Transactions of the American Fisheries Society 116:232–243, 1987

Use of a Log-Linear Model with the EM Algorithm to Correct Estimates of Stock Composition and to Convert Length to Age

JOHN MAURICE HOENIG

Rosenstiel School of Marine and Atmospheric Sciences
Cooperative Institute for Marine and Atmospheric Studies, University of Miami
4600 Rickenbacker Causeway, Miami, Florida 33149, USA

DENNIS M. HEISEY

NH Analytical Software, Post Office Box 13204, Roseville, Minnesota 55113, USA

Abstract.—The EM (expectation–maximization) algorithm was used to develop a general procedure for finding maximum likelihood estimates of population proportions when some observations cannot be assigned unambiguously to a population category. The method can be used to estimate the age composition of fish from length frequencies, to adjust biased estimates of age composition (e.g., scale ages that tend to be too low), and to correct biased estimates of unit stock composition. To implement the method, two samples are obtained. In the first sample, the items are cross-classified by their actual identity and by a second (possibly error-prone) surrogate classifying variable. In the second sample, the items are classified by only the surrogate variable. The information in the two samples is then used to estimate the population proportions in the second sample.

A variety of seemingly unrelated problems in fisheries science can be shown to be special cases of a general problem involving partially classified contingency tables. Consequently, these problems can be solved by a single, simple statistical method. Our work in this area was motivated by three important tasks in fishery science which serve to illustrate the generality of the basic problem.

The first task involves correcting error-prone estimates of stock composition. Suppose some classification rule is developed for identifying fish and the rule then is applied to known samples (e.g., fish collected from their spawning grounds). The resulting information can be used to estimate (mis)classification rates which, in turn, can be used to correct the results of a survey in which the classification rule is applied to animals of unknown identity.

The second task involves aging a sample of fish in order to cross-classify the animals by age and length. The resulting classification rates can be used to estimate the age composition of any population from its length-frequency distribution (provided that growth rates and gear selectivities do not vary among samples).

The third task is to correct the results of a growth study in which the age of animals is determined by an error-prone technique. For example, counting annuli on fish scales tends to underestimate the age of old animals. If some animals are aged by both the scale method and a more reliable method (such as otolith rings), then the estimate of age composition obtained from scale readings can be adjusted to the more reliable categorization.

A structure shared by these problems is that two samples are obtained. In the first sample, all items are completely cross-classified by two classifying variables. We will call the first (row) variable the accurate classifier and the second (column) variable the error-prone or, more generally, the surrogate classifier. In the second sample, the items are classified by only the error-prone classifier. If there are I accurate categories and J error-prone categories, the data can then be represented as a partially classified $I \times J \times 2$ contingency table (Figure 1).

In this paper, we describe the structure of the basic problem as a particular log-linear model associated with the partially classified contingency table. We then develop a procedure to find maximum likelihood estimates for the proportions in the unobserved accurate categories in the second sample by use of the EM (expectation–maximization) algorithm (Dempster et al. 1977). The procedure is illustrated and compared to existing methods in the literature described by Pella and Robertson (1978), Clark (1981), Bartoo and Parker (1983), and Cook (1983).

A fundamental distinction between our meth-

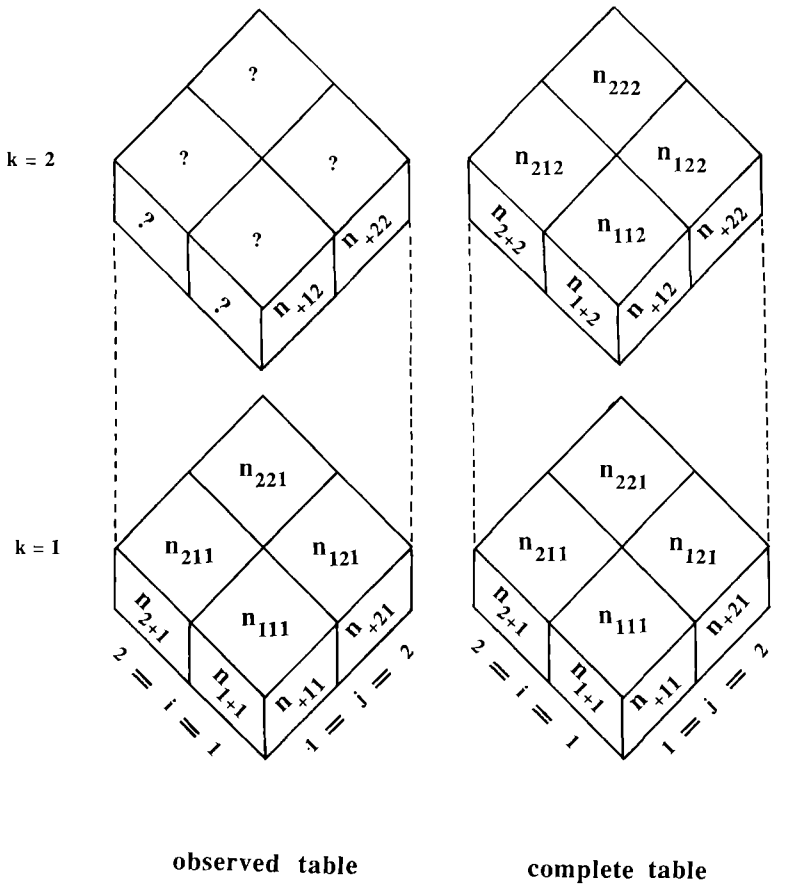

FIGURE 1.—Representation of observed data as a partially classified $I \times J \times 2$ contingency table (left), and appearance of the contingency table if the data were completely classified (right). Sample number is denoted by k.

