The American Journal of Human Genetics

Vol. 11 No. 1 *March 1959*

Genetic Tests Under Incomplete Ascertainment

NEWTON E. MORTON

Department of Medical Genetics, University of Wisconsin

GENETIC TESTS are commonly made without separation of the major sources of discrepancy, such as isolated cases and nonsegregating families. Often attention is directed toward estimation of the mean segregation frequency, rather than to specific tests of genetic hypotheses. This difference in emphasis is responsible for much computational difficulty, and therefore perhaps also for the failure of human geneticists to examine by stringent statistical methods the impressions obtained from family data. Only such tests can resolve discrepancies and discriminate among alternatives, such as phenocopies, mutations, and incomplete penetrance. Fortunately, it is possible by extension of existing formulae (Haldane, 1938, 1949; Finney. 1949) and by use of maximum likelihood scores (Rao, 1952) to obtain simple and efficient tests of a variety of genetic hypotheses (Morton, 1958).

Definitions, assumptions, and methods

A *proband* is an affected person who at any time was detected independently of the other members of the family, and who would therefore be sufficient to assure selection of the family in the absence of other probands. The first proband detected in a family may be designated the index case, but the index case is no more important than the other probands, and valuable information will be lost if the total number of probands is not recorded. The term propositus will be avoided as ambiguous, since it has been used by some authors to signify the index case, and by others to include all probands. In a sibship of size s, it will be convenient to let a be the number of probands, b the number of affected children not probands, $a + b = r$, and $c = s - r$ be the number of normal children.

Families with no affected children, one affected child, and more than one affected child are called *nonsegregating*, *simplex*, and *multiplex*, respectively. An affected child is called *isolated* or *familial* in simplex and multiplex families, respectively. Isolated cases are of two possible types. *Chance* isolated cases are of the same origin as familial cases, and the other children in such families have the same *a priori* probability of being affected. *Sporadic* cases are of different origin from the familial and chance isolated cases (mutation, diagnostic error, phenocopy, etc.), and are assumed to be rare and independent, so that the probability is negligible that a familial case be of the same origin as a sporadic case.

Ascertainment may include both selection of families for analysis and recognition of segregating and nonsegregating families. *Complete selection* signifies random sampling of families *through the parents*, without consideration of the phenotypes of the

This work was supported by a grant from the Rockefeller Foundation.
Received April 25, 1958.

children. *Incomplete selection* denotes selection of families *through the children*, with exclusion of nonsegregating families. Five methods of ascertainment will be considered.

Complete selection

1. Separation of segregating and nonsegregating families, with failure to distinguish between homozygous and heterozygous parents in nonsegregating families. This separation is appropriate with dominant-recessive gene pairs when the parental genotypes can be inferred with certainty only in segregating families, for which it is equivalent to truncate selection (see below).

2. Separation of homozygous and heterozygous parents by direct inspection (for codominant gene pairs and rare "dominants" not selected through the children) or from information about the grandparents. In families of a given mating type and size, the distribution of the number of affected children is a complete binomial.

Incomplete selection

3. *Truncate selection,* with random sampling of segregating families. Families with many affected children are no more likely to be selected than families with only one affected child, so that in sibships of a given mating type and size, the distribution of the number of affected children is a truncated binomial, with the first term missing. The phrase "complete selection of affected individuals" (Bailey, 1951) will be avoided as cumbersome and liable to confusion with complete selection (1 and 2 above).

4. *Single selection,* with the probability of ascertainment so small that there is virtually no chance of having two probands in one sibship, and the probability that a family be ascertained is proportional to the number of affected children.

5. *Multiple selection,* with a constant but arbitrary probability of ascertainment. The ascertainment probability π is the chance that an affected person be a proband. There may be from 1 to r probands in a family with r affected, and each proband may have $t \geq 1$ ascertainments. Multiple selection includes single and truncate selection as limiting cases.

In addition to the restrictions implicit in these definitions, the following assumptions are made.

1. The ascertainment probability π is constant, and all probands in a family are ascertained independently.

2. In multiplex and simplex families of the same origin, there is a constant *a priori* probability p that a child be affected (and the complementary probability $q = 1 - p$ that he not be affected).

3. Sporadic cases make up a proportion x of all cases in the population, and simplex families with sporadic cases constitute a proportion w of families of size s with affected children. Excluding sporadic cases, the mean number of affected children in a sibship of size s with at least one member affected is:

$$\bar{r} = \sum_{r=1}^{s} \frac{r\binom{s}{r} p^r q^{s-r}}{1 - q^s} = \frac{sp}{1 - q^s}$$

and

$$x = \frac{w}{w + (1 - w)\bar{r}}.$$

Substituting for $\bar{r}$ and rearranging,

$$w = \frac{xsp}{1 - q^s - x(1 - sp - q^s)}$$

If sporadic cases are not related to parity or parental age it may be shown that x is independent of s. For then the expected number of sporadic cases in a family of size s is $s\gamma$, and the expected number of nonsporadic cases is sf, where γ is the frequency of sporadic cases in the general population and f is the frequency of nonsporadic cases. Therefore the frequency of sporadic among all cases is $x = \gamma/(\gamma + f)$, which is independent of s. Since w increases with s, it is a less useful parameter than x. The condition for familial cases of sporadic origin to be negligible is $x\mu \ll (1 - x)\,p$, where μ is the probability of affection among sibs of sporadic cases. If the occurrence of sporadic cases is random among families, $\mu = \gamma$.

