LIKELIHOOD INFERENCE OF PATERNITY

To the Editor: While hesitating to enter the current debate concerning statistical approaches to paternity testing, the more general area of likelihood inference in the estimation of genealogical relationships [1, 2] has some bearing on the problem. In this context, some of the statistical statements made in the recent paper by Li and Chakravarti [3] are incorrect and require clarification if still greater confusion is not to reign. This is *not* to suggest that the posterior probability method based only on the fact of nonexclusion, advocated by [3], is either in error or unwise. Indeed, as discussed by Walker [4], it has merit, and may be optimal at the current time. However, the discussion of "fallacies" leading to the conclusion that it is the only possible approach contains several serious errors.

For convenience and brevity, I shall refer to [3] as L & C, to the penultimate sentence of their first paragraph as equation (1), to their formula for $\Sigma f_i^2/g_i$ on p. 811 as equation (2), and to that for $\Sigma g_i^2/f_i$ on p. 812 as equation (3). Further, the equation for W(3) on p. 815 will be equation (4), and their final recommendation, p_t (p. 816), will be equation (5). For clarity, insofar as it is possible, I shall comply with their notation and number my own equations from (6) upwards. In equation (1), L & C make two errors, which are the source of their later difficulties. A likelihood ratio is a ratio of likelihoods, *not* of probabilities of paternity, and the L that they discuss relates *not* to a comparison of individuals (fathers and non-fathers) but to two genealogical hypotheses ("father" and "unrelated") about a given man. The latter distinction is important and has implications reaching far beyond paternity testing [5, 6].

On p. 814, L & C correctly state that a likelihood ratio must be of the form

$$L(H_1{:}H_0) = P(X|H_1)/P(X|H_0) \tag{6}$$

for data X and hypotheses H_1 and H_0. The L of L & C is precisely such a ratio with X = genetic data on M, C, F; H_1 = M is mother of C, and F is father of C; H_0 = M is mother of C, and F is unrelated. There can, thus, be no doubt that L *is* a likelihood ratio. In that it is often misused and misinterpreted, I would agree with L & C; but they themselves are not immune to this.

Now L & C discredit L by showing that true fathers give mean values greater than 1, while "non-fathers" have mean value 1. But this is a trivial and well-known result for any likelihood ratio:

$$\left.\begin{aligned} E(L(H_1{:}H_0)|H_0) &= \int \frac{P(X|H_1)}{P(X|H_0)}\, p(X|H_0)\, dX = 1 \\ E(L(H_1{:}H_0)|H_1) &= \int \frac{P(X|H_1)}{P(X|H_0)}\, p(X|H_1)\, dX > 1 \end{aligned}\right\} \tag{7}$$

by an exactly analogous derivation to equation (2). It is not surprising that the ratio for H_1 against H_0 is higher, on average, if H_1 is true than if H_0 is so; likelihood inference would have little validity were this not true.

If log-likelihoods were used, one could go further and show that the log-likelihood-ratio (or log-likelihood difference) is maximal in expectation for the true hypothesis: precisely, $E(\log L(H_1:H_0):H^*)$ is maximal with respect to variation in H_1 for $H_1 = H^*$ (see, for example [7]).

Thus, the comments of L & C, concerning distributions of L under other genealogical hypotheses and genetic assumptions (for example, nonexclusion) also fall within a standard likelihood analysis of the properties of these statistics. A recent study of these properties in the general context of genealogy reconstruction is given by [8]. L & C also state that by reversing the hypotheses H_0 and H_1 they reverse the expectation [equation (3)]. Again, it would be surprising if it did not! The likelihood ratio $L(H_0:H_1)$ is the inverse of $L(H_1:H_0)$ and is a ratio for H_0 against H_1, so must, on average, be larger if H_0 is true than if H_1 is so [cf. equation (7)]. The ratio g_i/f_i in the notation of L & C is the likelihood *for* unrelatedness *against* paternity. The ratio f_i/g_i is the ratio *for* paternity *against* unrelatedness.