od and previous methods is that we utilize information in the second, incompletely classified, sample to help estimate classification rates. Others have assumed that these rates are perfectly known from the first sample. Of course, this ignores sampling errors in the first sample, which may be substantial if the sample sizes are not large. Our method allows one to examine the assumption of identical classification rates in the first and second sample. We emphasize that our method does not require the same accurate or surrogate population distributions in the populations from which the first and second samples were drawn.

Structure of the Problem

Throughout this section and the next, we will use as an example the estimation of an age-frequency vector (the "accurate" classification) from a length-frequency vector (the "surrogate" classification). Suppose an animal can fall into one of I age categories, indexed by i, and one of J length categories, indexed by j. Two samples are taken. In the first sample ($k = 1$), all animals are cross-classified by age and length. In the second sample ($k = 2$), only the lengths are observed (Figure 1, left).

The same estimation procedure may be used for a variety of sampling designs. The second sample may result from multinomial sampling, where the total sample size was fixed before sampling, or from Poisson sampling, where the total was a random variable. The first sample may result from product-multinomial, multinomial, or Poisson sampling. With product-multinomial sampling, the age marginal totals are preset. This type of sampling is generally not of interest for the length-to-age conversion problem described here. However, it may be desirable for the other examples discussed later to ensure adequate sampling of all categories.

We will assume that the probability that an animal of age i falls in length category j remains constant from sample to sample. That is,

$$P(j|i)_{k=1} = P(j|i)_{k=2}; \quad (1)$$

$P(j|i)_{k=1}$ is read as the "probability of length j given age i in sample 1." These probabilities are known as the classification rates. We will not assume that the two samples come from populations with the same age composition. As will be seen later, this distinguishes our approach from the classic age–length key, for which it is assumed that the distribution of age about length remains constant from sample to sample (Kimura 1977; Westrheim and Ricker 1978).

Let a_{ik} be the probability that an animal from sample k is in age class i. The probability that an animal in sample k is cross-classified as i, j, denoted by $P(i, j|k)$, is given by

$$P(i, j|k) = P(j|i)a_{ik}. \quad (2)$$

This structure can be seen to correspond to the hierarchical log-linear model (Fienberg 1980) given by

$$\log(\text{cell count}) = \mu + A_i + L_j + S_k + (A*L)_{ij} + (A*S)_{ik};$$

μ is the grand mean, A, L, and S are main effects for age, length, and sample, respectively, and an asterisk indicates an interaction between two factors. Each term in the log-linear model is subject to the usual constraint that it must sum to zero over any subscript (Fienberg 1970, 1980).

Maximum Likelihood Estimation

Maximum likelihood estimation consists of forming the likelihood function which describes the likelihood of obtaining the observed results in terms of the unknown parameters. The maximum likelihood (ML) estimates are those values of the parameters which maximize the likelihood of obtaining the observed data, i.e., which maximize the likelihood function.

The EM algorithm of Dempster et al. (1977) is a convenient procedure for obtaining maximum likelihood estimates for incomplete (e.g., grouped, censored, or truncated) data. Instead of writing a complicated likelihood function for the incomplete data, one works with the (generally) simpler likelihood function for the complete data. The procedure consists of first guessing at the expected values of the missing data, next computing the maximum likelihood (ML) estimates based on the now complete data (maximization or M step), and then using the ML estimates to revise the estimates of the missing data (expectation or E step). The procedure is iterated by alternating E and M steps until convergence is achieved. The EM algorithm was applied to partially classified contingency tables by Chen et al. (1984) and Espeland and Odoroff (1985). However, they assumed that the "accurate" population composition does not vary among samples. Chen and Fienberg (1976) developed a general theory for model building with partially classified categorical data but did not explicitly deal with the particular type of problem and the applications considered here.

For the length-to-age conversion problem, the EM algorithm proceeds as if both samples had been completely classified (Figure 1, right). Let n_{ijk} be the number of individuals in sample k classified as i, j. Of course, the n_{ij2} data are not actually observed. Initially, any values can be assigned to the n_{ij2} cells as long as the sums over i are equal to the observed length frequency in the second sample, n_{+j2} (the + sign in the subscript denotes summation over the variable i; e.g., $n_{+j2} = \sum_i n_{ij2}$). The kernel of the log-likelihood for the now complete data is (see Appendix A)

$$\begin{aligned} L &= \sum_{i=1}^{I}\sum_{j=1}^{J} \{n_{ij1}\log[P(j|i)a_{i1}] \\ &\qquad + n_{ij2}\log[P(j|i)a_{i2}]\} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} (n_{ij1} + n_{ij2})\log P(j|i) \\ &\quad + \sum_{i=1}^{I} n_{i+1}\log a_{i1} \\ &\quad + \sum_{i=1}^{I} n_{i+2}\log a_{i2}. \end{aligned} \quad (3)$$