4. Let h be the probability that a parent be of genotype TT if his phenotype is the same as Tt. If mating is random and there are no sporadic cases $(x = 0)$,

$$h = \frac{f_T^2}{f_T^2 + 2f_T f_t} = \frac{f_T}{f_T + 2f_t},$$

where f_T, f_t are the population gene frequencies of T, t. In a few cases involving multiple alleles, it will simplify the algebra to use h in a more general sense, as the probability that a parent either be homozygous, or that heterozygosity not be detectable because of the genotype of the other parent.

The distributions to be investigated arise from these assumptions and ancillary ones about p, π, x, and h. The null hypothesis specifies some theoretical value for p, and the other parameters either take theoretical values or are maximized subject to the hypothesis about p and the remaining parameters. For each independent observation, the maximum likelihood functions for any parameter, say θ, consist of a score whose expectation is zero on the null hypothesis,

$$u_\theta = \frac{\partial \ln L}{\partial \theta},$$

and its conditional variance, which is also the information about θ,

$$k_\theta = E\{u_\theta^2\} = -E\left\{\frac{\partial^2 \ln L}{\partial \theta^2}\right\}$$

where L is the probability of the observation, and u and k are evaluated at p_0, π_0, x_0, and h_0, the values of the parameters specified by the null hypothesis. Suppose the sample consists of m such observations and that none of the other parameters is estimated from the sample, and let $\Sigma u = U$ and $\Sigma k = K$. Then on the null hypothesis, U^2/K in the theory of large samples has the χ^2 distribution with one degree of freedom (testing the goodness of fit of θ_0), and, on the same assumptions, $\Sigma[u^2/k] - U^2/K$ has the χ^2 distribution with m − 1 degrees of freedom (testing homogeneity of θ). Furthermore, if the first χ^2 indicates a significant discrepancy, and this is thought not to be due to erroneous assumptions about other parameters, then θ may be estimated as

$$\theta^* = \theta_0 + U/K$$

with standard error $\sigma_{\theta^*} = \sqrt{1/K}$, if θ is constant, or approximately

$$s_{\theta^*} = \sigma_{\theta^*}\sqrt{\frac{\Sigma\left(\frac{u^2}{k}\right) - U^2/K}{m - 1}}$$

if θ varies among families. This is the first step in the iterative approach to the exact maximum likelihood solution, to which it is often a close approximation, (Rao, 1952, Chapter 4).

The above formulae may be generalized to the case where n parameters are to be estimated or tested against some null hypothesis. Let U_i be the total score for the i^th^ parameter, K_{ij} be the covariance between U_i and U_j, and K^{ij} be the corresponding element in the inverse matrix of the K_{ij}. Then to test a null hypothesis with respect to n parameters,

$$\chi^2_n = \sum_{i,j=1}^{n} U_i U_j K^{ij} = \sum_{i=1}^{n} U_i^2 K^{ii} + 2\sum_{i<j} U_i U_j K^{ij}.$$

To test homogeneity of θ_i, assuming homogeneity of n − 1 other parameters estimated from the sample, $\chi^2_{m-n} = \Sigma(u_i^2/k_i) - U_i^2/K_{ii}$. To estimate the ith parameter θ_i from an initial estimate θ_{0i}, $\theta_i^* = \theta_{0i} + \sum_{j=1}^{n} U_j K^{ij}$, omitting from the original K_{ij} matrix any parameters not estimated from the sample. The standard error of this estimate is $\sigma^*_{\theta i} = \sqrt{K^{ii}}$ if θ_i is constant and the assumptions about the other parameters are correct. If any parameter is heterogeneous, an empirical standard error for θ_i is $s_{\theta^* i} = \sqrt{\chi^2/(m - n)}$ (Rao, 1952, Chapter 4).

This empirical standard error is approximate in two ways: it is an estimate of the actual sampling error, to which it converges in large samples; with heterogeneity in θ_i, the maximum likelihood (M.L.) estimate is not necessarily unbiased, even in the limit for large samples, but converges to some other value different from, although usually near, the mean value $E(\theta_i)$. If the data cannot be separated into homogeneous groups, this bias is unavoidable, and the M.L. estimate is as satisfactory as any other.

The appendix gives formulae for the five modes of ascertainment. Where appropriate, each family is scored as five independent observations, corresponding to:

1. Separation of segregating and nonsegregating families.
2. Among segregating families, separation of simplex and multiplex families.
3. Among multiplex families, the distribution of r.
4. Among multiplex families under multiple selection, the distribution of probands among r affected.
5. Among probands, the distribution of t ascertainments.

Tests of homogeneity among these sources will detect discrepancies obscured in the pooled data and help to identify disturbing factors. Homogeneous data may always be pooled, since the scores and variances are additive and jointly exhaust the information in the sample, providing a fully efficient analysis in the neighborhood of the null hypothesis.

Incomplete ascertainment

It has been assumed that π is constant and ascertainments are independent. However, human data may depart from this model in several ways, which may be distinguished by a test of homogeneity of estimates of π from ascertainments, probands, and affected children.

