Estimation of Prevalence under Incomplete Selection*

I. BARRAI,[1] M. P. MI,[2] N. E. MORTON,[2] AND N. YASUDA[2]

[1]*Istituto di Genetica, Università di Pavia,*
Italy.
[2]*Genetics Department, University of Hawaii, Honolulu.*

FOR A COMMON TRAIT it is feasible and clearly desirable to estimate prevalence from a random sample of the general population (Lilienfeld, 1962). However, this is prohibitively expensive for the rare traits with which a geneticist often has to deal, since reliable estimates would require careful examination of a large fraction of the population (e.g., more than 100,000 individuals for a trait with a prevalence of 1/10,000). Such rare traits are ordinarily ascertained through pedigrees which contain one or more *probands* (selected through hospital records, death certificates, inquiries to physicians, examination of a population sample, or other direct means of ascertainment) and in addition may contain *secondary cases* not represented in these primary sources and detected only through a family study of probands. For this *incomplete selection,* several indirect estimates of prevalence are available, depending on the mode of inheritance and method of ascertainment.

THE FREQUENCY OF A RARE RECESSIVE GENE

In the general population, the inbreeding coefficient f_i has frequency c_i, mean $\alpha = \sum_i c_i f_i$, and variance $\sigma^2 = \Sigma\, c_i f_i^2 - \alpha^2$. The incidence of homozygotes at any stage preceding natural selection is $q[q + (1 - q)\alpha]$, where q is the recessive gene frequency. Among probands, the mean value of the inbreeding coefficient is

$$F = \frac{\sum_i c_i f_i\,(q + pf_i)}{\sum c_i\,(q + pf_i)} = \frac{q\alpha + (1 - q)\,(\sigma^2 + \alpha^2)}{q + (1 - q)\alpha}$$

Solving for q, we obtain

$$q = \frac{\sigma^2 + \alpha^2 - F\alpha}{F\,(1 - \alpha) - \alpha + \sigma^2 + \alpha^2} \qquad (1)$$

which reduces, when $\sigma^2 + \alpha^2 - F\alpha << F - \alpha$, to

Received November 23, 1964.

Supported initially at the University of Wisconsin by U. S. Atomic Energy Commission research contract and U. S. Public Health Service training grant from the National Institutes of Health and concluded at the University of Hawaii with the support of U. S. Public Health Service research grant GM 10424 from the National Institutes of Health and by National Science Foundation research grant G23190.

*Genetics Paper 2.51.

$$q \cong \frac{\sigma^2}{F - \alpha} - \alpha \tag{2}$$

with a standard error estimated (neglecting errors in α and σ^2) by

$$\sigma_q \cong \left| \frac{dq}{dF} \right| \sigma_F = \frac{\sigma^2 \sigma_F}{(F - \alpha)^2} = \frac{(q + \alpha)^2 \sigma_F}{\sigma^2}$$

If there are sporadic, nonrecessive cases in addition to the recessive ones, it is only necessary to exclude isolated cases in computing F. Incomplete penetrance and incomplete or irregular ascertainment are irrelevant if they are independent of the inbreeding coefficient, but it is an essential assumption that only one autosomal locus is involved and that the trait is recessive.

These results may be considered a generalization of Dahlberg's formula (1948) to include consanguineous marriages other than first cousins and a special case of Morton's formula (Chung, Robison, and Morton, 1959) when there is only a single locus. Kimura (1958) gave an iterative maximum likelihood solution which he considered "to be too complicated for practical use" and suggested a simpler but inefficient formula.

In Dahlberg's case there are assumed to be only two values of f in the population: $f = 0$ with frequency $1 - c$ and $f = 1/16$ with frequency c. Then $\alpha = c/16$ and $\sigma^2 = c\ (1 - c)/256$. If k is the proportion of probands whose parents are first cousins, $F = k/16$. Substituting in equation (1) we obtain Dahlberg's formula

$$q = \frac{c(1 - k)}{16k - 15c - ck}$$

while using equation (2) we obtain the Dahlberg approximation

$$q \cong \frac{c\,(1 - k)}{16\,(k - c)}$$

which works very well for $k >> c$.

Nei (1963) based his formula on the frequency of related parents instead of the mean inbreeding coefficient. He gave

$$q = \frac{f\,(1 - K)}{f\,(1 - K) + K - C}$$

where $f = \alpha =$ the mean inbreeding coefficient in the population, $K =$ the proportion of probands whose parents are related, $C =$ the proportion of consanguineous marriages in the general population from which the parents of probands are drawn.

This reduces to Dahlberg's formula when consanguinity other than first cousins is neglected but is inefficient if there are other types of consanguinity, since it does not distinguish between large and small values of the inbreeding coefficient (for example, between $f = 1/4$ and $f = 1/128$). It is also highly sensitive to subjective factors in the ascertainment of remote consanguinity, which different observers may report accurately or as "remote" or "no" consanguinity. This makes little contribution to F or α but may contribute heavily and erratically to the proportion of consanguinity.

The general treatment of the problem admits any distribution of consanguinity, more than one locus, and an admixture of nonrecessive cases (Chung, Robison, and Morton, 1959). We write

$$F = \frac{\sum\limits_i c_i f_i \, (A + Bf_i)}{\sum\limits_i c_i \, (A + Bf_i)} = \frac{A\alpha + B\,(\sigma^2 + \alpha^2)}{A + B\alpha}$$

Morton (1960) gave the maximum likelihood solution. In the general case, we must know the prevalence to estimate A and B, but for a single locus, $A = q^2$ and $B = q\,(1 - q)$, so we obtain equation (1).