The maximum likelihood estimates from the above likelihood function are (Fienberg 1970)

$$\hat{P}(j|i) = (n_{ij1} + n_{ij2})/(n_{i+1} + n_{i+2}); \quad (4a)$$

$$\hat{a}_{i1} = n_{i+1}/n_{++1}; \quad (4b)$$

$$\hat{a}_{i2} = n_{i+2}/n_{++2}. \quad (4c)$$

Under this model, estimated expected cell counts are

$$\hat{m}_{ijk} = \hat{P}(j|i)\hat{a}_{ik}n_{++k} = \hat{P}(j|i)n_{i+k}. \quad (5)$$

Finding the $\hat{m}_{ijk}$ values is the final step of the M step.

For the E step, updated values of the cells in the second sample are found by

$$n_{ij2} = \hat{m}_{ij2}n_{+j2}/\hat{m}_{+j2}. \quad (6)$$

This is simply a rescaling of the fitted values (equation 5) so that the sums over i agree with the observed length marginals in the second sample (see Espeland and Odoroff 1985).

The M and E steps are repeated in sequence until adequate convergence is achieved. Convergence can be examined by noting successive changes (or percent changes) in the log-likelihood function or the parameter estimates.

The above procedure can be simplified somewhat by the following considerations.

(1) Estimates of a_{i1} do not change from cycle to cycle and thus do not affect the maximization of the kernel of the log-likelihood. Hence, the a_{i1} values do not need to be computed and can be left out of the log-likelihood function.

(2) A computationally more convenient form of equation (6) is

$$n_{ij2} = \hat{P}(j|i)n_{i+2}n_{+j2} \Big/ \sum_{i=1}^{I} [\hat{P}(j|i)n_{i+2}]. \tag{7}$$

This eliminates the need to compute $\hat{m}_{ijk}$ values.

When some cell counts are zero, problems may be encountered in evaluating the log-likelihood function and also in estimating the variance–covariance matrix (see below). In this case, we recommend adding a small number (e.g., 10^{-4}) to all counts.

A flow chart for the entire procedure is presented in Figure 2. Note that a superscript notation has been added to denote quantities which are continually updated. Variance–covariance estimation based on the method of Louis (1982) is described in Appendix B.

Example

Suppose a sample of fish is aged, giving rise to the following table[1]:

	Length category					
Age	1	2	3	4	5	6
1	50	30	20	0	0	0
2	0	50	30	20	0	0
3	0	0	50	30	20	0
4	0	0	0	50	30	20

A second sample of fish is obtained, with the same sampling gear, from an area where fish are known to grow at the same rate as the fish in the first sample. Thus, the principal assumption, embodied in equation (1), is met. The second sample has the following length frequency for categories 1–6:

$$\text{length frequency} = [60\ 40\ 0\ 0\ 40\ 60].$$

The first step is to guess at the values of the age–length classification matrix for the second sample, n_{ij2}. A convenient way to do this is to apportion the lengths to age-classes in the same proportions as in the first sample, i.e.,

$$n_{ij2} = n_{+j2}n_{ij1}/n_{+j1},$$

provided all of the n_{+j1} values are nonzero.

The next step is to compute the ML estimates of $P(j|i)$ and a_{i2}. Given the ML estimates, it is possible to update the estimates of n_{ij2} with equation (7). Given the updated values of n_{ij2}, one can recompute the ML estimates, then recompute the values of n_{ij2}, etc. until convergence is adequate.

With this procedure, the ML estimates of the age composition in the second sample were found after about 30 iterations to be[2]

$$\text{age composition} = [100\ 0\ 0\ 100].$$

Model Evaluation

The main assumption of our procedure is described by equation (1); the misclassification rates are the same for samples 1 and 2. This can be tested with either a likelihood ratio or Pearson chi-square goodness-of-fit test with $J - I$ degrees of freedom. However, our experience suggests that these tests generally would not be very powerful. We found it more instructive to construct standardized residuals, computed as

$$r_{ij} = (n_{ij1} - \hat{m}_{ij1})/\sqrt{\hat{m}_{ij1}}.$$

For example, a large positive residual would suggest that, on the basis of the second sample, fewer items would be expected in the i, j cell of the first sample. Numerous residuals with large absolute values, or a systematic pattern in the residuals, would cast doubt on the assumption in equation (1).

[1] These data were provided by W. Clark (Washington State Department of Fisheries, Seattle, personal communication) to enable users to test and examine the performance of his restricted least-squares computer program for estimating age composition from length data.

[2] In this case, the procedure converges to a boundary solution, i.e., some of the estimated age proportions a_{i2} are zero. Haberman (1974) pointed out that, in this case, unique maximum likelihood estimates do not exist. This is largely of theoretical interest rather than of practical concern.

