BIOMETRICS 57, 975–981
September 2001

CORRESPONDENCE

Comment on Schwarz and Arnason: "Estimation of Age-Specific Breeding Probabilities from Capture–Recapture Data"

From: *Morten Frederiksen* and Roger Pradel*
CEFE/CNRS
1919 Route de Mende
F-34293 Montpellier Cedex 5, France
**Current address:*
NERI, Department of Coastal Zone Ecology
Kalø, Drenåvej 12
DK-8410 Rønde, Denmark

To the Editor of Biometrics:

In a recent article, Schwarz and Arnason (2000) describe a new approach for direct estimation of local recruitment from capture–recapture data. They provide estimates of b_{ij}, the age- and cohort-specific probabilities that an animal that survives until it starts breeding will do so at age $j+1$. They term the b_{ij} "age-specific breeding proportions" (or probabilities) and claim that the α_i estimated indirectly by the method of Clobert et al. (1994) are biased estimators of the same quantities. We argue here that the b_{ij} are not equivalent to the α_i and that the b_{ij}, relevant population parameters in their own right, should not be termed "age-specific breeding proportions."

As defined by Schwarz and Arnason (2000, p. 60), b_{ij} estimates the "probability that an animal in cohort i that survives until it starts to breed will start breeding at age $j+1$." Pradel and Lebreton (1999) define α_i as "the probability that an animal of age i is a first-time breeder" (p. 75). Even when the obvious differences in the definitions have been accounted for (no cohort effect in the definition of α_i, age $j+1$ = age i), it should be clear that these probabilities cannot be expected to be equal. Indeed, they are related through survival between the earliest age of breeding (k in Schwarz and Arnason (2000), y in Pradel and Lebreton (1999)) and age i ($j+1$). A simple numerical example illustrates this (Table 1).

Information about survival is thus necessary to estimate the α_i, and if prebreeders are nonobservable, the α_i can only be estimated under the assumption of equal (possibly age-specific) survival of prebreeders and breeders.

We suggest that the term "age-specific breeding proportions" should be applied to the $\Sigma_y^i \alpha_i$ (using the definitions of Pradel and Lebreton (1999)), which estimate the proportion of breeders among all animals aged i. This quantity is a component of the age-specific fecundity, which is used in Leslie matrices and other mathematical population models.

Table 1

A numerical example of the relationship between α_i and b_{j+1}. Parameter values are taken from Table 2 in Schwarz and Arnason (2000) *and illustrate recruitment in black-headed gulls* (Larus ridibundus) *at a colony in France (Clobert et al.,* 1994). *In this study, survival was constant over age (from age* 2) *and time at a level of* 0.806 *and the earliest age of breeding (y) was* 2. *The sum of the intermediate parameters d_i provides an estimate of the proportion of animals alive at age y that will breed at some time during their life (Frederiksen and Bregnballe,* 2001). *The values we obtain for b_{j+1} are exactly the same as the direct estimates of Schwarz and Arnason* (2000).

Age	α_i		d_i		b_{j+1}
2	0.214	Multiply by	0.214	Scale sum to	0.299
3	0.316	0.806^{i-2}	0.255	1 (multiply	0.356
4	0.001		0.001	by 1/0.716)	0.001
5	0.470	$\longrightarrow$	0.246	$\longrightarrow$	0.344
Sum	1		0.716		1

The term "age-specific proportions of first-time breeders" is ambiguous since it could apply both to the α_i (proportions of first-time breeders among all animals aged i) and to the b_{ij} (proportions of animals aged $j+1$ among all first-time breeders).

If an empirical estimate of the mean age at first breeding (recruitment) is required, it can, as argued by Schwarz and Arnason (2000), be calculated from the b_{ij} or, equivalently, from d_i (Table 1; Frederiksen and Bregnballe, 2001). The various quantities in use for estimating age-specific recruitment from capture–recapture data are thus not equivalent, and each should be used only for the purposes for which it is suitable.

REFERENCES

Clobert, J., Lebreton, J.-D., Allainé, D., and Gaillard, J.-M. (1994). The estimation of age-specific breeding probabilities from recaptures or resightings in vertebrate populations: II. Longitudinal models. *Biometrics* **50,** 375–387.

Fredericksen, M. and Bregnballe, T. (2001). Conspecific reproductive success affects age of recruitment in a great cormorant *Phalacrocorax carbo sinensis* colony. *Proceedings of Royal Society of London, Series B,* in press.