1. Ascertainments may not be independent, because referral from one source favors or precludes referral from another. If this cannot be avoided by careful definition of the sources of ascertainment, the method of §7 in the appendix will not be applicable to the distribution of the number of ascertainments. However, probands will still give valid information, if the probability that an affected individual be a proband is independent of the number of his affected sibs, the severity of their affection, and the number of other probands in the family.
2. The probands may not be correctly identified, either through failure to record probands after the index case or through counting as probands sibs who were in fact ascertained from other family members. When this is the only discrepancy from the ascertainment model, the analysis will be valid if the number of probands is neglected, and π estimated solely from the distribution of r in segregating families. However, much of the genetic information is lost if probands and ascertainments are not identified.
3. If a trait is more likely to be correctly diagnosed when it is familial, then isolated cases will be poorly or excessively represented. The analysis may be restricted to multiplex families.
4. The number of ascertainments (t) may be known even if the number of probands has been recorded incompletely or not at all. Providing ascertainments are independent, the method of §7 of the appendix may still be used.
5. The probability of ascertainment may be heterogeneous among families because the trait is a mixture of entities, which as far as possible should be separated before the analysis proceeds.
6. The ascertainment model may be systematically wrong if the cases are collected from occasional reports in the literature or other biased sources. In this event it is still possible that the distribution of r among segregating families, or at least among multiplex families, may be adequately described by some effective ascertainment probability π, and that a valid analysis of p may be carried out from the empirical standard error.

Obviously any test on p, x, h, or π depends on the accuracy of the ascertainment model. Unfortunately, analysis of incomplete ascertainment in the past has been so inadequate, that the magnitude of the error of this method cannot be assessed. However, it is hopeful that even data from the medical literature seem to fit fairly well for albinism (Haldane, 1949), and there is reason to suppose that a more systematic collection of cases would in general conform more closely to the ascertainment model of this paper.

Incomplete penetrance and delayed onset

The assumption that p is constant neglects interfamily heterogeneity in penetrance and age at onset. In this case the analysis will be approximate, but the use of an empirical standard error helps to protect against invalid conclusions.

For a common trait, incomplete penetrance so complicates the analysis that the methods of this paper are not always applicable, since several segregation ratios will occur within some phenotypic mating classes. However, if a trait is rare enough so that nearly all segregating matings are of one type, the analysis presents no difficulty. The expected segregation frequency will then be the product of the theoretical value p_0 and the average penetrance for the sample (y).

Delayed onset constitutes an important special case of incomplete penetrance. Let f(z) be the frequency of age z at death or last examination among normal and affected siblings, $f_1(z)$ be this frequency with the index cases excluded, and G(z) be the cumulative frequency of onset at age z among affected cases. Then if incomplete penetrance is entirely due to delayed onset, the estimate of the average penetrance in the sample is $y = \int f(z)G(z)dz$ for complete selection and $\int f_1(z)G(z)dz$ for single selection, where integration is over the range of z. Since these are the two limiting cases, the best estimate of y should lie between these values.

As with ascertainment, it is not clear how adequate this model for the segregation ratio will be. Reliable results may be expected if the data are homogeneous. However, the assumption of incomplete penetrance is so consistent with variable p, that it might in practice be difficult to recognize other kinds of heterogeneity. Only actual trial of these methods will determine their limits, but the more regular the sampling procedure and the higher the penetrance, the greater their precision will be.

An example of complete selection

Taylor and Prior (1938) and Race and others (1942) presented a series of 236 families tested for the A_1A_2BO blood group factors, and analyzed them by partition into segregating and nonsegregating families. Pooling reciprocals, there are 21 different mating types, in six of which there is no dominance and separation of homozygous and heterozygous parents is by direct inspection. The progeny distributions in the 15 remaining types give 28 degrees of freedom for tests of genetic hypotheses by calculation of expected numbers of families.

If we separate parental segregations where possible, and apply the methods of the present paper, there are only six segregation types, in two of which there is no dominance (h = 0). Considering backcrosses and intercrosses separately, there are eleven mating types, which require calculation of only four values of h (table 1). Letting p_1 , p_2 . q and r denote the gene frequencies of A_1 , A_2 , B, and O, respectively, and using the estimates of Ikin, Prior, Race, and Taylor (1939) from an English sample of 3,459 persons, the values of h are computed as follows:

Type 1. The probability that an A_1 parent is A_1A_1 and not A_1O or A_1A_2 is $h = p_1/[p_1 + 2(p_2 + r)] = .1252$

TABLE 1. CLASSIFICATION OF ABO MATINGS

Segregation type			Backcross	Segregants		Intercross
	T	t		T	t	
1	A_1	$A_2 + O$	$A_1 \times O$	A_1	A_2, O	$A_1 \times A_1$
			$\times A_2$	A_1	A_2, O	
			× B	A_1, A_1B	B, A_2B, O	
			$\times A_2B$	A_1, A_1B	B, A_2B	
			$\times A_1B$ exclude non-B progeny	A_1B	A_2B, B	
2	A_1	B	$A_1B \times O$	A_1	B	$A_1B \times A_1B$
			$\times A_2$	A_1	A_2B, B	
			× B	A_1, A_1B	B	
			$\times A_2B$	A_1, A_1B	A_2B, B	
			$\times A_1$	A_1	B, A_2B, A_1B	
3	A_2	O	$A_2 \times O$	A_2	O	$A_2 \times A_2$
			× B	A_2, A_2B	B, O	
			$\times A_2B$ exclude non-B progeny	A_2B	B	
			$\times A_1B$ exclude non-B progeny	A_2B	B	
4	A_2	O	$A_2 \times A_1$ exclude A_1 progeny	A_2	O	
5	A_2	B	$A_2B \times O$	A_2	B	$A_2B \times A_2B$
			× B	A_2B, A_2	B	
			$\times A_1B$	A_1, A_2B	A_1B, B	
			$\times A_1$	A_1, A_2	A_1B, B, A_2B	
			$\times A_2$	A_2	A_2B, B	
6	B	O	B × O	B	O	B × B
			$\times A_1B$ exclude B progeny	A_1B	A_1	
			$\times A_2B$ exclude B progeny	A_2B	A_2	
			$\times A_1$	A_2B, A_1B, B	A_1, O, A_2	
			$\times A_2$	A_2B, B	O, A_2	