Although our equations are less sensitive to subjective factors in recording consanguinity than Nei's results, they are not fully efficient. Neglecting (with Kimura, 1958) errors of estimate in the c_i, the maximum likelihood solution is not difficult. The likelihood of n_i probands with f_i is

$$L = \prod_i c_i^{n_i} \left\{ \frac{q + (1-q)f_i}{q + (1-q)\alpha} \right\}^{n_i}$$

and

$$\frac{d \ln L}{dq} = \frac{1}{q + (1-q)\alpha} \sum_i n_i \left[\frac{\alpha - f_i}{q + (1-q)f_i} \right] \equiv U$$

$$\sum \left(\frac{d \ln L}{dq} \right)^2 = \left(\frac{1}{q + (1-q)\alpha} \right)^2 \sum_i n_i \left[\frac{\alpha - f_i}{q + (1-q)f_i} \right]^2 \equiv K$$

These yield the iteration

$$q^* = q + U/K \qquad (3)$$

and the standard error of the final estimate is $\sigma_q = \sqrt{1/K}$. The frequency $q^2 + q\,(1 - q)\alpha$ has a standard error of $[\alpha + 2q(1 - \alpha)]\sigma_q$.

It is of some incidental interest that Smith's counting method (Smith, 1957) fails to converge in this problem.

Numerical Example

Kimura (1958) and Nei (1963) both used Furusho's data (1957) on deaf mutism. Since 1959, a substantial body of evidence has shown that there are sporadic and dominant cases and many loci controlling recessive deaf mutism (Chung, Robison, and Morton, 1959; Sank, 1963). For comparability we have analyzed these data in Table 1 as if there were only a single recessive gene. Equation (2) gives $q \cong .00988$, compared with .00877 by Nei's formula and .00761 by Kimura's. The exact maximum likelihood estimate is .00910, with a

TABLE 1. ANALYSIS OF FURUSHO'S DATA (1957) ON DEAF-MUTISM AS IF THERE WERE ONLY A SINGLE LOCUS AND ALL CASES RECESSIVE

Parental relationship (after Kimura)	f_i	c_i	n_i	$(\alpha - f_i)/\{q + (1-q)f_i\}$ $q = .00910$
Unrelated	0	1039	879	.504396
First cousins	1/16	77	484	−.815275
1½ cousins	1/32	6	26	−.665408
Second cousins	1/64	13	68	−.448891
2½ cousins	1/128	2	5	−.191344
Total		1137	1462	

FORMULA 2

$$\alpha = 77(1/16) + \ldots. + 2(1/128)/1137 = .00459$$

$$\sigma^2 = \frac{77(1/16)^2 + \ldots. + 2(1/128)^2 - 1137\alpha^2}{1136} = .000252$$

$$F = \{484(1/16) + \ldots. + 5(1/128)\}/1462 = .02200$$

$$\sigma_F^2 = \frac{484(1/16)^2 + \ldots. + 5(1/128)^2 - 1462F^2}{(1462)(1461)} = .57365 \times 10^{-6}$$

$$q \cong \frac{.000252}{.02200 - .00459} - .00459 = .00988$$

$$\sigma_q \cong \frac{(.000252)(.0007574)}{(.02200 - .00459)^2} = .000630$$

$$\cong \frac{(.00910 + .00459)^2(.0007574)}{.000252} = .000563$$

FORMULA 3

Maximum likelihood solution

$q = .00910 \qquad q + (1 - q)\alpha = .013648$

$U = \{879(.504396) + \ldots\}/.013648 = -.801$

$K = \{879(.504396)^2 + \ldots\}/(.013648)^2 = 3064028$

$q = .0091 - .801/3064028 = .00910$

$\sigma_q = \sqrt{1/K} = .000571$

standard error of .000571. Kimura's estimate, though farthest from the maximum likelihood solution, appears to have the smallest standard error, but this is presumably deceptive since the formulae for the standard errors are valid only in the limit for large samples and are very approximate in samples of the size commonly observed. This may be illustrated by the standard error for our equation (2), which is calculated to be .000630 by the formula $\sigma_q = [\sigma^2/(F - \alpha)^2]\sigma_F$ and .000563, using the maximum likelihood estimate of q in the asymptotically equivalent expression $\sigma_q = [(q + \alpha)^2/\sigma^2]\sigma_F$. This instability of the variance discourages comparison of the efficiency of the different methods, which may be expected to have high efficiency and to give almost identical results when consanguinity other than first cousins is nearly negligible. However, when remote consanguinity is relatively important, the maximum likelihood equation (3) should be used.

The frequency in a randomly mating population is estimated as $q^2 = 8.3 \times 10^{-5}$, which is less than half the frequency of recessive deaf mutism. For proof that the discrepancy is due to multiple loci, see Chung, Robison, and Morton (1959) and Sank (1963).

PREVALENCE UNDER INCOMPLETE SELECTION, WITH ISOLATED CASES INCLUDED

Prevalence (n) will be defined as the number of cases of a trait existing in a given area at a given time. From an estimate of the prevalence and the population size N we can determine the frequency at birth of individuals who will develop the trait and from this, if the genetic basis is simple, proceed to a calculation of gene frequency and mutation rate.