Pradel, R. and Lebreton, J.-D. (1999). Comparison of different approaches to the study of local recruitment of breeders. *Bird Studies* **46**(Suppl.), 74–81.

Schwarz, C. J. and Arnason, A. N. (2000). Estimation of age-specific breeding probabilities from capture–recapture data. *Biometrics* **56,** 59–64.

The authors replied as follows:

We would like to thank Frederiksen and Pradel for their comments on our article. These are helpful in distinguishing between the two estimators.

The β-parameterization adopted by Schwarz and Arnason (2000) indicates what fraction of all breeders started to breed at each age, i.e., if 1000 animals in total did breed, then the β's indicate that, on average, 299, 356, 1, and 344 started to breed at ages 2, 3, 4, and 5, respectively. As Frederiksen and Pradel pointed out in their letter, these are appropriate for such global parameters as the mean age of first breeding. This was the focus of Schwarz and Stobo (2000).

The α-parameterization of Pradel and Lebreton (1999) measures breeding proportions of animals alive. For example, if 1000 animals were alive at age 2, then about 214 would start to breed at age 2; of the remaining $1000 \times .806 = 806$ animals alive at age 3, then $.316 \times 806 = 254$ would start to breed at age 4, etc. We agree with the authors that the cumulative α-sum measures the "proportion of breeders among all animals aged i." For example, continuing with the example above, of the 214 animals that started to breed at age 2, $0.806(214) = 172$ animals are alive at age 3 and breeding; there are a total of $172 + 254 = 426$ animals who are breeders at age 3, which corresponds to $426/806 = .528$ as the proportion of breeders among all animals aged 3.

Neither method directly estimates the breeding proportion needed to form the Leslie matrix parameters, but appropriate estimates can be derived in both cases. Working out the derived estimates and their SE could be complex, especially if survival is time or age dependent.

The key difference between the two estimators is the conditioning involved, as noted in our article—it is not surprising that the two estimators measure different aspects of the problem.

We agree with their concern that the term "age-specific breeding proportions" is ambiguous and the estimates obtained under any of the methods need to be interpreted carefully and used appropriately.

References

Schwarz, C. J. and Stobo, W. T. (2000). Estimation of juvenile survival, adult survival, and age-specific pupping probabilities for the female grey seal (*Halichoerus gryprus*) on Sable Island from capture–recapture data. *Canadian Journal of Fisheries and Aquatic Sciences* **57**, 247–253.

C. J. Schwarz and A. N. Arnason
Department of Statistics and Mathematics
Simon Fraser University
Burnaby, British Columbia V5A 1S6
Canada

Comment on Stockmarr's "Likelihood Ratios for Evaluating DNA Evidence When the Suspect Is Found Through a Database Search"

From: *A. P. Dawid*
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT, U.K.

To the Editor of Biometrics:

In a paper in *Biometrics*, Stockmarr (1999) considers the case that a DNA database $\mathcal{D}$ of size n is searched, and a single individual S (say Smith) in $\mathcal{D}$ is found to have a DNA profile matching a trace found at the scene of a crime, which can be assumed to come from the true perpetrator TP. No other evidence is available. He argues that the strength of the evidence in favor of Smith's guilt, i.e., the hypothesis $TP = S$, is, under some simple assumptions, captured by a likelihood ratio of $1/np$, p being the frequency of the trace DNA profile in the population at large. When so measured, the evidence becomes rapidly weaker as the size n of the database increases. This is in agreement with recommendations of the U.S. National Research Council (National Research Council, 1996) but in serious conflict with other treatments (Balding and Donnelly, 1996; Dawid and Mortera, 1996), which conclude that the evidence against Smith becomes stronger, albeit typically only marginally, as n increases. The issue has been discussed by Donnelly and Friedman (1999) and, in a previous response to Stockmarr, by Evett, Foreman, and Weir (2000). Here I wish to elaborate on and extend some of their arguments for rejecting Stockmarr's position. Although his mathematics are essentially correct, his logic is faulty. Properly interpreted, his own analysis undermines, rather than supports, the conclusions he draws and supports instead the views he criticizes.