Now assuming prior probability p_0 of paternity, and that "nonpaternity" is equivalent to the hypothesis H_0, H_0, and H_1 being the only possibilities, Bayes theorem gives immediately a posterior probability

$$\begin{aligned} P(\text{paternity}|\text{data}) &= \frac{p_0\, P(\text{data}|H_1)}{p_0\, P(\text{data}|H_1) + (1 - p_0)\, P(\text{data}|H_0)} \qquad (8) \\ &= p_0/[p_0 + (1 - p_0)Y] \\ \text{where } Y &= P(\text{data}|H_0)/P(\text{data}|H_1) \\ &= L(H_0:H_1) = 1/L\ . \qquad (9) \end{aligned}$$

Note now the parallel between equations (8) and (5), the proposal of L & C. The "data" assumed by L & C are assumed only to be exclusion/nonexclusion and, thus, for a nonexcluded individual,

$$\begin{cases} P(\text{data}|H_1) = 1 & \text{(nonexclusion is certain)} \\ P(\text{data}|H_0) = \Pi_1^n(1 - x_i) & \text{in notation of L \& C} \end{cases}$$

and equation (8) becomes equation (5).

The full data, however, consist of the genetic marker types of the three individuals in question [M, C, and F]. For *these* data

$$\begin{cases} P(\text{data}|H_0) \propto \Pi_1^n g_i \\ P(\text{data}|H_1) \propto \Pi_1^n f_i \end{cases}$$

and $Y = 1/\Pi_1{}^n L(i)$ in the notation of L & C. Thus, equation (8) is then equivalent to equation (4) of L & C, if $p_0 = \frac{1}{2}$. However, $p_0 = \frac{1}{2}$ is not normally an appropriate prior: use of a p_0 based on population information, such as that advocated by L & C, is clearly preferable.

Now L & C dismiss the use of equation (4) as being not necessarily monotonic in the tests performed. But consider a man who has, as in the example of L & C, a very low value of L(4), despite not being excluded as a father by this test. Then $W(4) < W(3)$. However, by definition, $L(4) < 1$ simply because the joint phenotypes of M, C, and F at this locus do indeed arise with lower probability under hypothesis H_1 than under H_0; that is, the relative likelihoods of paternity and unrelatedness *for this man* are decreased by the data, although he is not excluded. The man might well argue his right to have this valid quantitative evidence taken into account. The result does *not* concern the relative likelihoods relating to any *other* man. Another man presumably has the right to request the same test.

In conclusion, equations (8) and (9) [the generalizations of equation (4)] provide a perfectly valid posterior probability, encompassing more information than does equation (5). In both cases, appropriate population frequencies are required, and other relevant information can be encompassed [4]. However, equation (8) is more sensitive to such items as population frequencies than is equation (5), and where there is debate as to appropriate values, equation (5) may be the more robust procedure. Nonetheless, it is not the only candidate, nor universally optimal.

E. A. THOMPSON[1]

REFERENCES

1. THOMPSON EA: Inference of genealogical structure, I–III. *Soc Sci Inform* 15:477–526, 1976
2. THOMPSON EA: The estimation of pairwise relationship. *Ann Hum Genet* 39:173–188, 1975
3. LI CC, CHAKRAVARTI A: Basic fallacies in the formulation of the paternity index. *Am J Hum Genet* 37:809–818, 1985
4. WALKER RH: Guidelines for reporting estimates on the probability of paternity (Letter to the Editor). *Am J Hum Genet* 37:819–825, 1985
5. THOMPSON EA: A paradox of genealogical inference. *Adv Appl Prob* 8:648–650, 1976
6. THOMPSON EA, MEAGHER TR: The joint distribution of likelihood statistics, and their use in genealogical inference. Manuscript in preparation
7. EDWARDS AWF: *Likelihood*. Cambridge, England, Cambridge University Press, 1972
8. MEAGHER TR, THOMPSON EA: The relationship between genetic likelihoods in genealogy reconstruction, *Theor Popul Biol* 29:87–106, 1986

Received November 26, 1985; revised April 10, 1986.

[1] Statistical Laboratory, 16 Mill Lane, Cambridge CB2 1SB, England. Current address: Department of Statistics, GN-22, University of Washington, Seattle, WA 98195.