Assume that the method of sampling is to collect a fixed number A of probands and then study their relatives. Since the population is finite, the chance of detecting a family with r affected is

$$1 - \frac{\binom{A}{0}\binom{n-A}{r}}{\binom{n}{r}} = 1 - \prod_{i=1}^{r}\left\{1 - \frac{A}{n-i+1}\right\}$$

If $r << n$, the effect of sampling without replacement is negligible and the chance of detecting the family approaches $1 - (1 - \pi)^r$, where $\pi = A/n$ is the ascertainment probability. Since A and n are both fixed numbers, the error of estimate for $n = A/\pi$ is

$$\sigma_n = A\sigma_\pi/\pi^2 \qquad (4)$$

where π is estimated by segregation analysis (Morton, 1959, 1962). This error does not include the effects of drift on gene frequencies nor of accidents of segregation and fertility on genotype frequencies.

Numerical Example

Morton and Chung (1959) reported 26 living probands with limb-girdle muscular dystrophy among 3,700,000 residents in Wisconsin. They estimated the ascertainment probability as .354 with standard error .0534. Therefore the number of cases of limb-girdle muscular dystrophy living in Wisconsin was estimated to be

$$n = A/\pi = 26/.354 = 73.4$$

with standard error

$$\sigma_n = A\sigma_\pi/\pi^2 = 26(.0534)/(.354)^2 = 11.1$$

The frequency is $n/N = 73.4/3700000 = 2.0 \times 10^{-5}$. From the age distribution of the general population and the ages of onset and death of affected persons, they calculated that only .304 of cases born had expressed the trait and were still living, so that the incidence at birth of persons who will develop limb-girdle muscular dystrophy is

$$I = n/.304N = 6.5 \times 10^{-5}$$

with standard error

$$\sigma_I = \sigma_n/.304N = 1.0 \times 10^{-5}$$

PREVALENCE UNDER INCOMPLETE SELECTION, WITH EXCLUSION OF ISOLATED CASES

If the trait is a mixture of sporadic and high-risk cases, data on the latter component may be collected by omitting isolated cases. The resulting sample of *multiplex* families (i.e., with two or more affected, $r > 1$) still permits estimation of prevalence. If A^* is the number of probands in multiplex families, the number of affected in multiplex families in the population is estimated by A^*/π. But to estimate prevalence we need to know θ, the ratio of probands in multiplex families to all probands. The probability that a proband should have $s - 1$ sibs is $sf(s)/\sum_{s=0}^{\infty} sf(s)$, where $f(s)$ is the frequency of families of size s capable of producing trait bearers. The probability that all $s - 1$ sibs will be unaffected is q^{s-1}, where q is the segregation frequency of unaffected children. The mean of this probability for all sibship sizes is $\Sigma sf(s)q^{s-1}/\Sigma sf(s)$, and the ratio of probands in multiplex families to all probands is the complement of this, or

$$\theta = 1 - \sum_{s=1}^{\infty} sf(s)q^{s-1}/\sum_{s=1}^{\infty} sf(s)$$

Then the prevalence is

$$n = A^*/\pi\theta$$

We may estimate θ either empirically or on the assumption that $f(s)$ has a Poisson, geometric, negative binomial, or other distribution.

The *empirical method* assumes that $f(s)$ can be represented with sufficient accuracy by a random sample of sibship sizes from the general population. This assumption is valid only for completed families if onset of the trait is delayed or if it causes premature mortality, for then age and consequently incomplete sibship size will be different from the general population. When the empirical method is applicable, the error of estimate of the prevalence n from v control families is

$$\sigma_n = \sqrt{\sigma_\pi^2\left(\frac{dn}{d\pi}\right)^2 + \sigma_\theta^2\left(\frac{dn}{d\theta}\right)^2} = n\sqrt{\sigma_\pi^2/\pi^2 + \left\{\frac{\Sigma\, sf(s)q^{2s-2}}{\Sigma\, sf(s)} - (1-\theta)^2\right\}/v\theta^2}$$

where the first term in brackets is the expectation of $(q^{s-1})^2$.

Kiser and Whelpton (1944) found some data on completed family size of cohorts to give a good fit to a Poisson distribution. We are concerned only with fertile families, since childless families do not contribute to θ. This truncated Poisson distribution is

$$f(s|s > \upsilon) = \frac{m^s e^{-m}}{s!(1 - e^{-m})} \qquad \begin{array}{l} s = 1, 2, \ldots\ldots \infty \\ m > 0 \end{array}$$

and so the frequency of probands in multiplex families is

$$\theta = 1 - \frac{\Sigma\, sm^s q^{s-1}/s!}{\Sigma\, sm^s/s!}$$

$$= 1 - e^{-mp\pi}$$

where $p = 1 - q$ is the segregation frequency of affected children.