Stockmarr rejects the usual approach of directly comparing the hypotheses he labels H'_p, H'_d, expressing the guilt or innocence of Smith, on the grounds that these are 'data-dependent' since the identification of Smith as an individual matching the crime trace cannot be made until after the search has been conducted. Instead, he sets up hypotheses H_p and H_d that TP is or is not in $\mathcal{D}$. He then proposes to measure the strength of the evidence against Smith by means of the likelihood ratio in favor of H_p as against H_d. Although I myself do not share Stockmarr's aversion to calculating likelihoods for data-dependent hypotheses, for the sake of further argument, I shall fully concede this point and consider only approaches involving prespecified hypotheses. But note in passing that, whereas Stockmarr's hypotheses indeed do not depend on the data, they nonetheless do (unlike H'_p, H'_d) depend on the database—and so change as n grows. If we change the question, it is hardly surprising to find that the answer changes. What we need to do is to see how the answers to the same question are affected by changing n.

Dawid and Mortera (1996) have given a general analysis covering the problem at hand, as well as more general ones where, e.g., the search is conducted sequentially, terminating when the first match is found. They consider the whole family of hypotheses $\{\mathrm{H}_i : i \in \mathcal{P}\}$, where $\mathcal{P}$ is the population of all possible perpetrators (including Smith and the other members of $\mathcal{D}$) and H_i denotes the hypothesis $TP = i$. Since these hypotheses are independent of both the data and the database, they "discriminate against no person in particular" (Evett et al., 2000), and Stockmarr's distaste for data-dependent hypotheses becomes irrelevant.

The full outcome of the search procedure will be a random subset $U \subseteq \mathcal{D}$, consisting of just those individuals in $\mathcal{D}$ providing a DNA match with the crime trace. (Stockmarr considers a reduced outcome, the size of U. In the absence of any good argument for reduction, I prefer to use the full data available, U—without which we could not, in any case, identify a suspect! However, it can be shown that, under Stockmarr's simple assumptions, replacing the data set U by its size merely scales all likelihoods by an unimportant constant and so does not affect inferences or my arguments. This exact equivalence does not extend to more general formulations, but the differences will typically be small.)

We are specifically concerned with the case that the database search results in a single hit, so that the observation has the form $U = \{S\}$. The impact of this evidence is expressed by the associated likelihood function over the hypotheses. Under Stockmarr's assumptions, this is

$$L_i \propto \text{prob}(U = \{S\} \mid \mathrm{H}_i) = \begin{cases} (1-p)^{n-1} & (i = S) \\ 0 & (i \in \mathcal{D} \setminus \{S\}) \\ p(1-p)^{n-1} & (i \in \mathcal{P} \setminus \mathcal{D}). \end{cases}$$

Using some fixed individual $i_0 \in \mathcal{P} \setminus \mathcal{D}$ as a reference point to set the arbitrary constant of proportionality, we can thus take

$$L_i = \begin{cases} 1/p & (i = S) \\ 0 & (i \in \mathcal{D} \setminus \{S\}) \\ 1 & (i \in \mathcal{P} \setminus \mathcal{D}). \end{cases} \tag{1}$$

Comparing the likelihood functions (1) with that arising in the case where Smith was arrested without trawling a database (obtainable from (1) by setting $\mathcal{D} = \{S\}$), we see that the effect of the database search is just to eliminate the excluded individuals in $\mathcal{D} \setminus \{S\}$ while leaving other relative likelihoods unchanged. If we accept, as Stockmarr would appear to, that the impact of the identification evidence is fully embodied in the likelihood function over the relevant hypotheses, it is intuitively clear that these exclusions can only increase, by a common factor, the overall evidence against any of the remaining nonexcluded individuals—including Smith. Any likelihood-based argument that reaches a contrary conclusion must be faulty.

In an attempt to pinpoint this faulty logic, I now turn a closer analysis of Stockmarr's approach, restricting attention, as he does, to a comparison between just two hypotheses.

Stockmarr's hypotheses H_p ($TP \in \mathcal{D}$) and H_d ($TP \in \mathcal{P} \setminus \mathcal{D}$) are composite. It is not possible to construct a single likelihood for a composite hypothesis without further ingredients or assumptions. I therefore proceed by incorporating a 'prior' distribution Π over $\mathcal{P}$, to be understood as representing a juror's uncertainty about the identity of TP in the light of any non-DNA evidence in the case. The specific form of this prior distribution will be left unspecified. In fact, it will turn out that our analysis will only involve $\delta := \Pi(\mathcal{D})$, the prior probability that TP is in the database, and $\pi := \Pi(\{S\})$, the prior probability that Smith is our man.