Type 3. The probability that an A_2 parent is A_2A_2 and not A_2O is $h = p_2/(p_2 + 2r) = .0501$

Type 4. The probability that an A_2 parent is A_2A_2 and not A_2O, *or* that the non-A_1 allele of a heterozygous A_1 parent is A_2 and not O, is

$$h = 1 - \{2r/(p_2 + 2r)\}\{r/(p_1 + r)\} = .1407.$$

This is a case where multiple allelism makes it convenient to define h, not as the probability of homozygosity, but as the probability either of homozygosity or of heterozygosity not detectable because of the genotype of the other parent. This mating is treated as a backcross, because segregation of the A_1 parent is eliminated by exclusion of A_1 progeny.

Type 6. The probability that a B parent be BB and not BO is $h = q/(q + 2r) = .0443$

In applying these formulae, the relationship of some parents *inter se* has been ignored, which has the effect of exaggerating deviations from the null hypothesis. Since all tests are in excellent agreement with hypothesis, no more elaborate treatment is required.

Table 2 summarizes the results. Only nine of the eleven possible mating types occur in these data. The frequencies of segregating and nonsegregating families agree well with the values of h calculated from the English gene frequency estimates on the assumption of random mating and negligible selection, as shown by the analyses of both p and h.

In the analyses of segregating families, types 3 and 4 are pooled, but type 1 is divided into A_1/O and A_1/A_2 segregations. This gives six possible segregation types, or 13 mating types when intercrosses and backcrosses are distinguished, of which nine are observed in these data. Agreement with hypothesis is again excellent.

It is noteworthy that the partition into segregating and nonsegregating families gives only 743 of the total of 1724 units of genetic information in these data, or 43 per cent, while the analysis of segregating families accounts for the remaining 57 per cent. Much of the value of laboriously collected data will be lost unless both sources of information are utilized.

The application of these methods to tests on reciprocal crosses and other matings within a segregation type is obvious.

TABLE 2. ABO DATA

Source	Analysis of p		
	Σk_{pp}	d.f.	x^2
1) $r = 0$ vs. $r > 0$	743		
Mating types		9	5.11
Families within types		210	214
2) $r = 1$ vs. $r > 1$	603		
Mating types		9	12.53
Families within types		103	106
3) r among $r > 0$	378		
Mating types		6	7.63
Families within types		35	38
Total	1724		
Mating types		9	9.92
Families within types		210	236
		Analysis of h	
		d.f.	x^2
Mating types		9	3.20
Families within types		210	216
$\Sigma k_{hh} = 969$			

DISCUSSION

There have been four stages in the development of segregation analysis in man. At first, tests of significance based on complete selection were applied to rare dominant pedigrees and later to codominant factors. Next, the disturbing effects of truncate selection were recognized, leading to the development of the *a priori* methods of Bernstein, Lenz, and others, and the *a posteriori* method of Haldane. (The methods of the present paper are *a priori* in the sense of starting with a test of some null hypothesis, but *a posteriori* in leading by iteration to the maximum likelihood estimate if the null hypothesis is rejected.) Thirdly, multiple selection was considered by Weinberg in the proband method, which is not fully efficient except in the limiting case of single selection, and more elaborately by later authors, none of whom used the large amount of information present in the number of ascertainments of probands.

Finally, interest in mutation and other sources of sporadic cases led to their inclusion in the general models of this paper, with separation from incomplete penetrance and other superficially similar phenomena. It seems remarkable that this generalization should have required half a century. This is perhaps understandable considering the small number of workers in formal human genetics and the greatly increased concern with sporadic cases in recent years in connection with mutation studies. However, a more cogent reason may be found in the development of adequate computing equipment.

Ten years ago, a geneticist might well have been discouraged by the equations of the present paper, requiring many days or weeks of desk calculation for their application. Fortunately, readily available computers have reduced this time by a factor of 100 or more, with incorporation of computing checks that insure accuracy. Once programmed, very little labor is required to tabulate scores for various values of the parameters and to perform the same type of analysis on other data. All of the methods of this paper, with others reported elsewhere (Morton, 1958) or still unpublished, have been programmed for the IBM 650 computer, checked exhaustively, and employed in many analyses. This program, written by Mr. R. A. Hedberg and Mrs. Nancy Jones, may be used by arrangement with the Department of Medical Genetics, University of Wisconsin Medical School.

Two special applications of these methods are of interest. The sometimes complex analysis of concordance in twins may be assimilated by considering each set of twins as a sibship of size 2, and similarly each set of triplets as a sibship of size 3, etc., where p is the concordance. The formulae may be used, with some extension (also programmed for the IBM 650), to distinguish between technical errors, illegitimacy, and disturbed segregation in the blood group systems.

These procedures, in studies to be published shortly, give remarkably good agreement with genetic theory and other sources of information. The apprehension expressed by Kempthorne (1957, p. 195), that segregation data in man are so complicated by family planning and other disturbances as not to be amenable to precise analysis by simple models, appears to be unfounded.