Under incomplete selection, with exclusion of isolated cases, the probability that a family of size s have at least one proband is $1 - (1 - p\pi)^s - sp\pi q^{s-1}$, and the distribution of sibship size under this condition when the distribution in the general population is a truncated Poisson is

$$f(s|r>1) = \frac{\{1 - (1-p\pi)^s - sp\pi q^{s-1}\} m^s/s!}{\sum_{s=1}^{\infty}\{1 - (1-p\pi)^s - sp\pi q^{s-1}\} m^s/s!}$$

$$= \frac{m^s e^{-m}\{1 - (1-p\pi)^s - sp\pi q^{s-1}\}}{s!\{1 - e^{-mp\pi} - mp\pi e^{-mp}\}}$$

The standard error of the prevalence is

$$\sigma_n = \sqrt{\partial K^{-1} \partial'}$$

where K^{-1} is the inverse of the informational K matrix for p, π, and m, and ∂ is the vector of derivatives,

$$\partial = \left(\frac{\partial n}{\partial p}, \frac{\partial n}{\partial \pi}, \frac{\partial n}{\partial m}\right) = \left(\frac{mn}{1-e^{mp}}, -\frac{n}{\pi}, \frac{pn}{1-e^{mp}}\right)$$

Elements for parameters specified by hypothesis are omitted from ∂ and K.

Accidents (Greenwood and Yule, 1920), abortions (James, 1963), and family size (Kojima and Kelleher, 1962) are often fitted by a negative binomial distribution, which Fisher (1941) derived by Poisson trials from a population with a gamma distribution of risks. The truncated negative binomial is

$$f(s|s > 0) = \frac{\binom{z}{s} m^s (1-m)^{z-s}}{1 - (1-m)^z} \qquad \begin{array}{l} s = 1, 2, \ldots\ldots \infty \\ m, z < 0 \end{array}$$

This gives a truncated geometric distribution when $z = -1$. It approaches a logarithmic distribution when z approaches 0 and a truncated Poisson when z approaches ∞, m approaches 0, and mz remains constant (Kendall and Stuart, 1958). The frequency of probands in multiplex families is

$$\theta = 1 - \frac{\Sigma s\binom{z}{s}m^s(1-m)^{z-s}q^{s-1}}{\Sigma\, s\binom{z}{s}m^s(1-m)^{z-s}}$$
$$= 1 - (1-mp)^{z-1}$$

Under incomplete selection with exclusion of isolated cases, the distribution becomes

$$f(s|r > 1) = \frac{\binom{z}{s}m^s(1-m)^{z-s}\{1-(1-p\pi)^s-sp\pi q^{s-1}\}}{\sum\limits_{s=1}^{\infty}\binom{z}{s}m^s(1-m)^{z-s}\{1-(1-p\pi)^s-sp\pi q^{s-1}\}}$$

$$= \frac{\binom{z}{s}m^s(1-m)^{z-s}\{1-(1-p\pi)^s-sp\pi q^{s-1}\}}{1-(1-mp\pi)^z-mp\pi z(1-mp)^{z-1}}$$

The standard error of the prevalence is

$$\sigma_n = \sqrt{\partial K^{-1}\partial'}$$

where K^{-1} is the inverse matrix for p, π, m, and z, and ∂ is the vector of derivatives,

$$\partial = \left(\frac{n(1-z)m(1-mp)^{z-2}}{1-(1-mp)^{z-1}}, -\frac{n}{\pi}, \frac{n(1-z)p(1-mp)^{z-2}}{1-(1-mp)^{z-1}}, \frac{-n\ln(1-mp)}{1-(1-mp)^{1-z}}\right)$$

Elements for parameters specified by hypothesis are omitted from ∂ and K.

For an example of this analysis, see Dewey *et al.* (1965).

THE INFORMATION GAINED ABOUT SEGREGATION PARAMETERS WHEN A DISTRIBUTION OF FAMILY SIZE IS ASSUMED

Gittelsohn (1960) introduced the concept of a prior distribution of family size into segregation analysis. Our first response was unsympathetic. "Attempts have been made to describe complete family size by a modified geometric or Poisson distribution, but in populations with mixtures of contraceptive and noncontraceptive groups and with biological variations in fertility there is no reliable approximation to the distribution of family size, especially incomplete size. Geneticists have preferred to avoid the unknown prior distribution of family size in favor of distributions conditional on fixed size, especially as a distribution of segregants within families is $f(s)P(r,a;s)/\Sigma f(s)P(r,a;s)$. Under and equivocal information about segregation" (Morton, 1962). However, a little reflection has shown that this reaction was premature. The introduction of a prior distribution of family size is useful, as we have seen, to estimate prevalence when isolated cases are excluded from the sample. We turn now to the effects of an assumed family size distribution on segregation analysis.

With any mode of selection of a probands among r affected individuals, the distribution of segregants within families is $f(s)P(r,a;s)/\Sigma f(s)P(r,a;s)$. Under complete selection, the denominator equals 1 and therefore estimation of fertility parameters provides no information about the segregation parameters. However, under incomplete selection, some of the information about segregation can be extracted only when $f(s)$ is given, as in the last section.

The conditional information about a parameter λ_i is $1/K^{ii}$, where K^{ii} is the corresponding diagonal element in the inverse of the K submatrix which includes that parameter and all others estimated from the sample. Let I_o denote this information when fertility parameters m and z are not estimated and I_m, I_{mz} be the information when m or both m and z are estimated simultaneously. Then I_0/I_m and I_0/I_{mz} are the efficiencies for λ_i of methods conditional on fixed sibship size, relative to simultaneous estimation of the fertility parameters.

