Chapter 3

Human Genetics and Evolution
Estimation of Recessive Gene Frequencies from Data on Consanguineous Marriages

MASATOSHI NEI

Division of Genetics, National Institute of Radiological Sciences, Chiba, Japan

Reprinted from THE AMERICAN JOURNAL OF HUMAN GENETICS Vol. 15, No. 1, March, 1963 Printed in U.S.A.
Estimation of Recessive Gene Frequencies from Data on Consanguineous Marriages

MASATOSHI NEI

Division of Genetics, National Institute of Radiological Sciences, Chiba, Japan

Dahlberg’s (1948) formula for estimating the frequency of recessive genes in man ignores information from consanguineous marriages other than those of first cousins. This limitation was removed by Kimura (1958), who developed a maximum likelihood estimation, but his method requires a tedious iterative solution. The formulas here presented are free of these two criticisms.

Large Populations

In small populations certain types of consanguineous marriages are expected to occur with sufficiently high frequencies even under random mating, but these frequencies rapidly decrease as the size of population increases (cf. Dahlberg, 1929; Nei and Imaizumi, 1963). We thus assume that the population under investigation is so large that all types of consanguineous marriages expected under random mating are negligible and therefore the consanguineous marriages observed are all due to non-randomness of mating.

Following Kimura’s (1958) terminology, let $C_i$ be the frequency of marriages between subjects related with coefficient of parentage $f_i$ where $f_0 = 0$ for unrelated subjects, $f_1 = 1/16$ for first cousins, $f_2 = 1/32$ for first cousins once removed or half-first cousins, $f_3 = 1/64$ for second cousins, etc. ($i = 0, 1, 2, \ldots, r$), and

$$\sum_{i=0}^{r} C_i = 1$$

Here consanguineous marriages having $f_i > 1/16$ are neglected because of their low frequencies in human populations.

In the offspring of parents related with $f_i$ the frequency of recessive homozygotes ($P_i$) for a trait controlled by a pair of genes is expected to be

$$P_i = q^2 + q(1 - q)f_i$$

where $q$ is the frequency of the recessive gene. Thus, the total frequency of recessive homozygotes ($P$) in the population is

$$P = \sum_{i=0}^{r} C_i P_i = q^2 + q(1 - q)f$$

where $f$ is $\sum_{i=0}^{r} C_i f_i$, i.e., the mean inbreeding coefficient in the offspring generation. The proportion of recessive homozygotes whose parents are related with $f_i$ is then given by

Received October 9, 1962.
\[ k_i = \frac{C_iP_i}{P} = \frac{C_i(q + (1 - q)f_i)}{q + (1 - q)f} \]

Hence,

\[ \sum_{i=1}^{r} k_i = \frac{q \sum_{i=1}^{r} C_i + (1 - q) \sum_{i=1}^{r} C_i f_i}{q + (1 - q)f} \]

\[ = \frac{Cq + (1 - q)f}{q + (1 - q)f} = K \]

where \( C \) is the total frequency of consanguineous marriages and \( K \) is the proportion of all recessive homozygotes whose parents are related. The gene frequency is therefore obtained directly from the above equation. That is,

\[ q = \frac{f(1 - K)}{f(1 - K) + K - C} \]

Assuming that \( C \) and \( K \) are distributed as binomials and \( f \) is constant, the variance of \( q \) in large samples is approximately given by

\[ \sigma_q^2 = \frac{(1 - K)(1 - C)f^2}{[f(1 - K) + K - C]^2} \left( \frac{K(1 - C)}{n_k} + \frac{C(1 - K)}{n_c} \right) \]

where \( n_c \) and \( n_k \) are the total numbers of observations on which \( C \) and \( K \) are based respectively.

If \( C_i = 0 \) for \( i \geq 2 \), the formula for \( q \) reduces to Dahlberg's and becomes equivalent to Kimura’s.