SUMMARY

Methods are developed for analysis of data with arbitrary segregation ratio, ascertainment frequency, and incidence of sporadic cases, with separation of mutations, phenocopies, and incomplete penetrance. Tests of consistency and estimates by maximum likelihood scores are provided for all parameters. Formulae and an example are given. The methods are also applicable to estimation of concordance in twins and natural selection in families.

APPENDIX. DERIVATION OF FORMULAE

1. *Separation of segregating and nonsegregating families. Complete selection, complete penetrance, no sporadic cases.*

The most important matings are possible backcrosses ($p_0 = 1/2$) and diallelic intercrosses ($p_0 = 1/4$). Possible multiple allelic intercrosses may be scored as backcrosses for each parent separately unless the parental phenotypes are identical, but only informative progeny should be scored. Thus if T_1 is dominant to t_2 and both dominant to t, then in matings of type T_1- × t_2- all s children may be scored for the T_1 parent, but only non-T_1 children for the t_2 parent, with

$$h = f_{T_1}^2/(f_{T_1}^2 + 2f_{T_1}f_{t_2} + 2f_{T_1}f_t) = f_{T_1}/[f_{T_1} + 2\,(f_{t_2} + f_t)]$$

in the first case, and $h = 1 - \{2f_t/(f_{t_2} + 2f_t)\}\{f_t/(f_t + f_{t_2})\}$ in the second, since segregation of the t_2- parent can be recognized only if the non-T_1 allele of a heterozygous T_1- parent is t. The same principles apply to analysis of other modes of selection.

In possible backcrosses of size s, the probability of a segregating family is $m = (1 - h)(1 - q^s)$, and $u_p = \frac{\partial\,(\ln m)}{\partial p} = \frac{1}{m}\frac{\partial m}{\partial p} = sq^{s-1}/(1 - q^s)$, similarly $u_h = \frac{1}{m}\frac{\partial m}{\partial h} = -1/(1 - h)$. The probability of a nonsegregating family $= 1 - m = h + (1 - h)q^s$, and $u_p = -(1 - h)\,sq^{s-1}/[h + (1 - h)q^s]$, $u_h = (1 - q^s)/[h + (1 - h)q^s]$. The conditional variances and covariance are

$$k_{pp} = \frac{1}{m}\left(\frac{\partial m}{\partial p}\right)^2 + \frac{1}{1 - m}\left(\frac{\partial(1 - m)}{\partial p}\right)^2$$

$$= (1 - h)s^2q^{2s-2}/(1 - q^s)[h + (1 - h)q^s],$$

$$k_{hh} = \frac{1}{m}\left(\frac{\partial m}{\partial h}\right)^2 + \frac{1}{1 - m}\left(\frac{\partial(1 - m)}{\partial h}\right)^2 = (1 - q^s)/(1 - h)[h + (1 - h)q^s],$$

and

$$k_{hp} = \frac{1}{m}\frac{\partial m}{\partial p}\frac{\partial m}{\partial h} + \frac{1}{1 - m}\frac{\partial(1 - m)}{\partial p}\frac{\partial(1 - m)}{\partial h} = -sq^{s-1}/[h + (1 - h)q^s].$$

In possible intercrosses, the probability of a segregating family is $(1 - h)^2(1 - q^s)$, and $u_p = sq^{s-1}/(1 - q^s)$, $u_h = -2/(1 - h)$. The probability of a nonsegregating family is $1 - (1 - h)^2(1 - q^s)$, and $u_p = -(1 - h)^2sq^{s-1}/\{1 - (1 - h)^2(1 - q^s)\}$, $u_h = 2(1 - h)(1 - q^s)/\{1 - (1 - h)^2(1 - q^s)\}$. The conditional variances and covariance are

$$\begin{aligned} k_{pp} &= (1 - h)^2s^2q^{2s-2}/(1 - q^s)\{1 - (1 - h)^2(1 - q^s)\}, \\ k_{hh} &= 4(1 - q^s)/\{1 - (1 - h)^2(1 - q^s)\}, \\ \text{and} \qquad k_{hp} &= -2(1 - h)sq^{s-1}/\{1 - (1 - h)^2(1 - q^s)\}. \end{aligned}$$

The parents of these sibships and unrelated individuals from the same population contribute other information about h. For the estimate of h from the gene frequencies f_T and f_t, $k_{hh} = (2f_t + f_T)^4/4\{f_t^2 \text{ var } (f_T) + f_T^2 \text{ var } (f_t) - 2f_Tf_t \text{ cov } (f_t, f_T)\}$. When $f_T = 1 - f_t$, this reduces to $k_{hh} = (1 + f_t)^4/4 \text{ var } (f_t)$. The sources of information about h are possible backcrosses, possible intercrosses, the parents, and the population sample, giving three degrees of freedom for testing homogeneity of h. Three cases arise.

Case 1. Homogeneous h, with the value of k_{hh} in the population sample much larger than the sum of the backcross and intercross values. Sampling error in h may be neglected. This is the method for a preliminary analysis, more refined tests being necessary only if there is an apparent deviation from the null hypothesis.

Case 2. Homogeneous h, the sampling error of h not negligible. The values of k_{hh} may be pooled, but the other scores are kept separate and the scores from backcrosses and intercrosses distinguished by 1 and 2 respectively. Then the scores and information matrix evaluated at p_{01}, p_{02}, and h give the required estimates and their variances.

Case 3. Heterogeneous h. This may arise from chance, incorrect gene frequencies, nonrandom mating, or disturbed segregation. These hypotheses can be examined separately by comparison of the parental distribution with the population sample, a contingency test of random association of parental phenotypes, and by tests on the segregating families (§3). If desired, h may be estimated as above on the evidence of the children and parents alone.

