phenotypic liability had a sharp threshold for affection. This seemed implausible to Edwards [2], who introduced a different model which allowed the risk to exceed unity. Smith [10] showed that Falconer's model could be derived by assuming a normal distribution of genetic liability, acted on by a cumulative normal risk function representing environmental liability. This removed any lingering doubts about the adequacy of Falconer's model to represent affection caused by additive genetic liability. Attempts to fit both Falconer's and Edwards' models to actual data revealed no advantage in the latter (Morton et al. [3]). Together with the device of polychotomizing the normal distriction, which makes it possible to replace multiple integration by summation, Smith's derivation of Falconer's model makes it the method of choice to represent quasi-continuous variation.

Let a normal distribution of genetic liability $f(x)$ be partitioned into n nonoverlapping classes, the kth of which has limits $L_{1k}<L_{2k}$ and therefore probability

$$p_k=Q(L_{1k})-Q(L_{2k}) \quad (\sum_k p_k=1), \tag{11}$$

where

$$Q(x)=\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt.$$

Let $L_k=(L_{1k}+L_{2k})/2$. Assume that the genetic liability within this class is a constant, which we take to be

$$\begin{aligned} x_k &= L_k && \text{for } -\infty<L_{1k},\ L_{2k}<\infty \\ &= f(L_{1k})/Q(L_{1k}) && \text{for } L_{2k}=\infty \\ &= -f(L_{2k})/Q(L_{2k}) && \text{for } L_{1k}=-\infty \end{aligned}$$

The standardized normal deviate x_k corresponds, in the distribution of genetic liability with variance T, to $x_k\sqrt{T}$ where T is the heritability. The standardized deviation of Z, the threshold for affection, is

$$x_k'=\frac{Z-x_k\sqrt{T}}{\sqrt{1-T}} \tag{12}$$

[10], table 1). Therefore, the incidence in the population is

$$A=\sum_k p_kQ(x_k')=Q(Z),$$

and the probability of liability class k among affected individuals is

$$a_k=p_kQ(x_k')=Q(Z).$$

(Note that the symbols T, x, and Z have different meanings for models 1 and 3, but A is always the incidence).

Similarly, the probability of liability class k among normals is

$$c_k=p_k[1-Q(x_k')]/(1-A)$$

If one parent belongs to class j and the other to class k, the mean genetic liability of the children is $(x_j+x_k)\sqrt{T}/4$, and the mean deviation of the threshold is

$$x_{jk}'=\frac{Z-(x_j+x_k)\sqrt{T}/4}{\sqrt{1-T/2}}$$

This defines the risk of affection

$$m_{jk}=Q(x_{jk}') \tag{13}$$

With these definitions of $p, a, c,$ and $m,$ we may use the equations of the first section to perform a segregation analysis of Falconer's model. For the calculations to be accurate with large $s,$ we must use many classes : we have taken $n=52$ (corresponding to 50 equally spaced classes from -4 to $+4$ and the two terminal classes) with the approximations of Hastings [11] to evaluate $Q(x')$ at

$$-4+.16(k-1)<x_k<-4+.16k, \quad \text{for } k=1, 2, \cdots, 50$$
$$x_{51}=f(4)/Q(4), \quad p_{51}=Q(4),$$
$$x_{52}=-f(4)/Q(4), \quad p_{52}=Q(4).$$

It is convenient to estimate the parameters A and $T,$ the population incidence and the heritability, respectively; note that A applies to the general population, and not merely a specific value of h as in model 2.

5. Inbreeding Effects

Let the incidence be $A+BF,$ where F is the coefficient of inbreeding and B is the genetic load. Morton [12, eq. 17] provided for quasi-continuity the approximation

$$B=A\,T\,(Z^2+1)/2, \tag{14}$$

where T is the heritability and Z the threshold for affection. As A approaches zero, the ratio B/A becomes large and incapable of discriminating between mutation and segregation loads.

Table 1. The Two-Allele Model

Genotype	GG	GG'	$G'G'$
Index, j	1	2	3
Frequency, p_j	q^2	$2q(1-q)$	$(1-q)^2$
Probability of affection, f_j	$t+z$	$td+z$	z

Note: Risk in parametric population, $A=q^2t+2q(1-q)td+z$.

For the generalized two-allele single locus, we have

$$B=q(1-q)t(1-2d)=qt-(1-x)A. \tag{15}$$

The B/A becomes large as q, x approach zero [13].

For any genetic model, complex segregation analysis predicts the effects of inbreeding. This provides an independent test of the model, which is of little power unless the inbreeding effects are large.

6. Affection in Relatives

In the most important case for genetic counseling, ego is the sib or child of a proband and equation (3) gives his risk. Sometimes more remote relationship is involved. If R is the coefficient of relationship, model 3 gives as the recurrence risk(*i.e.*, the probability that ego is affected, given a proband of relationship R)

$$Q(R)=\sum_k a_k\, Q(x_k'), \tag{16}$$

where

$$x_k'=-\frac{Z-x_kR\sqrt{T}}{\sqrt{1-R^2T}}$$

(Smith, [10]).