To study the information gained by simultaneous estimation, we must consider the probability that a family have at least one proband when isolated cases are included. Morton (1959) showed that this is proportional to $xsp\pi + (1-x)\{1-(1-p\pi^s\}$, where x is the probability that a case be sporadic (i.e., an isolated case due to mutation, phenocopy, or other nonrecurrent mechanism). Therefore the distribution of sibship size when the distribution in the general population is a truncated Poisson is

$$f(s|r > 0) = \frac{m^s[xsp\pi + (1-x)\{1 - (1-p\pi)^s\}]}{\sum_{s=1}^{\infty} m^s\{xsp\pi + (1-x)[1-(1-p\pi)^s]\}/s!}$$

$$= \frac{m^s[xsp\pi + (1-x)\{1-(1-p\pi)^s\}]}{[xmp\pi + (1-x)(1-e^{-mp\pi})]s!\, e^m}$$

For the negative binomial,

$$f(s|r > 0) = \frac{\binom{z}{s}m^s(1-m)^{z-s}[xsp\pi + (1-x)\{1-(1-p\pi)^s\}]}{\sum_{s=1}^{\infty}\binom{z}{s}m^s(1-m)^{z-s}[xsp\pi + (1-x)\{1-(1-p\pi)^s\}]}$$

$$= \frac{\binom{z}{s}m^s(1-m)^{z-s}[xsp\pi + (1-x)\{1-(1-\nu\pi)^s\}]}{xmzp\pi + (1-x)\{1-(1-mp\pi)^z\}}$$

To determine whether the information gained by assumption of a family size distribution justifies the inherent approximation, we have examined six bodies of data (Table 2). Two of these are ABO blood group segregations selected through the parents, with no sporadic cases. They are treated here under truncate selection, with nonsegregating families excluded. A third example of truncate selection is provided by deaf-mutism from normal parents, with a substantial frequency of sporadic cases. The remaining studies (of retinoblastoma, Duchenne muscular dystrophy, and spherocytosis) were

TABLE 2. DATA USED TO ESTIMATE FERTILITY PARAMETERS AND PREVALENCE–NEGATIVE BINOMIAL

Source	Number of families		Parameters					Prevalence ± S.E.		Goodness of fit		Reference
	Total $r, s > 0$	Multiplex $r, s > 1$	p	x	π	m	z	Incomplete selection	Multiplex families	χ^2	df	
Japanese A♂ × O♀	238	94	**1/2**	**0**	**1**	—.395	—4.464	384 ± 0	364 ± 28	6.02	6	1
Caucasian O♂ × A♀	341	190	**1/2**	**0**	**1**	—.324	—8.418	620 ± 0	662 ± 33	10.27	8	2
Deaf mutism, normal parents	309	90	**1/4**	.2406	**1**	—.743	—5.310	453 ± 0	423 ± 34	33.17**	11	3
Retinoblastoma	139	4	**1/2**	.915	.690	—.764	—2.162	206 ± 35	192 ± 36	4.97	8	4
Duchenne muscular dystrophy	263	63	.427	**.333**	.519	—.303	—4.092	509 ± 54	521 ± 112	5.75	6	5
Spherocytosis	60	20	**.45**	.266	.097	—.659	—1.960	639 ± 407	612 ± 487	20.33**	7	6
								Pooling to make $e > 5$:		5.42	3	

Parameters specified by hypothesis are in boldface.

$*P < .05$.

$**P < .01$.

[1] Chung, Matsunaga, and Morton (1960).

[2] Chung and Morton (1960).

[3] Chung, Robison, and Morton (1959).

[4] Macklin (1960).

[5] Morton and Chung (1959).

[6] Morton *et al.* (1962).

carried out under multiple selection with ascertainment probabilities ranging from .097 to .690 and the proportions of spcradic cases from .266 to .915. In these studies, some parameters were specified by genetic hypothesis, others had to be estimated empirically. Thus the material provides a variety of conditions under which the consequences of assuming a family size distribution can be examined.

We first fitted the geometric and Poisson distributions, the fertility parameter m being iterated with the relevant segregation parameters by the BINEG program for the IBM 7040 computer (see Appendix). Goodness of fit of the proposed distribution was tested by the likelihood ratio criterion,

$$\chi^2 = 2\ \Sigma o_i \ln(o_i/e_i), \qquad s_i,\ o_i > 0$$

where o_i is the observed number of families of size s_i, the expected number is e_i, and the degrees of freedom are taken as one less than the maximum observed family size. The Poisson distribution gave a total χ^2 of 183.82, df = 52, $P < .001$, and the fit was significantly poor in all six samples. The geometric distribution gave $\chi^2 = 196.20$, df = 52, $P < .001$, and the fit was significantly poor in five samples.

A further conclusion from these tests was that the information gained about segregation parameters by fitting a family size distribution is negligible. Averaging I_m for the Poisson and geometric distribution (which gave nearly identical results), the total amounts of information in the whole material are as follows:

	I_o	I_m	I_o/I_m
p, incomplete selection	7403	7430	.996
x, incomplete selection	5794	5807	.998
p, multiplex families	3917	3936	.995

Thus the sole advantage of fitting a family size distribution is that it permits estimation of prevalence when isolated cases are excluded. The reliability of this estimate depends on the goodness of fit of the fertility distribution. Since the Poisson and geometric are generally unsatisfactory, we turn now to the negative binomial.

