Reports

Segregation Analysis in Human Genetics

Thirty years ago, the infant science of human genetics was largely concerned with accumulating examples of regular Mendelian inheritance. By the methods of that period, devised by Weinberg, Bernstein, and others, it was not difficult to recognize traits that approximated monogenic inheritance, at least in some pedigrees. However, as larger and more representative samples were collected, many of these conditions appeared to deviate from simple Mendelian models in ways that suggested sporadic cases of mutational or nongenetic origin, mixture of genetic entities, differential viability, incomplete penetrance, or irregularities in the sampling ("ascertainment") procedure. Human geneticists have become increasingly interested in these discrepancies as critical evidence on the inheritance, gene frequencies, mutation rates, genetic risks, selection pressures, and homogeneity of genetic traits.

As soon as deviations from the simplest models were encountered, the traditional statistical methods of human genetics became inadequate. In some cases these methods have been found to be unreliable in small samples (*1*). More generally, they fail to provide tests of internal consistency or to distinguish among alternative hypotheses which in principle are quite distinct, such as partial sex-linkage and sex-biased manifestation, or mutations, phenocopies, and incomplete penetrance. Further progress in many aspects of human genetics must await the application, not only of better laboratory techniques, but also of a more refined methodology, made possible by modern developments in mathematical statistics.

Segregation and recombination, the basic data of formal genetics, pose contrasting analytical problems. (Segregation relates to the distribution of progeny with respect to a single genetic locus, and recombination, to the element introduced when two loci are considered simultaneously.) The alternative to 50 percent recombination is linkage with some smaller recombination value, whose a priori distribution is in principle specifiable and can in fact be approximated. Usually no other parameter need be estimated, nor the ascertainment of the main locus specified. Here is a problem differing from that usually encountered in scientific research in being amenable to the acceptance sampling theory of Neyman and Wald, which assumes a formal alternative to the null hypothesis. These small-sample methods, adapted to sequential analysis, are of all linkage tests in man the most precise, powerful, and utilitarian (*1, 2*).

Segregation, on the other hand, opposes a variety of alternatives to the null hypothesis that specifies the proportion affected p (for example, $p = 1/8$ for an intercross with 50-percent penetrance of the recessive homozygote). The true proportion may be larger or smaller than this, the ascertainment model may be incorrect, or segregation may be confused with sporadic cases of different etiology, and to these alternatives no a priori probabilities can reasonably be assigned. To such problems, characteristic of scientific research, acceptance sampling theory provides no solution. Fortunately, human families give much more information about segregation than about recombination, so that large-sample theory is appropriate. The maximum likelihood theory of Fisher is the method of choice, since to its excellent theoretical credentials may be added the advantage of analysis by simple addition of iterative scores. This application was initiated by Finney (*3*) and has now been extended to a number of common genetic situations. It is the purpose of the present report (*4*) to indicate the logical structure of this extension so that practitioners of human genetics may choose methods of analysis appropriate to their data. A full exposition, with derivation of formulae and tables of scores, is in preparation.

The expected proportion affected p (assumed constant among families of a particular mating type) is specified by Mendelian theory if penetrance is complete. With incomplete penetrance, a simple analysis must be restricted to matings with the same theoretical proportion, which is multiplied by the penetrance to obtain the effective proportion p. For example, if onset is delayed, the penetrance will lie between

$$\int f(z)\,G(z)\,dz$$

and

$$\int f_1(z)\,G(z)\,dz$$

where $f(z)$ is the frequency of age z at death or last examination among affected individuals and their normal sibs, $f_1(z)$ is this frequency with the first proband in each family excluded, and $G(z)$ is the cumulative frequency of onset at age z among affected cases. Any treatment of incomplete penetrance must be approximate if there is interfamily heterogeneity in penetrance or age of onset, but maximum likelihood scores provide empirical standard errors against which genetic hypotheses may be tested.