I. Guess at initial values for second sample cells, e.g.

$$n_{ij2}^{(0)} = n_{+j2}\, n_{ij1}/n_{+j1}$$

$$s = 0$$

II. M Step

(a) $$P(j|i)^{(s+1)} = (n_{ij1} + n_{ij2}^{(s)}) / (n_{i+1} + n_{i+2}^{(s)})$$

(b) $$a_{i2}^{(s+1)} = n_{i+2}^{(s)} / n_{++2}$$

III. E Step

$$n_{ij2}^{(s+1)} = \frac{P(j|i)^{(s+1)} n_{i+2}^{(s)} n_{+j2}}{\sum_i [P(j|i)^{(s+1)}\, n_{i+2}^{(s)}]}$$

IV. Compute the kernel of the log-likelihood, L, for nonzero estimates of $P(j|i)$ and a_{i2}

$$L = \sum_i \sum_j (n_{ij1} + n_{ij2}^{(s+1)}) \log P\,(j|i)^{(s+1)} + \sum_i n_{i+2}^{(s+1)} \log a_{i2}^{(s+1)}$$

V. Evaluation step

let $s = s + 1$
(a) If first cycle, then set L' = L, and to to II.
(b) else if $|L' - L| <$ tolerance, then stop.
(c) else, set L' = L, and go to II.

FIGURE 2.—The EM algorithm. Superscripts denote the cycle in which an iterated quantity is defined.

Discussion

The method of maximum likelihood estimation has desirable asymptotic properties when certain very general regularity conditions can be met (Bury 1975), as is the case for the problems considered here when the parameters are all positive. These asymptotic properties include minimum variance, unbiasedness, and normality. Several alternative approaches to the correction–conversion problem have been described in the fisheries literature. These approaches are reviewed below.

Age–Length Key Problem

Traditionally, age composition has been estimated from length-frequency data by use of the age–length key method developed by Fridriksson (1934). The basic idea is to stratify a large sample by length and to determine the age composition in a random subsample from each interval. The age–length data from the subsamples make up the key; the key is equivalent to what we call the first sample. The total number of animals at any given age i is taken to be the sum, over all L length intervals, of the number of animals in the interval (N_l) times the estimated proportion (from the key) that are age i (p_{li}), or

$$N_i = \sum_{l=1}^{L} N_l p_{li}.$$

When the assumptions of this method can be met, the classic key should be preferable to our more general model because fewer parameters need to be estimated. Further statistical development of this approach is presented in Tennenbein (1970, 1971).

Kimura (1977) and Westrheim and Ricker (1978) recognized the serious limitation that a classic age–length key can only be applied validly to samples from a population with the same age composition and growth rates as the one from which the key was derived, and only if the gear selectivity for the samples is the same. Clark (1981) and Bartoo and Parker (1983) made an important advance in technique by showing that it is possible

to avoid the assumption of constant age composition from sample to sample. Instead of working with the distribution of age about length, which depends on the age composition, they used the distribution of length about age, which is independent of the age composition. We call this kind of approach an inverse age–length key. They formulated the problem in matrix algebra as a linear model given by

$$\mathbf{L} = \mathbf{PA};$$

L and **A** are the length- and age-frequency vectors and **P** is a matrix of transition probabilities in which the elements p_{ji} represent the proportion of animals of age i which fall into length class j. Bartoo and Parker (1983) solved this system of equations by ordinary least squares. Clark (1981) pointed out that the ordinary least-squares solution can give infeasible estimates (i.e., negative proportions) and proposed a restricted least-squares approach in which each proportion is restricted to non-negative values. No method of estimating variances and covariances was provided in either paper. Both approaches assume that the **P** matrix is known exactly and that the only variability (random error) occurs in the length-frequency vector from the second sample. In practice, the length-frequency vector is usually based on a large sample and is thus known quite precisely. In contrast, the **P** matrix is usually based on a small number of age determinations, because age determinations are tedious to perform, and is thus subject to uncertainty. A basic assumption of the least-squares approaches is, therefore, called into question.

A consequence of the least-squares approach to the problem formulation is that the estimates do not depend on the relative sizes of the two samples. Since the classification matrix (based on sample 1) is assumed to be known perfectly, changes in the relative sizes of the two samples do not lead to a different weighting scheme for the information in the two samples.

The maximum likelihood approach, in contrast, allows for uncertainty in all of the data, the degree of uncertainty depending upon the sample sizes. This model is thus more realistic than the least-squares models. When the model fits the data exactly, the least-squares approaches give maximum likelihood estimates. That is, the least-squares estimates are maximum likelihood if it is possible to satisfy simultaneously the conditions that:

(1) the fitted values for the second half of the table ($k = 2$) are all non-negative and add up to the observed surrogate variable totals; and

(2) the fitted error rates in the second half of the table, i.e., fitted $P(j|i)_{k=2} = n_{ij2}/n_{i+2}$, are each equal to the corresponding observed error rates in the first half of the table ($=n_{ij1}/n_{i+1}$).

In other cases, the two approaches can give significantly different results. For example, the restricted least-squares estimates for the age-distribution example above are

$$[123.6\ 0\ 0\ 76.4],$$

whereas the maximum likelihood estimates are

$$[100.0\ 0\ 0\ 100.0].$$

The restricted least-squares estimates appear unreasonable since there are 100 small fish and 100 large fish, and the classification matrix indicates that no large fish are age 1, yet the predicted age composition implies that some large fish are, indeed, age 1.

