ON NEI AND ROYCHOUDHURY'S SAMPLING VARIANCES OF HETEROZYGOSITY AND GENETIC DISTANCE

S. MITRA

Emory University, Atlanta, Georgia 30322

Manuscript received August 9, 1974

ABSTRACT

NEI and ROYCHOUDHURY have developed an inequality relationship between the expected values of $(j_x - g)^2$ and $(\hat{g} - g)^2$ where j_x is a biased and $\hat{g}$ is an unbiased estimate of population homozygosity g . Their inequality is somewhat weak and can be improved, as is demonstrated in this paper.

1. The Problem

NEI and ROYCHOUDHURY's (1974) article presents an interesting inequality relationship between the expected values of $(j_x - g)^2$ and $(\hat{g} - g)^2$ where the sample homozygosity j_x is a biased estimator of g with a bias of $(1-g)/n$ and $\hat{g} = (nj_x - 1)/(n-1)$ is the corresponding unbiased estimate of $g = \sum p_i^2$, p_i being the relative frequency of the i th allele in the population such that $\sum p_i = 1$ and n is the number of genes sampled at random from the population ($n/2$ individuals in diploid organisms). In the APPENDIX of the said article they have shown that

$$E(\hat{g} - g)^2 - E(j_x - g)^2 = [4(2n-1)(n-2)(\sum p_i^3 - g^2) + (1-g)\{(n^2 + 3n - 2)g - n(n-1)\}]/[n^3(n-1)] \quad (1)$$

in which

$$\sum p_i^3 - g^2 = \sum_{i>j} p_i p_j (p_i - p_j)^2 \geq 0 \quad (2)$$

and the second term is positive if

$$g > \frac{n^2 - n}{n^2 + 3n - 2} \quad (3)$$

The average value of g being 0.9, equation (1) is greater than zero for $n < 36$. For $n \geq 36$, the authors are not clear about the inequality relationship.

2. Development of Stronger Inequality

It may be pointed out that (3) is a weak inequality and can be improved if the contribution of (2)—which is positive except for the special case when p_i s are all equal—is taken into account. An attempt has been made in the following to determine the minimum of (2), especially the minimum of

$$K = \sum p_i^3 \quad (4)$$

subject to the restrictions that

$$\sum p_i = 1 \quad (5)$$

and

$$\sum p_i^2 = g \quad (6)$$

This can be done by minimizing

$$\Sigma p_i^3 - \lambda_1 (\Sigma p_i^2 - g) - \lambda_2 (\Sigma p_i - 1) \quad (7)$$

where λ_1 and λ_2 are undetermined multipliers. Differentiating (6) with respect to p_i , one obtains

$$3p_i^2 - 2\lambda_1 p_i - \lambda_2 = 0. \quad (8)$$

If there are l alleles (1, 2 . . . l), there are l quadratic equations like (8). Since the coefficients of the quadratic are the same for all i , it is apparent that there are at most two different values of p_i s. In other words, there are, say, l_1 p s which are all equal to p_1 and another set of l_2 p s all equal to p_2 .

Thus,

$$l_1 + l_2 = 1 \quad (9)$$

$$l_1 p_1 + l_2 p_2 = 1 \quad (10)$$

$$l_1 p_1^2 + l_2 p_2^2 = g \quad (11)$$

$$\text{and } K = \Sigma p_i^3 = l_1 p_1^3 + l_2 p_2^3 \\ = p_1 g + p_2 g - p_1 p_2 \quad (12)$$

obtained by multiplying (11) with p_1 and p_2 , adding the results and collecting terms. Similarly, multiplying (10) with p_1 and p_2 and adding the results, one gets

$$g = p_1 + p_2 - l p_1 p_2 \quad (13)$$

and

$$g^2 = (p_1 + p_2)g - l g p_1 p_2 \quad (14)$$

so that

$$K - g^2 = p_1 p_2 (lg - 1). \quad (15)$$

Observe that (15) vanishes when $g = 1/l$ (the case of equal p values), as it should. In other cases, $g > 1/l$ and (15) is positive.