For the single-locus model, two coefficients are required, the relationship R and the probability of double identity by descent K, where $K=0$ for unilineal relatives, 1/4 for sibs, and 1/16 for double cousins. The recurrence risk, if neither proband nor ego is inbred, is

$$Q(R,K)=\sum_{j,k} a_j p_{jk} f_k \tag{17}$$

Here p_{jk} is an element of the matrix $P=(1-2R+K)T_0+2(R-K)T_1+KT_2$, where the T_n matrix is the conditional probability of K, given j, when there are exactly n alleles identical by descent, and

$$T_0=\begin{pmatrix} q^2 & 2q(1-q) & (1-q)^2 \\ q^2 & 2q(1-q) & (1-q)^2 \\ q^2 & 2q(1-q) & (1-q)^2 \end{pmatrix},$$

$$T_1=\begin{pmatrix} q & 1-q & 0 \\ q/2 & 1/2 & (1-q)/2 \\ 0 & q & 1-q \end{pmatrix},$$

$$T_2=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

(Li [14]; Elston and Campbell [15]). (The matrices T_0, T_1, T_2 correspond to Li's **0**, T, I, respectively, and have no relation to the paremeter T used in models 1, 2, and 3). Thus, segregation analysis predicts recurrence risks in more remote relatives which provide an independent test of the genetic model, although with less power than the first-degree relatives used for segregation analysis.

7. Discussion

Limited experience suggests that discrimination between discontinuous model 1 and quasi-continuous model 3 will often be difficult. Dominance and a high ratio of recurrence risk to incidence favor model 1. Lack of dominance and a low or moderate ratio of recurrence risk to incidence suggest quasi-continuity, but do not rule out a discontinuous model, which, with more parameters, is necessarily more flexible. Strong evidence for quasi-continuity requires that goodness-of-fit be no better for a discontinuous model with three parameters than for a quasi-continuous model with two. We doubt that such evidence is often feasible, and a decision may have to be suspended until more refined techniques of observation recognize major gene effects, leaving a residual which must come more and more to approximate quasi-continuity, even if the relevant familial factors are nongenetic.

Although the mode of inheritance may remain in doubt, complex segregation analysis leads to useful estimates of recurrence risks for genetic counseling. Here models with a significantly poor fit to the data are rejected. Among the remainder, models with a minimum number of parameters are preferred, and among these the models with the best fit to the data. When two or more models of the same rank fit about equally well, in the absence

of any firm decision about the mode of inheritance all relevant predictions may be used as a guide to genetic counseling. For example, if three acceptable models predict .09, .11, and .13 for a particular risk category, we might give the risk as "between .09 and .13" or more simply as "about .11".

Complex segregation analysis provides more powerful tests of genetic hypotheses and more reliable recurrence risks than can be obtained when sibships of different compositions are pooled. The equations of this paper have been incorporated into a computer program COMSEG, a description of which is available from the author.

SUMMARY

Segregation analysis has been extended to several complex models for qualitative attributes under two modes of ascertainment, providing a basis for genetic counseling.

REFERENCES

[1] Falconer, D. S., "The Inheritance of Liability to Diseases, Estimated from the Incidence among Relatives," *Annal of Human Genetics*, Vol. 29, 51-76.

[2] Edwards, J. H., *Linkage Studies of Whole Populations Proceedings, Third International Cog. Human Genetics*, edited by Crow, J. F. and Neel, J. V., Baltimore: Johns Hopkins Press, 1967, 483-489.

[3] Morton, N. E., Yee, S. and Elson, R. C., "Discontinuity and Quasi-Continuity:Alternative Hypotheses of Multifactorial Inheritance," *Clin Genetics* Vol. 1, 81-94, 1970.

[4] Morton, N. E., "Segregation Analysis," in *Computer Applications in Genetics*, edited by Morton, N. E., Honolulu: University of Hawaii Press, 1969, 129-139.

[5] Elandt-Johnson R. C., "Segregation Analysis for Complex Modes of

Inheritance," *American Journal of Human Genetics,* Vol. 22, 129-144, 1970.

[6] Barrai, I., Mi, M. P. and Yasuda, N., "Estimation of Prevalence under Incomplete Selection," *American Journal of Human Genetics,* Vol. 17, 221-236. 1965.

[7] Gini, C., "Considerazioni sulle Probabilita a Posteriori e Applicazioni al Raporto dei Sessi Nelle Nascite Umane," *Studi Economico-Giuridici,* Vol. 3, 5-41. 1911.

[8] Skellam, J. G., "A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials," *Journal of the Royal Statistical Society, B,* 10, 257-261, 1948.

[9] Grueneberg, H., "Genetical studies on the skelton of the mouse:IV. Quasi-continuous variations," *Journal of Genetics,* 51, 95-114, 1952.

[10] Smith, C., "Heritability of liability and concordance in monozygous twins," *Annal of Human Genetics,* 34, 45-91, 1970.

[11] Hastings, C., *Approximations for digital computers,* Princeton, N. J., Princeton University Press, 1955.

[12] Morton, N. E., "The detection of major genes under additive continuous variation," *American Journal of Human Genetics,* 19, 23-34, 1967.

[13] Morton, N. E., Crow, J. F., Muller, H. J., "An estimate of the mutational damage in man from data on consanguineous marriage," *Proceedings of National Academy of Sciences, U. S. A.,* 42, 855-863, 1956.

[14] Li, C. C., *Population Genetics,* University of Chicago Press, 1963.

[15] Elston, R. C., Camphell, M. A., Schizophrenia, "Evidence for the Major Gene Hypothesis," *Behavioral Genetics,* in press, 1971.