As this paper was being submitted for publication, we received a manuscript (Kimura and Chikuni, 1987) which describes another approach to applying the EM algorithm to the age–length key problem. Kimura and Chikuni obtained maximum likelihood estimates for a different model in which it is assumed that there are no sampling errors in the first sample. If the size of the first sample is large, the two methods will give similar results.

Correcting Estimates of Stock Composition

Pella and Robertson (1978) developed a procedure for correcting biased estimates of stock composition which generalizes the results of several earlier studies in the literature. Suppose an error-prone classification rule is developed and applied to fish of known identity. This gives rise to a matrix of classification rates, **P**, whose elements p_{ji} estimate the probability that an animal from stock i is classified as stock j. If the error-prone rule is applied to a sample of fish whose identities are unknown, it follows deterministically that the observed, error-prone stock composition vector, **E**, is related to the actual stock composition, **A**, by the matrix equation

$$\mathbf{E} = \mathbf{PA}.$$

Consequently, the actual stock composition vector can be estimated as

$$\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{E}.$$

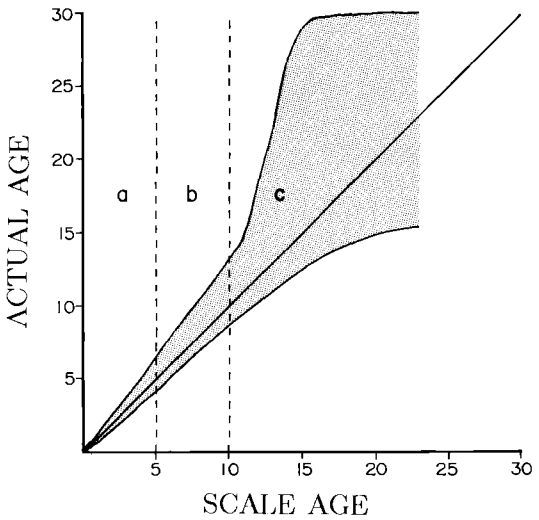

FIGURE 3.—Typical relationship between scale age and actual age for a long-lived fish species. In region "a," there is a close correspondence between actual and estimated ages; in region "b," there is a systematic underestimate of age by the scale method which can be accounted for with a bias correction procedure; in region "c," the relationship between scale age and actual age is so tenuous that the bias correction procedure will be ineffectual.

Variance–covariance estimation was accomplished by the delta method. The same matrix inversion procedure was used by Greenland and Kleinbaum (1983) to correct for misclassification errors in diagnostic test results obtained in epidemiological surveys.

Cook (1983) pointed out that the P-inverse method of Pella and Robertson sometimes produces infeasible (negative) estimates. Cook suggested for this case, with little statistical justification, that one should take as the estimate the feasible solution which is the shortest Euclidean distance from the P-inverse solution.

The EM algorithm described in our paper can be used to obtain maximum likelihood estimates of the actual stock composition. The only alterations required are to replace the words "length" with "error-prone" and "age" with "actual" in the above descriptions. If product multinomial sampling is used for the first sample, the actual marginals rather than the error-prone marginals must be fixed.

It can be shown (Hoenig and Heisey 1986) that whenever the P-inverse method gives feasible estimates, those estimates are maximum likelihood. When the P-inverse method gives infeasible results, Cook's (1983) method does not, in general, return maximum likelihood estimates except for the case where the population consists of just two stocks.

The choice of method can be significant. For example, if the classification data matrix is as follows:

	Classified		
Actual	A	B	C
A	10	20	0
B	0	10	20
C	10	20	30

and the error-prone estimates from a survey are

$$\mathbf{E}^T = [30\ 20\ 10],$$

then the three methods produce the following adjusted estimates:

Method	Stock A	Stock B	Stock C
P-inverse	0	−120.0	180.0
Cook	0	0	60.0
Maximum likelihood	41.5	0	18.5

Adjusting Error-Prone Estimates of Age Composition

It often happens that an investigator is faced with a choice of methods for aging animals. For example, fish scales can be obtained easily without killing the fish and are relatively easy to process. Otoliths and other internal calcified structures can only be obtained at the expense of killing the fish, and generally require more effort to process. However, internal structures generally provide more reliable estimates of age, particularly for older (larger) animals (e.g., Casselman 1983; Barnes and Power 1984; O'Gorman et al. 1987).

The same procedure described above for correcting biased estimates of stock composition can be used for the biased age problem. In the first sample, animals are classified to age by both the (more) reliable method (otoliths) and the error-prone method (scales). In the second sample, only the error-prone method is used. It should be noted, however, that this procedure is not a cure-all and will not provide reliable estimates for the oldest age classes of long-lived species. This is because, above a certain age, the error-prone method may break down to the point that very little information is provided about actual age. For example, it is apparent in Figure 3 that animals clas-

sified as being age 15 by the scale method may be anywhere from 12 to 30 years old. The same is largely true of animals classified as age 20 by the scale method. Thus, there is little information with which to distinguish the true ages of these animals on the basis of their apparent (scale) age. This will be reflected in the portion of the covariance matrix referring to the age composition of the oldest age groups. The correction method will, therefore, be useful only over a portion of the age range and can serve only to extend the usefulness of the scale-aging method. It will not obviate the need to use otoliths for the oldest animals in a long-lived stock.