2. *Separation of homozygous and heterozygous parents. Complete selection.*

Since segregation is not necessary for recognition of parental heterozygosity, all heterozygous parents are scored. The two important segregation ratios are 1:1 and 1:2:1, the latter being reduced to the former by comparing the two classes of homozygotes, then the pooled homozygotes with the heterozygotes. If there are no sporadic cases ($x = 0$), the distribution of r affected is $\binom{s}{r}p^rq^{s-r}$, and $u_p = r/p - (s - r)/q = r/pq - s/q$. Let $e = s/q$, so that $u_p = r/pq - e$. Clearly e is the expected value of r/pq on the null hypothesis. To obtain k_{pp}, note that

$$E(r/pq)^2 = E\{r(r - 1)\}/p^2q^2 + e/pq.$$

Substituting $s(s - 1)p^2$ for $Er(r - 1)$ we obtain: $s(s - 1)/q^2 + s/pq^2$, so that $k_{pp} = E(r/pq)^2 - e^2 = s/pq$.

This anticipated result required no derivation, but illustrates a method that will be used later for related distributions. The scores, although not needed in the analysis of this type of selection, are convenient for combination of these families with other data. If desired, each family may be partitioned into several items of information, the first corresponding to comparison of segregating and nonsegregating families and obtained as in the last section with $h = 0$. The analysis of segregating families proceeds as for truncate selection.

If there are nongenetic sporadic cases ($x > 0$), the most important matings with this method of ascertainment are possible backcrosses in which the affected parent is a proband and may be a phenocopy. Neglecting the possibility of two phenocopies or of a phenocopy and a genetic case in the same family, the probability of a segregating family is $(1 - x)(1 - q^s)$, and of a nonsegregating family is $x + (1 - x)q^s$. The scores and variances for x and p may be obtained as in the last section by substituting x for h. The analysis of segregating families is given in the next section.

3. *Truncate selection* ($\pi = 1$).

With the type of selection the distribution of r affected ($r > 0$) when $x = 0$ is $\binom{s}{r}p^r q^{s-r}/(1 - q^s)$, and $u_p = r/pq - e$, where $e = s/q(1 - q^s)$. We find $k_{pp} = s(s - 1)/q^2(1 - q^s) + e/pq - e^2 = s(1 - q^s - spq^{s-1})/pq(1 - q^s)^2$. Values of k have been tabulated by Finney (1949), who used the symbol W. However his "bias" B is not the same as our e, being equal to $e - kp$.

These values of u and k give an omnibus test of the null hypothesis that $x = 0$, $\pi = 1$, and $p = p_0$. More specific tests may be obtained from the separation of simplex and multiplex families and the distribution of r within multiplex families. The scores u_p and u_x and their conditional variances and covariance may be found in §5 for $\pi = 1$.

4. *Single selection* ($\pi \longrightarrow 0$).

The probability of selection of a family with r affected is $\lim\limits_{\pi \to 0} \{1 - (1 - \pi)^r\} = r\pi$, and when $x = 0$ the probability of selection of a family of size s is $\Sigma\binom{s}{r}\pi r p^r q^{s-r} = sp\,\pi = \lim\limits_{\pi \to 0} \{1 - (1 - p\pi)^s\}$. Therefore the distribution of r affected in families of size s is $r\pi\binom{s}{r}p^r q^{s-r}/sp\pi = \binom{s-1}{r-1}p^{r-1}q^{s-r}$, and single selection is equivalent to complete selection of the siblings of the index case.

With x arbitrary, the siblings give a family test of the sporadic or nonsporadic origin of the index case, and the frequency of simplex families is $x + (1 - x)q^{s-1}$. The scores u_x and u_p with their variances k_{xx} and k_{pp} and the covariance k_{xp} may be obtained as for possible backcrosses in §1 by substituting x for h and $s - 1$ for s. In multiplex families, the distribution of $r - 1$ affected among the $s - 1$ siblings of the index case is scored as for truncate selection ($r - 1 > 0$).

5. *Multiple selection* ($0 < \pi \leqq 1$).

When $x = 0$, the distribution of r affected is

$$\binom{s}{r}p^r q^{s-r}\{1 - (1 - \pi)^r\}/\{1 - (1 - p\pi)^s\},$$

since $\Sigma\binom{s}{r}p^r q^{s-r}(1 - \pi)^r = (1 - p\pi)^s$.

With x arbitrary, the frequency of simplex families is

$$m_1 = C\pi\{w + (1 - w)spq^{s-1}/(1 - q^s)\},$$

and the frequency of multiplex families is

$$m_2 = C(1 - w)\{1 - (1 - p\pi)^s - s\pi pq^{s-1}\}/(1 - q^s),$$

where C is a constant such that $m_1 + m_2 = 1$. We find

$$C = 1/\{w\pi + (1 - w)[1 - (1 - p\pi)^s]/(1 - q^s)\}.$$

Substituting for w,

$$m_1 = \frac{sp\pi\{x + (1 - x)q^{s-1}\}}{xsp\pi + (1 - x)\{1 - (1 - p\pi)^s\}} = sp\pi A/B,$$

say, and the scores for simplex families are

$$u_x = (BW - AY)/AB, \qquad u_\pi = (B - sp\pi Z)/\pi B,$$

and $u_p = (BX - sp\pi AZ)/pAB$, where

$$W = 1 - q^{s-1} \qquad Y = sp\pi - 1 + (1 - p\pi)^s$$

$$X = x + (1 - x)q^{s-1} - p(1 - x)(s - 1)q^{s-2} \qquad Z = x + (1 - x)(1 - p\pi)^{s-1}.$$

