

A NEW APPROACH TO AGE-LENGTH KEYS: USING LAST YEAR'S AND THIS YEAR'S DATA TO ESTIMATE AGE COMPOSITION

Hoening, J.M., D.M. Heisey**, R.C. Hanumara****

** Department of Fisheries and Oceans, P.O. Box 5667, St. John's, Newfoundland, A1C 5X1, Canada*

*** Madison Academic Computing Center, University of Wisconsin, Madison, Wisconsin, USA*

**** Department of Statistics, University of Rhode Island, Kingston, Rhode Island 02881, USA*

NOTE: Results still preliminary, not to be cited without permission of the senior author.

SUMMARY

Age composition is often estimated by measuring the lengths of a large number of fish and ageing a small portion of the measured fish. In statistical terms, this approach can be described as noting the value of a covariate (length) on each of a large number of randomly selected units (fish) and noting the value of the covariate and the exact class (age) to which the unit belongs for a smaller sample. The small, cross-classified sample can be used to estimate the classification rates and these, in turn, can be used in conjunction with the estimated distribution of the covariate to obtain an improved estimate of the population composition. There are two approaches to this problem characterized by the way in which the classification rates are defined. The simplest approach uses estimates of the probability $P(i|j)$ that the fish is actually age i given that the length is j . The more complicated approach uses estimates of the probability $P(j|i)$ that the length is j given that the fish is actually age i . The latter approach involves estimating more parameters than the former but avoids the necessity for the cross-classified sample to be a random sample from the same population as the population from which the large sample was drawn for estimating the length composition. In this paper, we show that the two approaches can be combined when there are multiple samples. For example, one might have two samples obtained by random sampling of the population, and a third cross-classified sample from another population, such as the one in the previous year, with different composition but identical classification probabilities $P(j|i)$. We also show how to modify the method to allow for fixing the numbers aged in each length category.

RESUME

La structure démographique est souvent estimée en mesurant la taille d'un grand nombre de poissons et en déterminant l'âge d'une fraction réduite du poisson mesuré. En termes statistiques, cette méthode peut être décrite comme notant la valeur d'une covariance (longueur) sur chacun d'un grand nombre d'unités (poissons) sélectionnées de façon aléatoire, et notant la valeur de la covariance et la classe (âge) exacte à laquelle appartient l'unité pour un échantillon plus réduit. L'échantillon réduit, contre-classifié, peut servir à estimer les taux de classement, qui peuvent à leur tour être utilisés conjointement avec la distribution estimée de la covariance pour obtenir une estimation améliorée de la composition de la population. Il y a deux façons d'aborder ce problème, caractérisées par la façon dont les taux de classement sont définis. La méthode la plus simple utilise des estimations de la probabilité $P(i/j)$ que le poisson est en réalité d'âge i étant donné que la longueur est j . La méthode plus complexe utilise des estimations de la probabilité $P(j/i)$ que la longueur est j étant donné que le poisson est en réalité d'âge i . Cette dernière méthode met en jeu l'estimation de plus de paramètres que la première, mais évite la nécessité de ce que l'échantillon contre-classifié soit un échantillon aléatoire de la même population que celle dont a été extrait l'échantillon ample pour estimer la composition de taille. Dans le présent document, nous avançons qu'il est possible de combiner les deux méthodes lorsqu'il s'agit d'échantillons multiples. Par exemple, on pourrait avoir deux échantillons obtenus par échantillonnage aléatoire de la population, et un troisième échantillon contre-classifié d'une autre population, telle que celle de l'année précédente, avec une composition différente mais des probabilités identiques de classification $[P(j/i)]$. Nous montrons également la façon de modifier la méthode pour permettre d'ajuster dans chaque catégorie de taille les nombres dont l'âge a été déterminé.