Application of this model was at first unsatisfactory because of the tendency of simultaneous estimates of m and z to diverge, often overshooting into the proscribed range of positive numbers. Divergence was decelerated and overshooting stopped by limiting the absolute value of increments to .7 of the estimates. While this braking rule slows divergence, it did not produce convergence in two of 12 analyses (multiplex families of deaf-mutism and retinoblastoma). We therefore introduced *regula falsi* interpolation. Given initial values for m and z, the scores have one of four patterns: $U_m, U_z > 0$ $(++)$; $U_m, U_z < 0$ $(--)$; $U_m > 0, U_z < 0$ $(+-)$; cr $U_m < 0, U_z > 0$ $(-+)$. Ordinary iteration with the braking rule is continued until a complementary pattern is reached, for example $+-$ following an initial pattern of $-+$.

Let the observation preceding the complement which gives the smallest χ^2 be $(m_0, z_0, U_{m_0}, U_{z_0})$ and the complementary set be $(m_1, z_1, U_{m_1} U_{z_1})$. Then interpolation gives improved estimates

$$m_2 = m_0 + \frac{(m_1 - m_0)|U_{m_0}|}{|U_{m_0}| + |U_{m_1}|}$$

$$z_2 = z_0 + \frac{(z_1 - z_0)|U_{z_0}|}{|U_{z_0}| + |U_{z_1}|}$$

from which scores U_{m_2}, U_{z_2} are calculated. If convergence is not obtained in 20 iterations, the computer takes a new trial value from the six estimates with the smallest values of χ^2. These are sufficient to determine a bivariate quadratic,

$$\chi^2 = A + Bm + Cz + Dm^2 + Ez^2 + Fmz$$

which yields the pair of simultaneous equations

$$\frac{\partial \chi^2}{\partial m} = B + 2Dm + Fz = 0$$

$$\frac{\partial \chi^2}{\partial z} = C + 2Ez + Fm = 0$$

whose root is

$$m = (CF - 2BE)/(4DE - F^2)$$
$$z = (BF - 2CD)/(4DE - F^2)$$

By these methods we verified that the m, z matrix in the neighborhood of of root is singular for multiplex families of deaf-mutism and retinoblastoma, the two problem cases, in both of which the Poisson distribution was accepted because it gave a smaller χ^2 than any of the other methods. We have adopted the practice, when iteration fails to converge, of taking as final the estimates (obtained by *regula falsi* or quadratic interpolation or any other method) which give the smallest χ^2.

Having removed the practical difficulties in fitting the negative binomial, we obtained the analyses shown in Table 2. The negative binomial fits splendidly in four of the six samples, but the total χ^2 is 80.51 with 46 degrees of freedom ($P < .01$). Since the Poisson and geometric distributions are special cases of the negative binomial, the superiority of the latter is tested with six degrees of freedom by the difference in χ^2, which is 103.31 for the Poisson and 115.69 for the geometric distribution. This indicates a highly significant superiority of the negative binomial over the Poisson and geometric distributions.

The two samples for which the negative binomial gives a significantly poor fit are spherocytosis and deaf-mutism. In the first case, pooling the largest family sizes (to make the expected frequency exceed 5) reduces χ^2 to 5.42 with 3 df, suggesting that the problem is with the reliability of the χ^2 test in small samples. For deaf-mutism, the observed frequencies exceed their negative binomial expectations for family sizes 1 and 2 and 6 to 8, agree well for families of 9 or more, but are considerably below expectation for family size 3 to 5. This alternation of deviations suggests a mixture of two distributions, and in fact the sample represents two generations, with about equal numbers of deaf-mutes above and below age 30. Goodness of fit of the negative binomial should generally be better in data stratified by generation. The multiplex estimates relate only to high-risk cases, who make up a proportion $1 - x$ of the total prevalence, including sporadic cases. For comparability with the estimates from incomplete selection, in Table 2 we have divided the multiplex estimates and their standard errors by $1 - x$.

In all six samples, the prevalence estimated from multiplex families on the assumption of a negative binomial distribution agrees remarkably well with the prevalence determined from incomplete selection without assuming a family size distribution. It would appear that samples of multiplex families give as precise information about the prevalence of high-risk cases as about their segregation frequency (Morton, 1959, 1962) and that this method of sampling should be more widely used by geneticists to study the etiology of high-risk cases when sporadics are frequent.

SUMMARY

Methods of estimating prevalence are derived for the case of a rare recessive gene and for a trait under incomplete selection, with isolated cases included or not. In the latter event, with the sample restricted to multiplex families, the prevalence estimate requires fitting of a family size distribution. The Poisson and geometric distributions are found to give a poor fit to six bodies of genetic data. However, the negative binomial fits much better and gives a reliable estimate of prevalence. The simultaneous estimation of fertility parameters contributes virtually nothing to information about segregation parameters. These methods, besides being useful to estimate prevalence in the general case of incomplete selection, make it possible to extract full genetic information from samples of multiplex families and therefore may profitably be applied to traits with a high incidence of sporadics whenever it is desired to concentrate on high-risk mechanisms.

ACKNOWLEDGMENT

We are grateful to P. Juetten and V. Pratt for writing the SEGRAN program for the CDC 1604, which was the forerunner of the BINEG program for the IBM 7040, with which the computations for this paper were performed.

REFERENCES

Chung, C. S., Matsunaga, E., and Morton, N. E. 1960. The ABO polymorphism in Japan. *Jap. J. Hum. Genet.* 5: 124–134.

CHUNG, C. S., and MORTON, N. E. 1960. Selection at the ABO locus. *Amer. J. Hum. Genet.* 13: 9–27.