Similarly,

$$m_2 = \frac{(1 - x)\{1 - (1 - p\pi)^s - \pi spq^{s-1}\}}{xsp\pi + (1 - x)\{1 - (1 - p\pi)^s\}} = (1 - x)D/B,$$

say, and the scores for multiplex families are $u_x = -\{(1 - x)Y + B\}/(1 - x)B$, $u_\pi = sp(BJ - DZ)/BD$, and $u_p = s\pi(BK - DZ)/BD$, where

$$J = (1 - p\pi)^{s-1} - q^{s-1} \qquad K = J + p(s - 1)q^{s-2}.$$

The conditional variances and covariances are $k_{xx} = \Sigma mu_x^2$, etc.

In multiplex families the probability of r affected is $m_r = \binom{s}{r}p^rq^{s-r}\{1 - (1 - \pi)^r\}/\{1 - (1 - p\pi)^s - \pi spq^{s-1}\}$, with scores $u_p = r/pq - e_p$, $u_\pi = r(1 - \pi)^{r-1}/\{1 - (1 - \pi)^r\} - e_\pi$, where $e_p = s(D + q\pi K)/qD$ and $e_\pi = spJ/D$.

$$k_{pp} = s(s - 1)\{1 - (1 - \pi)^2(1 - p\pi)^{s-2}\}/q^2D + e_p/pq - e_p^2$$

$$k_{\pi\pi} = \sum_{r=2}^{s} m_r u_\pi^2$$

$$k_{\pi p} = s(s - 1)p(1 - \pi)(1 - p\pi)^{s-2}/qD + e_\pi/pq - e_pe_\pi$$

Among r affected, the distribution of *a* probands $(a > 0)$ is $\binom{r}{a}\pi^a(1 - \pi)^{r-a}/\{1 - (1 - \pi)^r\}$, which corresponds to truncate selection of probands among affected sibs.

6. *Multiple selection* $(0 < \pi \leq 1)$ *with at least one affected girl*

If a rare recessive trait is a mixture of autosomal and sex-linked cases, families with autosomal or sporadic cases will be recognized if they contain at least one affected girl. When $x = 0$, the distribution of r affected under this condition is

$$\binom{s}{r}p^r q^{s-r}\{1-(1-\pi)^r\}\{1-(1/2)^r\}/ \\ \{1-(1-p\pi)^s-(p/2+q)^s+(p/2+q-p\pi/2)^s\}.$$

If x is the frequency of sporadic cases among affected girls, the frequency of simplex families is $m_1 = C(\pi/2)\{w + (1-w)spq^{s-1}/(1-q^s)\}$, and the frequency of multiplex families is

$$m_2 = C(1-w)\{1-(1-p\pi)^s-(p/2+q)^s+ \\ (p/2+q-p\pi/2)^s-\pi spq^{s-1}/2\}/(1-q^s),$$

where

$$C = 1/\{w\pi/2+(1-w)\,[1-(1-p\pi)^s- \\ (p/2+q)^s+(p/2+q-p\pi/2)^s]/\,(1-q^s)\}.$$

Substituting for w,

$$m_1 = \frac{sp\pi\{x+(1-x)q^{s-1}\}}{xsp\pi+2(1-x)\{1-(1-p\pi)^s-(p/2+q)^s+(p/2+q-p\pi/2)^s\}} = sp\pi A/B,$$

say, and the scores for simplex families are

$$u_x = (BW - AY)/AB, \qquad u_\pi = (B - sp\pi Z)/\pi B,$$

and $u_p = (BX - spAV)/pAB$, where

$$V = x\pi + 2(1-x) \\ \{\pi(1-p\pi)^{s-1}+(p/2+q)^{s-1}/2-(1+\pi)(p/2+q-p\pi/2)^{s-1}/2\}$$

$$W = 1 - q^{s-1}$$

$$X = x + (1-x)q^{s-1} - p(1-x)(s-1)q^{s-2}$$

$$Y = sp\pi - 2\{1-(1-p\pi)^s-(p/2+q)^s+(p/2+q-p\pi/2)^s\}$$

$$Z = x + 2(1-x)\{(1-p\pi)^{s-1}-(p/2+q-p\pi/2)^{s-1}/2\}.$$

Similarly,

$$m_2 = \frac{2(1-x)\{1-(1-p\pi)^s-(p/2+q)^s+(p/2+q-p\pi/2)^s-\pi spq^{s-1}/2\}}{xsp\pi+2(1-x)\{1-(1-p\pi)^s-(p/2+q)^s+(p/2+q-p\pi/2)^s\}} \\ = 2(1-x)D/B,$$

say, and the scores for multiplex families are $u_x = -\{(1-x)Y + B\}/(1-x)B$, $u_\pi = sp(BJ - DZ)/BD$, and $u_p = s(BK - DV)/BD$, where

$$J = (1-p\pi)^{s-1}-(p/2+q-p\pi/2)^{s-1}/2-q^{s-1}/2$$

$$K = \pi(1-p\pi)^{s-1}+(p/2+q)^{s-1}/2-(1+\pi)(p/2+q-p\pi/2)^{s-1}/2-\pi q^{s-1}/2 \\ +(s-1)\pi pq^{s-2}/2$$

The conditional variances and covariances are $k_{xx} = \Sigma u_x^2$, etc.