For the case of unequal p 's, let $p_1 > p_2$. Multiplying (10) by g and subtracting from (11)

$$l_1 (p_1^2 - p_1 g) + l_2 (p_2^2 - p_2 g) = 0. \quad (16)$$

Because $p_1 > p_2$, $p_1^2 > p_1 g$ due to (16), or

$$p_1 > g. \quad (17)$$

Again, multiplying (11) by l_1 and l_2 , adding and simplifying, one can write,

$$(l_1 p_1 + l_2 p_2)^2 + l_1 l_2 (p_1 - p_2)^2 = l g$$

so that

$$(p_1 - p_2)^2 = \frac{lg - 1}{l_1 l_2}. \quad (18)$$

The maximum value of (18) is obtained when the denominator $l_1 l_2$ is minimum. Since none can be zero, the minimum value is $l - 1$ for which

$$(p_1 - p_2)^2 \leq \frac{lg - 1}{l - 1} \quad (19)$$

and further, for $g < 1$

$$(p_1 - p_2)^2 < g \quad (20)$$

so that an upper bound of $p_1 - p_2$ is established. It may be pointed out for $g > .5$, l_1 must equal 1 because of (10), which is usually the case.

For $l_1 = 1$ and $l_2 = l-1$

$$p_2 = \frac{1-p_1}{l-1} \quad (21)$$

from (10). Equation (20) can be rewritten as

$$p_1 - p_2 < \sqrt{g}$$

or

$$p_1 - \frac{1-p_1}{l-1} < \sqrt{g}$$

or

$$p_1 < \frac{1}{l} + \frac{l-1}{l} \sqrt{g} \quad (22)$$

and

$$p_2 > \frac{1}{l} (1 - \sqrt{g}) \quad (23)$$

The minimum value of $p_1 p_2$ or $p_1(1-p_1)/(l-1)$ is obtained when p_1 assumes its maximum value (for $p_1 > .5$). That is to say,

$$p_1 p_2 > \frac{1}{l^2} [(1 + (l-1)\sqrt{g}) [1 - \sqrt{g}]]$$

or

$$> \frac{1}{l^2} [1 + (l-2)\sqrt{g} - (l-1)g] \quad (24)$$

or

$$> \frac{1-g}{l^2} \quad (25)$$

since $\sqrt{g} > g$.

Thus (15) can be rewritten as

$$K - g^2 > \frac{(1-g)(lg-1)}{l^2} \quad (26)$$

NEI and ROYCHOUDHURY's inequality shown in (3) can now be improved as

$$g > \frac{n^2 - n + 4(2n-1)(n-2)/l^2}{n^2 + 3n - 2 + 4(2n-1)(n-2)/l} \quad (27)$$

An interesting feature of (26) is that it also depends on l , the number of alleles. For $l=5$, the above holds when g is as low as .5, even for $n=100$.

3. Concluding Observations

It may be noted that the inequality shown in (25) is still weak but is developed and used for its elegance. As such, the expression (24) is the best. Several other inequalities can also be generated from (24) which are better than (25) but inferior to (24). For example, the right-hand side of (24) is

$$\begin{aligned} & [1 + (l-2)\sqrt{g} - (l-1)g]/l^2 \\ & = [(1-g) + (l-2)(\sqrt{g}-g)]/l^2 \end{aligned}$$

$$\begin{aligned}
&= (1-g) \left[1 + \frac{(l-2)g}{\sqrt{g} + g} \right] / l^2 \\
&> \frac{(1-g)g}{l(\sqrt{g} + g)} \quad (28)
\end{aligned}$$

since $g \leq \sqrt{g} \leq 1$.

Expression (28) is also greater than

$$\frac{(1-g)g}{2(l)} \quad (29)$$

which, though somewhat weaker than (28), is simple and more elegant.

LITERATURE CITED

NEI, MASATOSHI and A. K. ROYCHOUDHURY, 1974 Sampling variances of heterozygosity and genetic distance. *Genetics* **76**: 379-390.

Corresponding editor: R. C. LEWONTIN