CHUNG, C. S., ROBISON, O. W., AND MORTON, N. E. 1959. A note on deaf mutism. *Ann. Hum. Genet.* 23: 357–366.

DAHLBERG, G. 1948. *Mathematical Methods for Population Genetics.* London: Interscience.

DEWEY, W. J., BARRAI, I., MORTON, N. E., AND MI, M. P. 1965. Recessive genes in severe mental defect. *Amer. J. Hum. Genet.* 17: 237–256.

FISHER, R. A. 1941. The negative binomial distribution. *Ann. Eugen.* (Lond.) 11:182–187.

FURUSHO, T. 1957. A genetic study on the congenital deafness. *Jap. J. Hum. Genet.* 2: 35.

GITTELSOHN, A. M. 1960. Family limitation based on family composition. *Amer. J. Hum. Genet.* 12: 425–433.

GREENWOOD, M., AND YULE, G. Y. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *J. Roy. Stat. Soc.* 83: 255.

JAMES, W. H. 1963. Notes towards an epidemiology of spontaneous abortion. *Amer. J. Hum. Genet.* 15: 223–240.

KENDALL, M. G., AND STUART, A. 1958. *The Advanced Theory of Statistics.* New York: Hafner Publ. Co.

KIMURA, M. 1958. Theoretical basis for the study of inbreeding in man. *Jap. J. Hum. Genet.* 3: 51–70.

KISER, C. V., AND WHELPTON, P. K. 1944. Variations in the size of completed families of 6,551 native white couples in Indianapolis. *Milbank Mem. Fund Quart.* 22: 72–105.

KOJIMA, K., AND KELLEHER, T. M. 1962. Survival of mutant genes. *Amer. Naturalist* 96: 329–346.

LILIENFELD, A. M. 1962. Sampling techniques and significance tests. In *Methodology in Human Genetics,* W. J. Burdette (ed.). San Francisco: Holden-Day, pp. 3–16.

MACKLIN, M. T. 1960. A study of retinoblastoma in Ohio. *Amer. J. Hum. Genet.* 12: 1–43.

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

MORTON, N. E. 1960. The mutational load due to detrimental genes in man. *Amer. J. Hum. Genet.* 12: 348–364.

MORTON, N. E. 1962. Segregation and linkage. In *Methodology in Human Genetics,* W. J. Burdette (ed.). San Francisco: Holden-Day, pp. 17–52.

MORTON, N. E. 1964. Models and evidence in human population genetics. *Proc. XI Internat. Cong. Genet.*: 935–951.

MORTON, N. E., AND CHUNG, C. S. 1959. Formal genetics of muscular dystrophy. *Amer. J. Hum. Genet.* 11: 360–379.

MORTON, N E., MCKINNEY, A. A., KOSOWER, N., SCHILLING, R. F., AND GRAY, M. P. 1962. Genetics of spherocytosis. *Amer. J. Hum. Genet.* 14: 170–184.

NEI, M. 1963. Estimation of recessive gene frequencies from data on consanguineous marriages. *Amer. J. Hum. Genet.* 15: 86–89.

SANK, D. 1963. Genetic aspects of early total deafness. In *A Deaf Population,* J. D. Rainer, K. Z. Altschuler, and F. J. Kallman (eds.). New York: New York State Psychiatric Inst., Columbia University, pp. 28–81.

SMITH, C. A. B. 1957. Counting methods in genetical statistics. *Ann. Hum. Genet.* (Lond.) 21: 254–276.

APPENDIX

Models for family size distribution under incomplete selection.*

s = family size (i.e. number of examined children).

r = number of affected children.

a = number of children who are probands.

*These models are included in BINEG, a Fortran IV program for the IBM 7040. For a general treatment of segregation analysis, see Morton (1959, 1962, 1964).

$p =$ the segregation frequency (i.e. the probability of affection in a high-risk family), $q = 1-p$.

$x =$ the proportion of cases in the population that are sporadic due to mutation, phenocopies, technical errors, extramarital children, rare instances of heterozygous expression of a recessive gene, chromosomal nondisjunction, polygenic complexes, etc.

$\pi =$ the ascertainment probability (i.e. the probability that a case in the population be a proband).

$m =$ the base parameter in a Poisson or negative binomial distribution of family size.

$z =$ the exponential parameter in a negative binomial distribution of family size.

$u_i =$ the maximum likelihood score with respect to the i^{th} parameter.

$K_{ij} = \Sigma u_i u_j =$ the i, j element of the informational K matrix.

Incomplete selection, Poisson distribution (with isolated cases included).

$$P(s|r > 0; p,x,\pi,m) = \frac{m^s [xsp\pi + (1-x) \{1-(1-p\pi)^s\}]}{[xmp\pi + (1-x)(1-e^{-mp\pi})]s!\, e^m}$$

$$u_p = \frac{s\pi[x + (1-x)(1-p\pi)^{s-1}]}{xsp\pi + (1-x)[1-(1-p\pi)^s]} - \frac{m\pi[x + (1-x)e^{-mp\pi}]}{xmp\pi + (1-x)[1-e^{-mp\pi}]}$$

$$u_x = \frac{[sp\pi - 1 + (1-p\pi)^s]}{xsp\pi + (1-x)[1-(1-p\pi)^s]} - \frac{(mp\pi - 1 + e^{-mp\pi})}{xmp\pi + (1-x)[1-e^{-mp\pi}]}$$

$$u_\pi = \frac{sp[x + (1-x)(1-p\pi)^{s-1}]}{xsp\pi + (1-x)[1-(1-p\pi)^s]} - \frac{\{mp[x + (1-x)e^{-mp\pi}]\}}{xmp\pi + (1-x)[1-e^{-mp\pi}]}$$

$$u_m = \frac{s-m}{m} - \frac{x\pi p + (1-x)p\pi e^{-mp\pi}}{xmp\pi + (1-x)[1-e^{-mp\pi}]}$$