Other Applications

The procedure presented here should be useful for a wide variety of applications. Survey response errors can be estimated by follow-up studies or other forms of independent corroboration and then used to correct survey results. For example, deer harvest registration in Minnesota was monitored by having biologists make spot checks at registration stations. A systematic bias in reporting was detected and accounted for by the correction procedure described here (Hoenig and colleagues, personal observations). Similarly, misclassification errors in disease surveys can be accounted for if one knows the error properties of the diagnostic test used (e.g., Greenland and Kleinbaum 1983). In population genetics studies, genotypic frequencies can be estimated from phenotypic frequencies if an appropriate conversion procedure can be developed. The estimation of unit stock composition is an example. If some animals are cross-classified by results of electrophoretic examination (i.e., by genotypic traits) and morphometric–meristic analysis (i.e., based on phenotypic and genotypic traits), then population composition can be estimated from a survey of the phenotypic traits (see also Haberman 1974). Finally, a variety of error-prone techniques is available for classifying animals to sex, species, age, and maturity (Hoenig and Heisey 1984). Estimates of proportions derived by these methods can be adjusted with the procedure in this paper.

Acknowledgments

We thank William Clark for supplying us with an example for comparing methods and Daniel Kimura for kindly sending us a prepublication copy of his manuscript. The anonymous reviewers provided helpful comments which are greatly appreciated. Partial support for this study was provided through the Cooperative Institute for Marine and Atmospheric Studies by National Oceanic and Atmospheric Administration Cooperative Agreement NA85-WCH-06134.

References

Barnes, M. A., and G. Power. 1984. A comparison of otolith and scale ages for western Labrador lake whitefish, *Coregonus clupeaformis*. Environmental Biology of Fishes 10:297–299.

Bartoo, N. W., and K. R. Parker. 1983. Stochastic age-frequency estimation using the von Bertalanffy growth equation. U.S. National Marine Fisheries Service Fishery Bulletin 81:91–96.

Bury, K. 1975. Statistical models in applied science. Wiley, New York.

Casselman, J. 1983. Age and growth assessment of fish from their calcified structures—techniques and tools. NOAA (National Oceanic and Atmospheric Administration) Technical Report NMFS (National Marine Fisheries Service) 8:1–17.

Chen, T., and S. Fienberg. 1976. The analysis of contingency tables with incompletely classified data. Biometrics 32:133–144.

Chen, T., Y. Hochberg, and A. Tennenbein. 1984. On triple sampling schemes for categorical data analysis with misclassification errors. Journal of Statistical Planning and Inference 9:177–184.

Clark, W. G. 1981. Restricted least-squares estimates of age composition from length composition. Canadian Journal of Fisheries and Aquatic Sciences 38:297–307.

Cook, R. 1983. Simulation and application of stock composition estimators. Canadian Journal of Fisheries and Aquatic Sciences 40:2113–2118.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood via the EM algorithm. Journal of the Royal Statistical Society Series B: Methodological 39:1–22.

Dinse, G. 1982. Nonparametric estimation for partially-complete time and type of failure data. Biometrics 38:417–431.

Espeland, M. A., and C. L. Odoroff. 1985. Log-linear models for doubly sampled categorical data fitted by the EM algorithm. Journal of the American Statistical Association 80:663–670.

Fienberg, S. 1970. The analysis of multidimensional contingency tables. Ecology 51:420–433.

Fienberg, S. 1980. The analysis of cross-classified categorical data, 2nd edition. MIT Press, Cambridge, Massachusetts.

Fridriksson, A. 1934. On the calculation of age distribution within a stock of cod by means of relatively few age-determinations as a key to measurements on a large scale. Rapports et Procès-Verbaux des Réunions, Conseil Permanent International pour l'Exploration de la Mer 86(6).

Greenland, S., and D. G. Kleinbaum. 1983. Correcting

for misclassification in two-way and matched-pair studies. International Journal of Epidemiology 12: 93–97.

Haberman, S. 1974. Log-linear models for frequency tables derived by indirect observation: maximum likelihood equations. Annals of Statistics 2:911–924.

Hoenig, J. M., and D. M. Heisey. 1984. Some uses of bias reduction matrices in fisheries surveys. International Council for the Exploration of the Sea, C.M. 1984/D:15, Copenhagen.

Hoenig, J. M., and D. M. Heisey. 1986. Reducing bias in estimates of stock composition. International Council for the Exploration of the Sea, C.M. 1986/D:14, Copenhagen.

Kimura, D. 1977. Statistical assessment of the age–length key. Journal of the Fisheries Research Board of Canada 34:317–324.

Kimura, D., and S. Chikuni. 1987. Mixtures of empirical distributions: an iterative application of the age–length key. Biometrics 43:23–35.

Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B: Methodological 44:226–233.

O'Gorman, R., D. H. Barwick, and C. A. Bowen. 1987. Discrepancies between ages determined from scales and otoliths for alewives from the Great Lakes. Pages 203–210 *in* R. C. Summerfelt and G. E. Hall, editors. Age and growth of fish. Iowa State University Press, Ames.