Incomplete ascertainment is one of the most troublesome problems in human genetics and requires a special vocabulary. A proband is an affected person ascertained independently of the other affected members of the family. The ascertainment probability π (assumed to be independent of the number of affected sibs, the severity of their affection, and the number of probands among sibs) is the probability that an affected person be a proband. The ascertainment probability approaches zero if very few affected family members are probands (*single selection*) and unity if there is complete selection of affected children (*truncate selection*). In the general case of multiple selection ($0 < \pi \leq 1$), there will be a probands in a sibship with r affected, with probability

$$P_{(a)} = \frac{\binom{r}{a} \pi^a (1-\pi)^{r-a}}{1-(1-\pi)^r}$$

where $1 \leq a \leq r$. Each proband may be independently ascertained t times

$$P_{(t)} = \frac{(1-\pi)[-\ln(1-\pi)]^t}{t!\pi}$$

with $t \geq 1$. Comparison of scores for π from $P_{(a)}$ and $P_{(t)}$ provides a test of the hypothesis that sources of ascertainment are independent. This granted, the scores may be pooled for comparison with a third sum of scores, obtained from the distribution of r among families. The more information about π that can be derived in these ways, the more precise the tests on other parameters will be. Information from the distribution of t has apparently never before been used.

At this point it is necessary, for clar-

All technical papers and comments on them are published in this section. Manuscripts should be typed double-spaced and be submitted in duplicate. In length, they should be limited to the equivalent of 1200 words; this includes the space occupied by illustrative or tabular material, references and notes, and the author(s)' name(s) and affiliation(s). Illustrative material should be limited to one table *or* one figure. All explanatory notes, including acknowledgments and authorization for publication, and literature references are to be numbered consecutively, keyed into the text proper, and placed at the end of the article under the heading "References and Notes." For fuller details see "Suggestions to Contributors" in *Science* 125, 16 (4 Jan. 1957).

ity, to distinguish between *isolated* and *sporadic* cases. Isolated cases may be of the same origin as familial cases, and the absence of affected sibs due to chance, or they may be sporadic, of different origin (mutation, phenocopy, and so forth). To distinguish between incomplete penetrance and sporadic cases, we may recognize a quantity x, the frequency of sporadic cases among affected cases in the population, and test the hypothesis that no isolated case is sporadic ($x = 0$). For example, if a trait is sometimes due to a rare dominant gene, the probability that an affected parent with s children have at least one affected child is

$$P_{(r>0)} = (1-x)(1-q^s)$$

where $q = 1 - p$, and the scores for x give a specific, and the most efficient, test for the existence of sporadic cases not transmissible to the progeny.

If families are ascertained through parents or other relatives without regard to the children, and if $x = 0$, the probability of a segregating family ($r > 0$) is

$$P_{(r>0)} = (1-h)(1-q^s)$$

for possible backcrosses, and

$$P_{(r>0)} = (1-h)^2(1-q^s)$$

for possible intercrosses, where h, the probability that a parent of dominant phenotype be homozygous, is determined under random mating by the gene frequencies. Using this theoretical value of h, we may compare the scores for p from segregating and nonsegregating families with the scores from the distribution of r in segregating families as a test of the hypotheses that $x = 0$ and that mating is at random with the specified gene frequencies.

If $x\mu \ll (1-x)p$, where μ is the proportion affected among sibs of sporadic cases, then familial cases of sporadic origin may be neglected, and the probabilities for isolated and familial cases, respectively, are

$$P_{(r=1|r>0)} = \frac{sp\pi[(x+(1-x)q^{s-1}]}{xsp\pi+(1-x)[1-(1-p\pi)^s]}$$

and

$$P_{(r|r>1)} = \frac{\binom{s}{r}p^r q^{s-r}[1-(1-\pi)^r]}{1-(1-p\pi)^s-\pi spq^{s-1}}$$

The distribution of r for familial cases gives a third sum of scores for π and a test of the null hypothesis about p, independent of x, while the distribution of isolated and familial cases tests the hypothesis that $x = 0$, or permits an estimate if this hypothesis is rejected.

Sex-linked genes present interesting problems, such as that of distinguishing the mutation rates in the two sexes. However, for a decisive analysis, more penetrating methods are required than have so far been applied to this question. For a deleterious sex-linked recessive trait, the probability that an affected male be sporadic is

$$x = mu/(2u+v)$$

where m is the coefficient of selection against affected males, u and v are the mutation rates in egg and sperm, respectively, and it is assumed that carrier females have normal fertility and that all cases are sex-linked. The test of the hypothesis that $u = v$ reduces to testing whether $x = m/3$.