The likelihood L_p for H_p is then given by

$$\begin{aligned} L_p &= \sum_{i \in \mathcal{D}} L_i \times \text{prob}(TP = i \mid TP \in \mathcal{D}) \\ &= \frac{1}{p} \times \frac{\pi}{\delta}. \end{aligned}$$

Since $L_i = 1$ for all $i \in \mathcal{P} \setminus \mathcal{D}$, the likelihood for H_d is $L_d = 1$, irrespective of the choice of Π. The likelihood ratio for comparing H_p with H_d, on observing $U = \{S\}$, is thus

$$L(\mathrm{H}_p : \mathrm{H}_d) = \frac{1}{p} \times \frac{\pi}{\delta}. \tag{2}$$

In the case that every member of $\mathcal{D}$ has, *a priori*, the same probability of being guilty, this reduces to $1/np$, in agreement with Stockmarr.

If instead, using the approach Stockmarr criticizes, we had calculated the likelihoods, based on the observation $U = \{S\}$, for the 'data-dependent' hypotheses H'_p ($TP = S$) and H'_d ($TP \neq S$), we would have obtained

$$\begin{aligned} L'_p &= \frac{1}{p} \\ L'_d &= \frac{1-\delta}{1-\pi}, \end{aligned}$$

yielding the likelihood ratio

$$L(\mathrm{H}'_p : \mathrm{H}'_d) = \frac{1}{p} \times \frac{1-\delta}{1-\pi}. \tag{3}$$

When Π is uniform over $\mathcal{P}$, so that each $i \in \mathcal{P}$ has the same prior probability of being guilty (the only prior specification considered by Stockmarr), this becomes $(1/p) \times (N-1)/(N-n)$, where N denotes the size of $\mathcal{P}$. More generally, (3) will be close to $1/p$ whenever $\delta \ll 1$, which will often be a reasonable assumption.

As Stockmarr is at pains to point out for his special case, (2) and (3) can be very different. But that is not surprising since they address different questions. We must instead address the overall impact of the DNA evidence on the issue at hand: the guilt or innocence of the suspect Smith.

After we have observed $U = \{S\}$, the two hypotheses H_p and H'_p become logically equivalent: we may term them conditionally equivalent. But they were not equivalent before the search. In particular, the respective prior odds in favor of H_p and H'_p (each against its contrary, H_d and H'_d) differ, being, respectively, $\delta/(1-\delta)$ and $\pi/(1-\pi)$. (When Π is uniform over $\mathcal{P}$, these become, respectively, $n/(N-n)$ and $1/(N-1)$, or approximately n/N and $1/N$ if $n \ll N$.)

Applying Bayes's Theorem, multiply the prior odds and likelihood ratio for H_p as against H_d to obtain the posterior odds on H_p, i.e.,

$$\frac{1}{p} \times \frac{\pi}{1-\delta}. \tag{4}$$

Now perform a similar calculation of the posterior odds on H'_p. This yields the identical expression (4). This is hardly surprising since the two hypotheses H_p and H'_p are conditionally equivalent: after observing $U = \{S\}$, they are saying the same thing and so must have the same posterior probability. The effect of working in terms of H_p rather than H'_p was simply to transfer a factor $\delta(1-\pi)/\pi(1-\delta)$ between the likelihood ratio and the prior odds. We learn from this that, when we do not fully specify which conditionally equivalent hypotheses are being considered, neither prior nor likelihood can be regarded as meaningful in themselves: only their combination, the posterior, which is insensitive to the specific formulation of the hypotheses, could be so regarded. In particular, one should avoid talk of "the likelihood ratio" as if this term des-

ignated a well-defined objective measure of evidence: at best, it can only be regarded as such relative to a chosen specification of the hypotheses. And it is not appropriate to compare such likelihood ratios across differing specifications of the hypotheses, even when these are conditionally equivalent, without simultaneously taking into account counterbalancing changes to the prior probabilities.

Stockmarr himself, in his equation (3), notes the invariance of the posterior of the framing of the hypotheses for the case that Π is uniform, when the factor transferred becomes $n(N-1)/(N-n)$, or approximately n for $n \ll N$. However, he fails to appreciate its significance.