Pella, J. J., and T. L. Robertson. 1978. Assessment of composition of stock mixtures. U.S. National Marine Fisheries Service Fishery Bulletin 76:415–423.

Tennenbein, A. 1970. A double sampling scheme for estimating from binomial data with misclassifications. Journal of the American Statistical Association 65:1350–1361.

Tennenbein, A. 1971. A double sampling scheme for estimating from binomial data with misclassifications: sample size determination. Biometrics 27:935–944.

Westrheim, S. J., and W. E. Ricker. 1978. Bias in using an age-length key to estimate age-frequency distributions. Journal of the Fisheries Research Board of Canada 35:184–189.

Received October 29, 1986
Accepted April 14, 1987

Appendix A: Derivation of the Likelihood Equation (3)

For each sample, the probability that an animal lies in cell i, j is given (from equation 2) as

$$P(i, j \mid k) = P(j \mid i)a_{ik}.$$

Thus, the joint likelihood for observing n_{ij1} animals in cell i, j of the first sample ($i = 1, \ldots, I$; $j = 1, \ldots, J$) when the sample size is fixed at N_1 is

$$N_1! \prod_{i=1}^{I} \prod_{j=1}^{J} [P(j \mid i)a_{i1}]^{n_{ij1}}/n_{ij1}!$$

for multinomial sampling. Analogous results hold for the second sample, with a_{i2}, n_{ij2}, and N_2 replacing a_{i1}, n_{ij1}, and N_1, respectively.

Since the first and second samples are drawn independently, the joint likelihood is equal to the product of the likelihoods for each sample. Thus,

$$\Lambda = N_1!N_2! \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{2} [P(j \mid i)a_{ik}]^{n_{ijk}}/n_{ijk}!.$$

Taking logarithms gives

$$\log \Lambda = \text{constant term} + \sum_{i=1}^{I} \sum_{j=1}^{J} \{n_{ij1}\log[P(j \mid i)a_{i1}] + n_{ij2}\log[P(j \mid i)a_{i2}]\},$$

for which the constant term is

$$\log\left[N_1!N_2!/\left(\prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{2} n_{ijk}!\right)\right].$$

Dropping the constant term because it does not affect the maximization yields the kernel of the log-likelihood L given by equation (3).

Appendix B: Calculation of the Estimated Variance–Covariance Matrix

The computational procedure described here is based on the method of Louis (1982), which allows one to estimate the covariance matrix from the "complete" data likelihood function. The interested reader is referred also to Dinse (1982).

The likelihood function for the "complete" data was derived in Appendix A under the assumption of product multinomial sampling. The kernel of the log-likelihood can be written (from equation 3) as

$$L = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij1}\log[P(j\,|\,i)] + \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij1}\log a_{i1}$$
$$+ \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij2}\log[P(j\,|\,i)] + \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij2}\log a_{i2}.$$

We note that the above parameters are subject to certain constraints:

$$\sum_{j} P(j\,|\,i) = 1; \quad \sum_{i} a_{i1} = 1; \quad \sum_{i} a_{i2} = 1.$$

That is, the sum of probabilities over the appropriate subscript must equal unity. Thus, we can write

$$P(J\,|\,i) = 1 - \sum_{j=1}^{J-1} P(j\,|\,i); \quad a_{I1} = 1 - \sum_{i=1}^{I-1} a_{i1}; \quad a_{I2} = 1 - \sum_{i=1}^{I-1} a_{i2}.$$

We need only consider $I - 1$ of the a_{i1} parameters, $I - 1$ of the a_{i2} parameters, and $I(J - 1)$ of the $P(j\,|\,i)$ parameters.

Define $\mathbf{S}$ to be a vector of parameters of length

$$(I - 1) + (I - 1) + I(J - 1) = I(J + 1) - 2,$$

with the parameters ordered as

$$\mathbf{S}^T = [a_{11}, a_{21}, \ldots, a_{I-1,1}, P(1\,|\,1), P(2\,|\,1), \ldots, P(J-1\,|\,I), a_{12}, a_{22}, \ldots, a_{I-1,2}].$$

Then define the gradient vector of partial derivatives $\mathbf{G}$ with elements

$$g_l = \partial \log L/\partial \mathbf{S}_l$$

for $l = 1, \ldots, I(J+1) - 2$. This is accomplished as follows.

(I) If $l \le I - 1$,

$$g_l = (n_{l+1}/a_{l1}) - (n_{I+1}/a_{I1}).$$

(II) If $I - 1 < l \le (I - 1) + I(J - 1)$, let

$$i = 1 + \text{trunc}\{[l - (I - 1) - 1]/(J - 1)\};$$

trunc means truncate to an integer;

$$j^* = [l - (I - 1)] \text{ modulo } (J - 1);$$

$$j = \begin{cases} j^* \text{ if } j^* \neq 0 \\ J - 1 \text{ otherwise} \end{cases};$$

then

$$g_l = [n_{ij+}/P(j\,|\,i)] - [n_{iJ+}/P(J\,|\,i)].$$

(III) If $(I - 1) + I(J - 1) < l$, let

$$i = 1 - [(I - 1) + I(J - 1)];$$

then

$$g_l = (n_{i+2}/a_{i2}) - (n_{I+2}/a_{i2}).$$

Now, for each of the n_{+++} fish, construct an $I \times J \times 2$ indicator matrix $\mathbf{X}$. If the fish was in the i, j cell of sample $k = 1$, let $x_{ij1} = 1$, and set all other cells in $\mathbf{X}$ equal to 0. If the fish was in sample $k = 2$ and belonged to surrogate category j, then let

$$x_{ij2} = P(j|i)a_{i2}/\sum_{l=1}^{I} P(j|l)a_{l2}$$

for all i, and set all other cells in $\mathbf{X}$ equal to 0.