**Numerical example** Furusho (1957) obtained the following data for congenital deafism in Japan.

<table>
<thead>
<tr>
<th>Parental relation</th>
<th>Unrelated</th>
<th>First cousins</th>
<th>1½ cousins</th>
<th>Second cousins</th>
<th>2½ cousins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency (%)</td>
<td>879</td>
<td>484</td>
<td>26</td>
<td>68</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(60.12)</td>
<td>(33.11)</td>
<td>(1.78)</td>
<td>(4.65)</td>
<td>(0.34)</td>
</tr>
</tbody>
</table>

\[ C = 0.0862 \quad f = 0.0046 \]

If we assume that congenital deafness is caused by an autosomal recessive gene (cf. Chung et al., 1959 for detail), \( q \) is obtained as follows:

\[ q = \frac{0.0046 \times 0.6012}{0.0046 \times 0.6012 + 0.3988 - 0.0862} = 0.00877 \]

This value is smaller than that obtained by Dahlberg's formula (0.01063) and larger than Kimura's (1958) estimate (0.00761). The standard error of our estimate is calculated to be 0.000586, while in Kimura's method it is 0.00035.

**Comparatively Small Populations**

In populations which are small but not so small as to make random genetic drift important, we represent the frequencies of consanguineous marriages by \( C_i \) as before and those which are expected to occur under random mating by \( c'_i \) (\( i = 0, 1, 2, \ldots, r \)). The excess of \( C_i \) over \( c'_i \) is denoted by \( c_i \), so that \( C_i = c'_i + c_i \). Thus,

\[ \sum_{i=0}^{r} c'_i = 1 \]

\[ \sum_{i=0}^{r} c_i = 0 \]
The frequency of recessive homozygotes in the offspring of parents related with \( f_i \) is given again by

\[ P_i = q^2 + q(1 - q)f_i \]

The excess of recessive homozygote frequency over \( q^2 \) due to inbreeding is

\[ \Delta P = \sum_{i=0}^{r} c_i [q^2 + q(1 - q)f_i] = q(1 - q)f \]

where \( f \) is \( \sum_{i=0}^{r} c_if_i \). Now, the total frequency of recessive homozygotes is

\[ P = q^2 + q(1 - q)f \]

On the other hand, under random mating it is expected that

\[ q^2 = \sum_{i=0}^{r} c'_i \left[ (q^2 - \Delta q^2) + q(1 - q)f_i \right] \]

\[ = q^2 - \Delta q^2 + q(1 - q)f' \]

where \( f' \) is \( \sum_{i=0}^{r} c'_if_i \). Hence,

\[ \Delta q^2 = q(1 - q)f' \]

The proportion of the recessive homozygotes whose parents are related with \( f_i \) is therefore given by

\[ k_i = \frac{c_i[q^2 - q(1 - q)f' + q(1 - q)f'_i]}{q^2 + q(1 - q)f} \]

\[ = \frac{c_i[q + (1 - q)(f_i - f')]}{q + (1 - q)f} \]

and

\[ \sum_{i=1}^{r} k_i = \frac{C_q + (1 - q)[f + f'(1 - C)]}{q + (1 - q)f} = K \]

From this equation we have

\[ q = \frac{f(1 - K) + f'(1 - C)}{f(1 - K) + f'(1 - C) + K - C} \]

The comparison of this formula with the previous one shows that if the consanguineous marriages expected under random mating are neglected in spite of their sufficiently large values, the gene frequency is expected to be under-estimated. The frequencies of consanguineous marriages under random mating are the function of mean and variance of family size and age differences between mates and sibs. They are obtainable under specified conditions by the formulas developed by Nei and Imaizumi (1963), and in general they are sufficiently large for a population of size less than 1,000.

Further, it may be noted that if the frequencies of consanguineous marriages are so small as is expected under random mating the formula given above reduces to

\[ q = \frac{f'(1 - C)}{f'(1 - C) + K - C} \]
If there are no consanguineous marriages other than first cousins, this becomes

\[ q = \frac{c'_1 - c'^2}{16k_1 - 15c'_1 - c'^2} \]

while Dahlberg's formula equivalent to this is

\[ q = \frac{c'_1}{16k_1 - 15c'_1} \]

The difference between our formula and Dahlberg's is due to the fact that in our formula a finite correction has been made for \( q^2 \).

**SUMMARY**

Dahlberg's (1948) formula for estimating the recessive gene frequency in man is extended to cases where information is available on consanguineous marriages other than those of first cousins.

**REFERENCES**