In the legal case against Smith, the court is directly concerned with evidence in favor of H'_p: $TP = S$ against H'_d: $TP \neq S$. Stockmarr makes a fundamental logical error when he suggests that the court can replace these hypotheses by H_p and H_d and still use the resulting likelihood ratio as if it were directly relevant to the case against Smith. Superficially, this might seem reasonable since the new hypotheses are conditionally equivalent to the ones they replace. But we are not entitled to treat mere conditional equivalence on the same footing as full logical equivalence, and there is no justification for taking the likelihood ratio (2) in favor of H_p as a valid measure of evidence for the distinct hypothesis H'_p before the court—it is, simply, addressing a different issue. The posterior odds, on the other hand, which fully takes into account both the probabilistic and the logical import of the data, does address the identical issue under either formulation of the hypotheses. And when we apply Stockmarr's own analysis to calculate, in terms of posterior odds, the evidence in favor of his replacement hypothesis H_p, we find that the resulting answer agrees exactly with that of the workers he seeks to criticize.

It is generally regarded as desirable that expert testimony of statistical evidence be phrased in terms of likelihoods rather than posterior probabilities. Thus, Evett et al. (2000) state: "The weight of the evidence is represented by the likelihood ratio, *not* the posterior odds." I agree with this insofar as it implies that it should be left to the juror to assess and incorporate his or her own prior probabilities rather than have these supplied, explicitly or implicitly, by expert witnesses; but I disagree with the implicit suggestion that the phrase "the likelihood ratio" always has a clear and unambiguous meaning. When likelihoods are calculated in a nonstandard way, such as that proposed by Stockmarr, it is vital that the hypotheses being compared are carefully specified so that it is clear to the juror exactly which prior probabilities need to be assessed and incorporated. Once the limited and relative role of likelihood is appreciated, Stockmarr's criticisms are seen lacking in substance, and his reformulation of the problem in terms of 'data-independent hypotheses' as making absolutely no difference to the only thing that matters: the juror's posterior probability, in the light of all the evidence, that Smith is guilty. In view of this and the fact that jurors and judges may well have difficulty appreciating the subtleties involved in the correct treatment of Stockmarr's hypotheses, it seems appropriate to recommend the continued presentation in court of the likelihood ratio L'_p in favor of the hypothesis H'_p. Notwithstanding any data dependence, this is the correct factor to combine with the juror's prior odds that Smith is guilty in order to obtain the posterior probability of Smith's guilt.

ACKNOWLEDGEMENT

I am grateful to Marjan Sjerps and David Balding for valuable comments.

REFERENCES

Balding, D. J. and Donnelly, P. J. (1996). DNA profile evidence when the suspect is identified through a database search. *Journal of Forensic Science* **41,** 603–607.

Dawid, A. P. and Mortera, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society, Series B* **58,** 425–443.

Donnelly, P. and Friedman, R. D. (1999). DNA database searches and the legal consumption of scientific evidence. *Michigan Law Review* **97,** 931–984.

Evett, I. W., Foreman, L. A., and Weir, B. S. (2000). Letter to the editor. *Biometrics* **56,** 1274–1275.

National Research Council. (1996). *The Evaluation of Forensic DNA Evidence.* Washington, D.C.: National Academy Press.

Stockmarr, A. (1999). Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics* **55,** 671–677.

The author replied as follows:

Ironically, Dawid (2000) (henceforth APD) emphasizes the need for writing my paper Stockmarr (1999) (henceforth AS), as misunderstandings still exist. For the opportunity to discuss the misunderstandings, I welcome the letter. However, I do not welcome the extent of misquotations to which AS is subject. I simply must comment on a number of issues in APD.

The result of a DNA database search is, as APD writes, a random subset of the database consisting of the matches to what has been searched for, the profile of the true perpetrator, TP. But APD's formula (1) for the database search likelihood function does not incorporate this, like the paper that he elaborates on (Evett, Foreman, and Weir (2000) and their references). When APD writes that his formula (1) is calculated "under Stockmarr's assumptions," then it simply isn't correct. If the profile of the true perpetrator is not in the database (APD indexes the contents of the database by the persons that donated the samples rather than by the profiles themselves) and a single match is observed, then the match is obtained by chance. The probability that the random subset, the person S, has the matching profile A that is searched for is not p but one. That is, if S exists at all as a unique match in the database, the probability of that is $np(1-p)^{n-1}$, and we arrive at the expression for the probability of the data that was derived in AS, conditionally on the profile of the TP. APD's formula (1) corresponds to independence between the DNA profiles of S and that of the TP, and APD arrive at something different. Thus, he thinks of S as the fixed [S]mith rather than the random [S]uspect. While a fixed, prespecified person may be thought of as having a profile that is stochastically independent of the profile of the TP, a person that is selected from a database search on the grounds that his profile matches that of the TP will not have this independence relation. The model that APD's formula (1) corresponds to is that [S]mith is singled out in forehand, and this does not go