The next step is to evaluate the $\mathbf{G}$ vector for each fish f resulting in n_{+++} vectors $\hat{\mathbf{G}}_f$. The $\hat{\mathbf{G}}_f$ vectors are constructed from the $\mathbf{X}$ matrices and the final parameter estimates from the EM algorithm as follows.

(1) Everywhere the letter "n" appears in a formula for g_l, replace it with the letter "x."

(2) Everywhere a parameter appears, replace it with the estimate of the parameter.

Thus, the formula for the first element of $\mathbf{G}$,

$$g_1 = (n_{i+1}/a_{i1}) - (n_{I+1}/a_{I1}),$$

would become

$$\hat{g}_1 = (x_{i+1}/\hat{a}_{i1}) - (x_{I+1}/\hat{a}_{I1});$$

and would be evaluated for the fth fish by substituting in the appropriate values obtained from the $\mathbf{X}$ matrix for fish f.

For each fish f, compute the information matrix $\mathbf{I}_f$ by

$$\mathbf{I}_f = \hat{\mathbf{G}}_f \hat{\mathbf{G}}_f^T.$$

The Fisher information matrix for the entire data set is the sum of the $\mathbf{I}_f$ matrices. Thus,

$$\mathbf{I} = \sum_{f=1}^{n_{+++}} \mathbf{I}_f.$$

Of course, it is computationally more efficient to calculate $\mathbf{I}_f$ only once for each observed cell and then calculate $\mathbf{I}$ as the sum of the $\mathbf{I}_f$ matrices weighted by the observed cell counts.

Finally, the estimated variance–covariance matrix ($\hat{\mathbf{V}}$) of the parameter estimates is the inverse of the information matrix, $\hat{\mathbf{V}} = \mathbf{I}^{-1}$.

Example

Suppose the following set of age–length data is to be used as the basis for an inverse key:

	Length class									
Age	1	2	3	4	5	6	7	8	9	10
1	42	12	8	2	0	0	0	0	0	0
2	1	13	20	56	24	12	2	0	0	0
3	0	0	8	16	29	38	27	15	7	3
4	0	0	0	0	2	5	11	27	13	4

and suppose the following length-frequency distribution is obtained in the second sample:

[143 71 122 214 161 172 118 133 59 21].

Then the estimated age composition is as follows:

age	1	2	3	4
number	213	368	444	188
proportion	0.18	0.30	0.37	0.15

There are 42 separate parameters estimated for this model. Rather than present the full 42 × 42 covariance matrix, we focus attention on the estimated age composition for the second sample. The estimated covariance matrix for the proportion at ages 1, 2, and 3 in the second sample is given below (E denotes scientific notation).

	Age		
Age	1	2	3
1	3.57 E−4	−4.5 E−4	1.94 E−4
2	−4.5 E−4	2.29 E−3	−2.7 E−3
3	1.94 E−4	−2.7 E−3	4.51 E−3

To find the estimated variance of the proportion at age 4, $\hat{V}(\hat{a}_{42})$, we note that

$$\hat{a}_{42} = 1 - \sum_{i=1}^{3} \hat{a}_{i2}.$$

Hence,

$$\begin{aligned}\hat{V}(\hat{a}_{42}) &= \hat{V}\left[1 - \sum_{i=1}^{3} (\hat{a}_{i2})\right] \\ &= \sum_{i=1}^{3} \hat{V}(\hat{a}_{i2}) + 2 \sum_{i<j<4} \sum \widehat{\mathrm{Cov}}(\hat{a}_{i2}, \hat{a}_{j2}) \\ &= 1.2\ \mathrm{E}{-3}.\end{aligned}$$

By similar reasoning, the estimated covariance between $\hat{a}_{42}$ and the other estimated proportions is

$$\widehat{\mathrm{Cov}}(\hat{a}_{42}, \hat{a}_{i2}) = -\sum_{j=1}^{3} \widehat{\mathrm{Cov}}(\hat{a}_{i2}, \hat{a}_{j2})$$

($\widehat{\mathrm{Cov}}[\hat{a}_{i2}, \hat{a}_{i2}]$ is, by definition, $\hat{V}[\hat{a}_{i2}]$). Thus, $\widehat{\mathrm{Cov}}(\hat{a}_{42}, \hat{a}_{12}) = -9.4\ \mathrm{E}{-5}$, $\widehat{\mathrm{Cov}}(\hat{a}_{42}, \hat{a}_{22}) = 8.6\ \mathrm{E}{-4}$, and $\widehat{\mathrm{Cov}}(\hat{a}_{42}, \hat{a}_{32}) = -2.0\ \mathrm{E}{-3}$.