Incomplete selection, negative binomial distribution (with isolated cases included).

$$P(s|r > 0; p,x,\pi,m,z) = \frac{\binom{z}{s} m^s (1-m)^{z-s} \{sxp\pi + (1-x)[1-(1-p\pi)^s]\}}{xmzp\pi + (1-x)[1-(1-mp\pi)^z]}$$

$$u_p = \frac{s\pi[x + (1-x)(1-p\pi)^{s-1}]}{xsp\pi + (1-x)[1-(1-p\pi)^s]} - \frac{mz\pi[x + (1-x)(1-mp\pi)^{z-1}]}{xmzp\pi + (1-x)[1-(1-mp\pi)^z]}$$

$$u_x = \frac{sp\pi - 1 + (1-p\pi)^s}{xsp\pi + (1-x)[1-(1-p\pi)^s]} - \frac{mzp\pi - 1 + (1-mp\pi)^z}{xmzp\pi + (1-x)[1-(1-mp\pi)^z]}$$

$$u_\pi = \frac{sp[x + (1-x)(1-p\pi)^{s-1}]}{xsp\pi + (1-x)[1-(1-p\pi)^s]} - \frac{mzp[x + (1-x)(1-mp\pi)^{z-1}]}{xmzp\pi + (1-x)[1-(1-mp\pi)^z]}$$

$$u_m = \frac{s}{m} - \frac{z-s}{1-m} - \frac{zp\pi[x + (1-x)\,(1-mp\pi)^{z-1}]}{xmzp\pi + (1-x)\,[1-(1-mp\pi)^z]}$$

$$u_z = \sum_{i=0}^{s-1} \frac{1}{z-i} - \frac{xmp\pi - (1-x)\,(1-mp\pi)^z \ln(1-mp\pi)}{xmzp\pi + (1-x)\,[1-(1-mp\pi)^z]} + \ln(1-m)$$

Multiplex families, Poisson distribution (with isolated cases excluded).

$$P(s|r > 1; p,\pi,m) = \frac{m^s[1-(1-p\pi)^s - sp\pi q^{s-1}]}{s!e^m\,[1-e^{-mp\pi} - mp\pi e^{-mp}]}$$

$$u_p = \frac{s\pi[(1-p\pi)^{s-1} - (1-sp)q^{s-2}]}{1-(1-p\pi)^s - sp\pi a^{s-1}} - \frac{m\pi[e^{-mp\pi} - (1-mp)e^{-mp}]}{1-e^{-mp\pi} - mp\pi e^{-mp}}$$

$$u_\pi = \frac{sp[(1-p\pi)^{s-1} - q^{s-1}]}{1-(1-p\pi)^s - sp\pi q^{s-1}} - \frac{mp(e^{-mp\pi} - e^{-mp})}{1-e^{-mp\pi} - mp\pi e^{-mp}}$$

$$u_m = \frac{s}{m} - 1 - \frac{p\pi\,[e^{-mp\pi} - (1-mp)e^{-mp}]}{1-e^{-mp\pi} - mp\pi e^{-mp}}$$

Multiplex families, negative binomial distribution (with isolated cases excluded).

$$P(s|r > 1; p,\pi,m) = \frac{\binom{z}{s} m^s\,(1-m)^{z-s}[1-(1-p\pi)^s - sp\pi q^{s-1}]}{1-(1-mp\pi)^z - mp\pi z(1-mp)^{z-1}}$$

$$u_p = \frac{s\pi[(1-p\pi)^{s-1} - q^{s-2}(1-sp)]}{1-(1-p\pi)^s - sp\pi q^{s-1}} - \frac{mz\pi[(1-mp\pi)^{z-1} - (1-mp)^{z-2}(1-mpz)]}{1-(1-mp\pi)^z - mp\pi z(1-mp)^{z-1}}$$

$$u_\pi = \frac{sp[(1-p\pi)^{s-1} - q^{s-1}]}{1-(1-p\pi)^s - sp\pi q^{s-1}} - \frac{mpz[(1-mp\pi)^{z-1} - (1-mp)^{z-1}]}{1-(1-mb\pi)^z - mp\pi z(1-mp)^{z-1}}$$

$$u_m = \frac{s}{m} - \frac{z-s}{1-m} - \frac{p\pi z[(1-mp\pi)^{z-1} - (1-mp)^{z-2}\,(1-mpz)]}{1-(1-mp\pi)^z - mp\pi z(1-mp)^{z-1}}$$

$$u_z = \sum_{i=0}^{s-1} \frac{1}{z-i} + \ln(1-m) + \frac{[(1-mp\pi)^z \ln(1-mp\pi) + mp\pi(1-mp)^{z-1}\{1 + z\ln(1-mp)\}]}{1-(1-mp\pi)^z - mb\pi z(1-mp)^{z-1}}$$