along very well with APD's claim that he considers prespecified hypotheses only. As APD does not know the identity of [*S*]mith prior to the search, his model that corresponds to (1) is not a proper description of the database search experiment, regardless of his formulation in single-person hypotheses, because it merely models the profiles of the persons and does not take into account that the matching person is selected as a suspect, regardless of the name (i in APD) of the person.

There are two issues that must be considered when capturing the effect of a database search. The first is that multiple comparisons with a number of persons in the database increase the probability of a match by chance. This relates to the DNA typing of the individuals. The second is that the persons in the database excluded by the search are eliminated as possible perpetrators, decreasing the set of these. This relates to the sampling process leading to the database. APD's analysis only deals with the second issue, although the effect of this is marginal (APD's formulation) unless the database consists of a considerable part of the possible perpetrators. The first issue, however, decreases the weight of the evidence with the inverse of the database size. DNA databases are at present at a size on the order of 10^5 and growing, so the impact is huge.

The diverging expressions for the probability of the data and what the data actually are is the real difference between APD and AS, and I am surprised that APD does not state this more clearly. The difference is absolutely not, as APD suggests when he writes that his equation (2) is "in agreement with Stockmarr" under uniform priors on the members of $\mathcal{D}$, a question of which priors to choose. APD's formula (2) has inherited his problematic likelihood (1), and that his formula (2) has the same value as the database search likelihood ratio, if uniform priors apply, does not make it "in agreement with Stockmarr"; it is a different formula based on a different concept. The database search likelihood ratio derived in AS is conceptually free of prior probabilities, which APD's formula (2) relies on completely.

APD continues by claiming that the hypotheses H_p and H_d presented in AS are composite. This is certainly a statement that only refers to APD's formulations in judicial hypotheses about which person the TP is rather than the formulation in statistical models for DNA profiles that is being used in AS. In the context in AS, his claim is simply not correct. The formulation of the hypotheses in AS involves the selection of a unique probability measure under each of the two hypotheses, which are not formulated as Dawid translates them, "$TP \in \mathcal{D}$" and "$TP \in \mathcal{P} \setminus \mathcal{D}$," but as distributional descriptions. It is an essential part of the evaluation of DNA evidence in a forensic context that you can present the evidence as the probability of obtaining it under two competing sets of circumstances (hypotheses), i.e., that you are dealing with a simple hypothesis versus a simple alternative. This is so in AS but not so in the approach by APD.

APD then argues that the posterior odds of $\mathrm{H}_p : \mathrm{H}_d$ and $\mathrm{H}'_p : \mathrm{H}'_d$ are identical and suggests that the common expression is a proper approach to the decision problem of the court. The first is, of course, in the setting of AS, incorrect. APD claims that I note the invariance of the posterior for the two formulations under uniform priors. This is not correct. With the notation of APD, the posterior odds of $\mathrm{H}_p : \mathrm{H}_d$ are $(1/np)[\delta/(1-\delta)]$, while an experiment corresponding to H'_p and H'_d being true hypotheses and with the same single match as result would yield the posterior odds $(1/p)[\pi/(1-\delta)]$. These expressions agree if uniform priors apply, but in general, they are not equal, and I do not discuss that in AS. That the posteriors are equal in the setting of APD is, on the other hand, obvious because the mentioned independence relation in APD's formula (1) yields that his two models basically model the same thing. This is not the database search experiment, however, and probabilistic use of DNA profile data must be accompanied by a proper description of how they are obtained.

