THE ESTIMATION OF THE FREQUENCY OF A RECESSIVE SEX-LINKED GENE BY THE METHOD OF MAXIMUM LIKELIHOOD

PAULINUS F. FORSTHOEFEL

Department of Biology, University of Detroit, Detroit 21, Michigan

Neel and Schull in their recent book *Human Heredity* (1954) consider the problem of the estimation of genetic parameters by the method of maximum likelihood. The reader is referred to their book for a brief clear discussion of the theory behind this method and its advantages. Considering the case of the estimation of the frequency of a recessive sex-linked gene, Neel and Schull derive the following formula wherein n_{dd} stands for the observed number of females with the recessive phenotype, n_{DY} for the number of males with the dominant phenotype, n_{dY} for the number of males with the recessive phenotype, N_f for the total number of females observed, and N_m for the total number of males observed:

$$q^* = \frac{-n_{DY} + \sqrt{n_{DY}^2 + 4(2N_f + N_m)(2n_{dd} + n_{dY})}}{2(2N_f + N_m)}$$

It is the purpose of this paper to describe an alternative solution of the problem which yields an equivalent formula. The variance of the formula will also be calculated. These formulae may be considered to have an advantage over those derived by Neel and Schull in that they can be reduced to simpler forms when the ratio of males to females is taken to be unity. The question as to how far from unity the sex ratio may deviate and yet for practical purposes be regarded as unity in the carrying out of computations will be considered later on in this paper.

DERIVATION OF THE FORMULA FOR THE MAXIMUM LIKELIHOOD ESTIMATE OF THE FREQUENCY OF A RECESSIVE SEX-LINKED GENE

Let N stand for the observed number of females in the population. Let x stand for the sex ratio, viz., the number of males divided by the number of females. Then the number of males can be represented by xN, and N plus xN will be the number of individuals in the total population. Let k stand for the proportion among females of females with the character. Let g stand for the proportion among males of males with the trait. Let q stand for the true frequency of the recessive sex-linked gene a, a numerical value of which is to be estimated from the available data. In Table 1 the population is divided into four classes according to genotypes, Aa and AA among females being grouped since these are phenotypically indistinguishable. The relative frequency, the numerical frequency, and the probability of occurrence of each class are also given.

The exact probability (P) of getting the observed sample can be expressed as:

$$P = \frac{(N+xN)!}{Nk!\quad N(1-k)!\quad xNg!\quad xN(1-g)!} \cdot (q^2)^{Nk} \cdot (1-q^2)^{N(1-k)} \cdot q^{xNg} \cdot (1-q)^{xN(1-g)}$$

The specific problem here is to derive an expression for q such that for this value of q the probability of getting the observed sample is maximum. A solution giving a maximum logarithm of the probability will be identical with the desired solution. Let C stand for the first factor in the right-hand member of the equality.

Then, taking natural logarithms of both sides of the equality, we have:

$$\log P = \log C + 2Nk \log q + N(1-k) \log (1-q^2) + xNg \log q + xN(1-g) \log (1-q)$$

The next step is to take derivatives of both sides of the equality with respect to q:

$$\frac{d \log P}{dq} = \frac{2Nk}{q} - \frac{2N(1-k)q}{(1-q^2)} + \frac{xNg}{q} - \frac{xN(1-g)}{1-q}$$

The third step is to set the derivative equal to zero and to solve for q. When the derivative has zero as its value, the estimated value of q will give the maximum value for the probability of getting the observed sample, and is the value of q desired. This maximum likelihood estimate of the parameter q may be designated as above by q^*, and is:

$$q^* = \frac{-x(1-g) + \sqrt{x^2(1-g)^2 + 4(2+x)\quad(2k+xg)}}{4+2x}$$

By making the appropriate substitutions, it can be shown easily that the formula derived by Neel and Schull can be reduced to this formula. In the form derived here, the formula admits of a simplification when the sex ratio, x, is unity. In this special case, the formula reduces to:

$$q^* = \frac{1}{6}\left(g - 1 + \sqrt{1 + 10g + g^2 + 24k}\right)$$

TABLE 1

The Relative Frequency, the Numerical Frequency, and the Probability of Occurrence of Four Classes Resulting from a Recessive Sex-Linked Gene

Class	Relative Frequency	Numerical Frequency	Probability
aa	k	Nk	q^2
$A-$	$1-k$	$N(1-k)$	$1-q^2$
*a*Y	g	xNg	q
*A*Y	$1-g$	$xN(1-g)$	$1-q$

VARIANCE OF THE ESTIMATE

The variance of the maximum likelihood estimate can be found by use of the following formula (Hogben, 1946):

$$-\frac{1}{\sigma^2_{q^*}} = \Sigma\left(n f_i \frac{d^2 \log f_i}{dq^2}\right)$$

Table 2 contains the quantities needed to calculate the variance. Applying Hogben's formula, we have:

$$\sigma^2_{q^*} = \frac{q(1-q^2)}{N(4q+x+xq)}$$

When x is unity, the formula for the variance reduces to:

$$\sigma^2_{q^*} = \frac{q(1-q^2)}{N(1+5q)}$$

TABLE 2

Theoretical Probabilities and Derived Quantities Needed for the Calculation of the Variance