Corroboration of this test may be sought in the distribution of affected maternal uncles, assuming ascertainment through nephews. If two or more nephews are affected, the probability that at least one of s maternal uncles be affected is

$$P_{(r>0)} = (1-x')(1-q^s)$$

where the expected value of x' is ½ on the hypothesis of sex linkage. Similarly, if only one nephew is affected in a sibship of n nephews, the probability that at least one maternal uncle be affected, if ascertainment is through the nephew, is

$$P_{(r>0)} = \frac{(1-x)(1-x')q^{n-1}(1-q^s)}{x+(1-x)q^{n-1}}$$

These distributions provide a test of homogeneity of x and of x', of the deviation of x' from ½, and of x from $m/3$, the last deviation being accepted as significant only if the others are nonsignificant.

These ancillary tests are particularly important if sex-linked and autosomal cases may be confused, in which event families with autosomal or sporadic cases may be recognized if they contain at least one affected girl. The probabilities for isolated and familial cases under this condition are given in Fig. 1.

Genetic tests in man have never been carried out with the precision of these methods, the practical limitations of which cannot therefore be specified. Considerable caution should be exercised in analysis of medical literature or other biased sources, or if correct diagnosis is more likely in familial cases. However, it is still possible that the distribution of r for familial cases may be adequately described by the formulae of this report.

$$P_{(r=1|r>0)} = \frac{sp\pi[x+(1-x)q^{s-1}]}{xsp\pi+2(1-x)[1-(1-p\pi)^s-(\frac{1}{2}p+q)^s+(\frac{1}{2}p+q-\frac{1}{2}p\pi)^s]}$$

$$P_{(r|r>1)} = \frac{\binom{s}{r}p^r q^{s-r}[1-(1-\pi)^r][1-(\frac{1}{2})^r]}{1-(1-p\pi)^s-(\frac{1}{2}p+q)^s+(\frac{1}{2}p+q-\frac{1}{2}p\pi)^s-(\pi spq^{s-1})/2}$$

Fig. 1. Probabilities for different numbers of affected children under multiple selection with at least one affected girl.

With careful enumeration of probands and ascertainments when π is less than 1, these methods should be sufficiently general for almost any Mendelian analysis, provided that the mating types to be analyzed do not contain a mixture of segregation frequencies. (Common genes with low penetrance will continue to defy precise analysis.) For the typical case of a rare or highly penetrant gene, the method of maximum likelihood scores for p, x, h, and π, applied to the distributions of this report, will provide a simple and more powerful analysis than has been previously available.

NEWTON E. MORTON
Department of Medical Genetics, University of Wisconsin, Madison

References and Notes

1. N. E. Morton, *Am. J. Human Genet.* 9, 55 (1957).
2. ———, *ibid.* 7, 277 (1955).
3. D. J. Finney, *Ann. Eugenics* 14, 319 (1949).
4. Genetics paper No. 663. This work was supported by a grant from the Rockefeller Foundation.

31 July 1957

Structure of Adenine Polynucleotide

The x-ray diffraction pattern of untreated individual fibers of enzymatically synthesized adenine polynucleotide has been presented and discussed by Watson (*1*). A more crystalline preparation of the polymer has been obtained from material synthesized by Roland F. Beers, Jr., (*2*), spun into a bundle of fine fibers, and stretched while bathed in a heated mixture of ethanol and dichloroacetic acid (DCA) (*3*). The distribution of intensity on the x-ray diagram of these fibers is similar to that given by untreated fibers, showing that the form of the molecule is not greatly altered; differences of detail are observed, however, chiefly on the equator and on the first layer line. The 13 reflections here observed indicate a tetragonal unit cell, 22.6 by 22.6 by 15.0 A. A fourfold screw axis is evident parallel to the fiber axis, with translation of 3.75 A. The observed density of these fiber bundles (~1.60) suggests 16 nucleotide residues and 16 molecules of DCA per unit cell (calculated density, 1.66).