APD discusses conditional equivalence of hypotheses and notes that H_p and H'_p are conditionally equivalent, in the sense that conditional equivalent hypotheses correspond to the same models after the experiment, or rather, this experiment. If [*S*]eaman had turned out to match instead of [*S*]mith, then a new conditionality principle would have to be formulated. However, in the setup of AS, H_d and H'_d are not conditionally equivalent regardless of which identity [*S*] reveals; they leave the same possibilities open for who the true perpetrator is, but they model the data differently. I find it hard to use this concept as an argument for the use of H'_p and H'_d, as the concept is just as data dependent as $\mathrm{H}'_{p/d}$. That APD wishes to use a Bayesian framework to evaluate forensic DNA evidence is part of an ongoing discussion, his view is one out of at least two and is disputed, as I described in Stockmarr (2000); this is further discussed by Roeder (1994) in her response to Balding, Donnelly, and Nichols (1994). One should note that the spirit in APD, that his approach is established and the approach of AS is nonstandard, is far from the truth. APD's analysis is in conflict with the recommendations of the U.S. National Research Council (National Research Council, 1996), and, not surprisingly, I find APD rather than AS as the controversial part in this, and that APD makes the logical error rather than AS by abandoning the concept of describing the data by a proper statistical model and formulating the results in a likelihood ratio that relates to two competing circumstances. Instead, APD recommends the use of data-dependent statements through L'_p (or, to put it in other words, statements invented for the occasion, which is what H'_p and H'_d are) as hypotheses for the evaluation of the evidence. Data-dependent statements do not make sense as statistical hypotheses, and I have described in AS how they may lead to improper presentations of the evidence in this situation. If the decision makers (jurors) do not wish to assert prior probabilities (I quote Roeder (1994): "The process by which jurors (or justices) reach a decision is complex, and formal probability arguments undoubtedly never enter the process"), they are faced with a situation that resembles the situation from AS, pages 676: You search a database the size of a million for a profile with a profile probability of one in a million. The example was repeated in my previous response to Evett, Foreman, and Weir (2000), and I have not found the answer to nor a discussion of the appropriateness of recommending an evidence weight of 1,000,000:1 in a situation that occurs more than one out of three times if the [*S*]uspect is innocent and exactly the same number of times if (s)he is the TP, from APD.

References

Balding, D. J. and Donnelly, P. (1996). Evaluating DNA profile evidence when the suspect is found through a database search. *Journal of Forensic Science* **41,** 603–607.

Balding, D. J., Donnelly, P. D., and Nichols, R. A. (1994). Some causes for concern about DNA profiles. Comments in Roeder, K. (1994). DNA fingerprinting: A review of the controversy. *Statistical Science* **8,** 222–278.

Dawid, A. P. (2001). Letter to the editor. *Biometrics* **57,** 979–982.

Evett, I. W., Foreman, L. A., and Weir, B. S. (2000). Letter to the editor. *Biometrics* **56,** 1274–1275.

National Research Council Committee on DNA Forensic Science. (1996). *An Update: The Evaluation of Forensic DNA Evidence.* Washington, D.C.: National Academy Press.

Roeder, K. (1994). DNA fingerprinting: A review of the controversy. *Statistical Science* **8,** 222–278.

Stockmarr, A. (1999). Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search. *Biometrics* **55,** 671–677.

Stockmarr, A. (2000). Rejoinder. *Biometrics* **56,** 1275–1276.

ANDERS STOCKMARR
DSI Danish Institute for Health Services Research
P.O. Box 2595, Dampfærgevej 22
DK-2100 Copenhagen, Denmark
and Department of Biostatistics
University of Copenhagen
email: ast@dsi.dk

Letter to the Editors

From: Stephen D. Walter, Ph.D., Professor
Clinical Epidemiology and Biostatistics
McMaster University
Hamilton, Ontario, Canada

To the Editor of Biometrics:

Dr Carroll's analysis of the long periods of time required to review biostatistics papers is timely and sobering. I (along with many others, I am sure) have often suffered through this tortuous process, as the following two examples might illustrate.

In the first case, I recently submitted a paper to *Biometrics* and received the usual acknowledgement, stating that I should expect to have reviews "in 2 to 6 months." Having heard nothing more than 7 months after submission, I contacted the *Biometrics* office and was told to expect a decision in about 2 weeks. After further delays and communications back and forth, I eventually received my long-awaited result—9 months after my submission. It consisted of a single review, 1/2 page long, containing 6 relatively minor comments. On average, the reviewer had managed to write approximately one line of text for every 4 weeks since I submitted the paper; but even that was better than the reviewer who never responded at all.