Class i	Probability f_i	$\frac{d^2 \log f_i}{dq^2}$	$\frac{n f_i \, d^2 \log f_i}{dq^2}$
1: *aa*	q^2	$\frac{-2}{q^2}$	$-2N$
2: *A*−	$1-q^2$	$\frac{-2(1+q^2)}{(1-q^2)^2}$	$\frac{-2N(1+q^2)}{(1-q^2)}$
3: *a*Y	q	$\frac{-1}{q^2}$	$\frac{-xN}{q}$
4: *A*Y	$1-q$	$\frac{-1}{(1-q)^2}$	$\frac{-xN}{(1-q)}$
Total	—	—	$\frac{-N(4q+x+xq)}{q(1-q^2)}$

AN EXAMPLE OF ESTIMATION

Waaler (1926) studied the inheritance of several kinds of color blindness in the human population of Oslo, Norway. In the course of his investigation, he recorded the color vision of 9049 boys and 9072 girls. Among these he found 725 color blind boys and 40 color blind girls. The quantities needed for the application of the formulas derived in this paper are x, the sex ratio, which is here 0.9975; g, the proportion of males having the trait, here 0.08012; k, the proportion of females with the trait, here 0.00441. Substituting these values in the formula for q^*, we calculate that q^* equals 0.0772. Neel and Schull after applying their formula to the same data arrived at the same result. Using the formula for the variance derived above, we get 0.000006112 as the variance of q^*. Neel and Schull do not explicitly give a formula for the variance of their estimate of q, but they give the formula for the invariance. The formula for the variance is the reciprocal of the formula for the invariance. It may be noted here in passing that the formula they give for the variance of q estimated for the case of allelic sex-linked genes lacking dominance can not be used in the present case of a recessive sex-linked gene. Returning to Waaler's data, and recalling that the standard deviation is the square root of the variance, we get 0.00247 as the standard deviation of q^*. This is identical with the value given for the same quantity by Neel and Schull.

USE OF THE SIMPLIFIED FORMULAS WHEN THE SEX RATIO DEVIATES FROM UNITY

The question arises as to how far from unity the sex ratio may be and yet be regarded as unity in calculating the maximum likelihood estimates of q and its variance, and thus permit the use of the simplified formulas given above in which the sex ratio, x, does not occur. The answer to this question will depend on how accurately one wishes or needs to know the value of q^* and its variance. Concrete examples may help one to decide in a particular case. Let us assume that the proportion of males with the recessive sex-linked trait (g) is 0.2 and the proportion

of females with it (k) is 0.01. With these values constant, let x, the sex ratio, vary from unity to 0.9 to 0.8 to 0.7 to 0.6 to 0.5. The corresponding values of q^* by the long formula will be 0.168, 0.166, 0.164, 0.161, 0.157, and 0.154. The value of q^* by the simplified formula which assumes the value of x to be unity remains 0.168. It may be concluded that use of the simplified formula when x ranges between 0.9 and unity gives a value close enough to the true maximum likelihood value to serve as a useful approximation. Of course, the closer x actually is to unity, the closer will be the approximation. Thus, using the simplified formula for Waaler's data, taking x to be unity and rounding off g to 0.08 and k to 0.0044, we get q^* to be 0.0772. This is the same value we got when using the more exact figures given above in the long formula, x being 0.9975.

It may be noted here that whenever k is equal to the square of g, the maximum likelihood estimate of q is always equal to g, no matter what value x has. This fact can be easily verified by substituting various values for x in the formula derived in this paper, but always keeping k equal to the square of g. It follows that in this special case the formula for the maximum likelihood estimate of q can be simplified to: q^* equals g.

A few words may suffice on the use of the simplified formula for the variance. Supposing N (the total number of females) to be 1000, g to be 0.2, k to be 0.01, and x to be 0.9, q^* by the long formula is 0.166. The variance by the long formula is 0.000094 while by the simplified formula which assumes that x is unity, the value is 0.000088. The difference is 0.000006. The corresponding standard deviations are 0.0097 and 0.0094, differing by 0.0003. It may be concluded again that when x ranges between 0.9 and unity, the simplified formula for the variance can give a useful approximation to the exact maximum likelihood estimate. Thus using the simplified formula for Waaler's data, taking x to be unity, N to be 9072, and q^* to be 0.0772, we calculate the variance as 0.000006099. This value differs by 0.000000013 from the exact figure given above using the long formula with x as 0.9975.

SUMMARY

This paper describes solutions of the problems of deriving the maximum likelihood estimate of the frequency of a recessive sex-linked gene and its variance. The formulas derived are equivalent to those derived by Neel and Schull, but admit of simplification when the sex ratio is unity. The question is discussed as to how far from unity the sex ratio may deviate and yet be regarded as unity in computations.

ACKNOWLEDGMENT

The author wishes to express his appreciation to Dr. Earl L. Green of the Department of Zoology of The Ohio State University for many helpful suggestions in preparing this paper.

REFERENCES

Hogben, L. 1946. An introduction to mathematical genetics. W. W. Norton and Co., New York.

Neel, J. V., and **W. J. Schull.** 1954. Human heredity. The University of Chicago Press, Chicago.

Waaler, G. H. M. 1926. Über die Erblichkeitsverhältnisse der vershieden Arten von angeborener Rotgrünblindheit. *Zeitschr. f. Abstgs. u. Vererbgsl.* 45: 299–333.