In multiplex families the probability of r affected is

$$m_r = \binom{s}{r} p^r q^{s-r} \{1 - (1 - \pi)^r\} \{1 - (1/2)^r\}/D,$$

with scores $u_p = r/pq - e_p$,

$$u_\pi = r(1 - \pi)^{r-1}/\{1 - (1 - \pi)^r\} - e_\pi ,$$

where $e_p = s(D + qK)/qD$ and $e_\pi = spJ/D$.

$$\begin{aligned} k_{pp} = {} & s(s - 1)\{1 - (1 - \pi)^2(1 - p\pi)^{s-2} - (p/2 + q)^{s-2}/4 \\ & + (1 - \pi)^2(p/2 + q - p\pi/2)^{s-2}/4\}/q^2D + e_p/pq - e_p^2 \end{aligned}$$

$$k_{\pi\pi} = \Sigma m_r u_\pi^2$$

$$k_{\pi p} = s(s - 1)p(1 - \pi)\left\{(1 - p\pi)^{s-2} - \left(\frac{p}{2} + q - \frac{p\pi}{2}\right)^{s-2} \Big/ 4\right\} \Big/ qD + e_\pi/pq - e_\pi^2$$

As in §5, the distribution of *a* probands ($a > 0$) among r affected corresponds to truncate selection of probands among affected sibs.

If sex-linked and autosomal cases can be distinguished phenotypically, the distribution of r in families of the autosomal phenotype with no affected girl is the same as in §5, letting $p_0 = 1/7$ and defining x as the frequency of sporadic cases among affected boys. With girls excluded, $p_0 = 1/4$ on the same conditions.

7. *Estimation of π from the number of ascertainments*

Sometimes an investigator reports the number of times a family is ascertained instead of the number of probands. This extracts information from isolated cases, but requires for an estimate of π the assumption that ascertainments are independent.

Let there be t ascertainments of a family with r affected ($r > 0$), and let m be the mean number of ascertainments per affected individual, so that $\pi = 1 - e^{-m}$. The distribution of $t > 0$ is

$$P(t) = \frac{(mr)^t e^{-mr}}{t!(1 - e^{-mr})} = \frac{[-r \ln (1 - \pi)]^t (1 - \pi)^r}{t![1 - (1 - \pi)^r]} .$$

Then

$$u_\pi = -t/(1 - \pi) \ln (1 - \pi) - e_\pi$$

$$e_\pi = r/(1 - \pi)\{1 - (1 - \pi)^r\}$$

$$k_{\pi\pi} = \frac{r\{[1 - r \ln (1 - \pi)](1 - \pi)^r - 1\}}{(1 - \pi)^2[1 - (1 - \pi)^r]^2 \ln (1 - \pi)} .$$

The expected number of probands in a family with r affected is

$$a^* = r\pi/[1 - (1 - \pi)^r] = \pi(1 - \pi)e_\pi$$

When the probands are not designated, this method is open to objection because of uncertainty of the assumption that ascertainments are independent. Human geneticists have not hitherto realized that the best analysis is possible only if the probands are designated, and in addition, ***the number of ascertainments of each proband is recorded*** and analysed by the method of this section for the case $r = 1$. These two estimates of π provide a test of the assumption that ascertainments are independent; this granted, the pooled estimate is more precise than probands or ascertainments alone could give.

It has been assumed above that ascertainment is sufficient to bring an individual into the record. This will not be true if some ascertained cases refuse to release their records or cooperate in other essential ways. However, the method is easily modified to adjust for this. Let there be N persons with at least one ascertainment, of whom n cooperate in the study. We agree to consider as a proband only patients who cooperate. Then if π' is the unadjusted ascertainment probability based on the distribution of t, and π the adjusted value,

$$\pi = n\pi'/N$$

$$K_{\pi\pi} = 1\Big/\left\{\frac{n^2}{N^2K_{\pi'\pi'}} + \frac{(N-n)\pi^2}{Nn}\right\}.$$

REFERENCES

Bailey, N. T. J. 1951. A classification of methods of ascertainment and analysis in estimating the frequencies of recessives in man. *Ann. Eugen.* 16: 223–225.

Finney, D. J. 1949. The truncated binomial distribution. *Ann. Eugen.* 14: 319–328.

Haldane, J. B. S. 1938. The estimation of the frequencies of recessive conditions in man. *Ann. Eugen.* 8: 255–262.

Haldane, J. B. S. 1949. A test for homogeneity of records of familial abnormalities. *Ann. Eugen.* 14: 339–341.

Ikin, Elizabeth, Eileen Prior, R. R. Race, and G. L. Taylor. 1939. The distribution of the A_1A_2BO blood groups in England. *Ann. Eugen.* 9: 409–411.

Kempthorne, O. 1957. *An introduction to genetic statistics.* John Wiley and Sons.

Morton, N. E. 1958. Segregation analysis in human genetics. *Science* 127: 79–80.

Race, R. R., Elizabeth Ikin, G. L. Taylor, and Eileen Prior. 1942. A second series of families examined in England for the A_1A_2BO and MN blood-group factors. *Ann. Eugen.* 11: 385–394.

Rao, C. R. 1952. *Advanced statistical methods in biometric research.* John Wiley and Sons.

Taylor, G. L. and Eileen Prior. 1938. Blood groups in England. III. Discussion of the family material. *Ann. Eugen.* 9: 18–44.