I must mention that there were also some helpful editorial remarks from Dr Carroll, with (to his credit) an offer to handle the paper personally if I chose to revise and resubmit. However, on balance, I felt that it would not be worthwhile for me to invest possibly another year of my life in this effort, and so I decided to try another journal.

The second case involved another well-known statistics journal, that I will not name here. At the request of the editor, my paper was submitted electronically, followed by a paper copy through the regular mail. After approximately 9 months, I learned that the editor had received two referee reports stating that they were both unable to understand my paper and that something seemed to be "missing." It transpired that the editorial office had printed off my electronic submission and distributed it to the referees, but a computer gremlin had suppressed the printing of all symbols and equations, leaving large swaths of empty space in the manuscript they were asked to review! Even a cursory examination of the first page should have revealed a problem before the paper was sent out for review.

I suspect that these are not isolated examples. In both cases, it was the author (me!) who suffered long delays at the hands of an inefficient and error-prone system. I believe the solution to these difficulties will involve (a) provision of adequate funding and other resources to the editor and his staff to prevent administrative hold-ups and errors of this kind and (b) a personal commitment by reviewers to do their job promptly or to turn down the assignment if they cannot do so. Peer review is an imperfect process, but it is the best we have at the moment, and as a profession, we *need* to make it work.

Editor's Note:

Dr Walter's personal anecdotes only serve to reinforce the points made by Dr Carroll and to remind us that the culture and practice of reviewing in our profession must be improved, not only out of respect for the authors who submit their work to our journals in good faith but to ensure the timely dissemination of important contributions so that they may be put to immediate use advancing our science and that of the applications with which we are involved.

Dr Walter identifies two main components that contribute to lengthy review times. Point (a) is one that may be addressed at the journal level. As Dr Carroll notes, *Biometrics* has in place a thorough system to track the progress of reviews for each paper we receive and has a superb editorial assistant, Ms Ann Hanhart, who not only is adept at circumventing administrative problems like that mentioned by Dr Walter but who keeps a detailed database on all submissions. Ms Hanhart sends regular reminders to associate editors for each manuscript they handle and keeps coeditors continually informed of the status of outstanding manuscripts so that we may intervene in difficult cases before too much time is lost. We believe that such an administrative system is essential, and we encourage all journals to adopt a similar strategy as well as to publish statistics on review times.

Dr Walter's point (b) emphasizes that administration can only go so far—shortened review times, not to mention thoughtful, helpful, quality reviews, are the responsibility of all of us. Experiences such as that Dr Walter cites with *Biometrics* can only be eliminated by a widespread commitment by everyone involved with the editorial process to alter the culture of reviewing in our profession that has made waiting months to submit a review or, worse, failing to respond at all,

an almost "acceptable" practice. At *Biometrics*, we define "time to review" as time from receipt of the paper in our office until time the reviews are mailed to the author, as we cannot hope to keep track of the additional time these materials spend in transit (e.g., by postal mail), which can sometimes themselves be substantial. Using this convention, it is only mildly satisfying to report that, of the 1490 papers submitted to *Biometrics* from February 1, 1997–December 31, 2000, of which Dr Walter's paper was one, 92, or 6%, experienced a time to initial review longer than 6 months, and 6 of these 1490 took longer than the approximately 8 months from receipt of the paper to mailing of the reviews for Dr Walter's submission. Although these data suggest Dr Walter's review time with *Biometrics* is unusual, we believe that such lengthy times to review simply should never happen with any statistics journal and should indeed be viewed as unacceptable by the profession. Dr Walter also notes that the quality of the review (when he did finally receive it) was neither useful nor consistent with the length of time taken to generate it. Lengthy times to review are often accompanied by less-than-helpful comments, as often the review is obtained only after considerable badgering of reviewers. As a profession, we should also be unwilling to accept reviews that are not thoughtful, helpful, and respectful of the time the author has devoted to his/her work.

Biometrics is committed to addressing these problems and playing a role in effecting a positive change of culture of reviewing in our profession. We are grateful to our many associate editors and referees who are committed likewise. We are currently exploring new approaches to our editorial structure that may involve more people who agree to provide quality reviews in a timely fashion. We would like to encourage strongly all members of the statistical profession to consider the implications of continuing to tolerate lengthy review times and less-than-careful reviews and to take an active role in changing the culture at the individual level.