RESUMEN

Con frecuencia, la composición por edad se estima midiendo las tallas de un gran número de peces y determinando la edad de una pequeña porción de los peces medidos. En términos estadísticos, puede describirse este enfoque como uno que toma en cuenta el valor de una covariada (talla) de cada una de un gran número de unidades seleccionadas (peces) aleatoriamente, y que toma en cuenta el valor de la covariada y la clase exacta (edad) a la cual pertenece la unidad, para una muestra más pequeña. La muestra pequeña, obtenida por clasificación cruzada, puede utilizarse para estimar las proporciones de clasificación y éstas, a su vez, puede utilizarse conjuntamente con la distribución estimada de la covariada para obtener una mejor estimación de la composición de la población. Hay dos enfoques para este problema, que se caracterizan por la forma en la cual se definen las proporciones de la clasificación. El enfoque más sencillo utiliza estimaciones de la probabilidad $P(i/j)$ de que el pez tenga actualmente la edad i si la talla es j . El enfoque más complejo usa estimaciones de la probabilidad $P(j/i)$ de que la talla es j , si el pez tiene actualmente la edad i . Este último enfoque implica estimar más parámetros que el anterior, pero evita la necesidad de que la muestra obtenida por clasificación cruzada sea una muestra aleatoria de la misma población de la cual se extrajo la muestra grande para estimar la composición por tallas. En este documento, demostramos que los dos enfoques pueden combinarse cuando hay muestras múltiples. Por ejemplo, se podrían tener dos muestras obtenidas por muestreo aleatorio de la población, y una tercera muestra obtenida por clasificación cruzada de otra población, como la del año anterior, con una composición diferente pero idénticas probabilidades de clasificación (P_j/i). También se muestra cómo modificar el método para permitir fijar los números a los cuales se les ha determinado la edad en cada categoría de talla.

1. INTRODUCTION

Each year, fishery scientists examine otoliths and other skeletal hard parts from thousands of fish in order to determine the age of individual fish; this information is then used to estimate the age composition in the catches. Typically, lengths are determined for a large sample of fish and the lengths and corresponding ages are determined for a much smaller sample. The length frequency information can be obtained at low cost; it provides information on the age composition in the population since age and length are correlated.

Information is often available from previous years but this prior information is not combined with the current data on ages and lengths in order to estimate the age composition in the current year. This seems a pity since the data from previous years is essentially free and can be used to help interpret the length frequency data from the current year.

In this paper, we review the two basic approaches to age-length keys - forward or classic keys and inverse keys. The forward key requires information from the current year whereas the inverse key can make use of information from previous years. We show how to combine the two approaches in a single analysis. We then generalize the inverse key approach to allow for length stratification. The goal is to use information from previous years to aid in the estimation of age composition in the current year. An added benefit is that information from the current year provides for revised estimates with higher precision of the age composition in previous years.

We illustrate the logic of the two approaches by considering a simple case in which the age and length are determined for each of n fish; each examination assigns an age of 1 or 2 to the fish and a length of 1 or 2 to the fish. The result is a 2×2 cross-classified table as in Figure 1. There is also a sample of size N on which observations are made on only the length.

1.1 Approach 1 - the forward or classic key

Assume that the n cross-classified fish and the N fish examined for just the length are simple random samples from the same population. Then, the probability $P(i|j)$ that a fish is actually of age i ($i = 1$ or 2), given that it has length j ($j = 1$ or 2), is the same for both samples. One can estimate these conditional probabilities by

$$\hat{P}(i|j) = q_{ij} = n_{ij}/n_j$$

where the $\hat{}$ symbol denotes an estimate and the rest of the notation is as in Figure 1. Denote the 2×2 matrix with elements q_{ij} by Q and the vector of length proportions by

$$E = \begin{bmatrix} \frac{Y_1 + n_{\cdot 1}}{N + n} \\ \frac{Y_2 + n_{\cdot 2}}{N + n} \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

Then, intuitively, the age composition A might be estimated by

$$\hat{A} = Q E$$

where $\hat{A} = [\hat{a}_1, \hat{a}_2]^T$ is the vector of estimated proportions at age. Thus,

$$\hat{a}_i = \sum_{j=1}^2 \hat{P}(ij) c_j.$$

In this example, the conditional probabilities are estimated from the first sample and the marginal probabilities $P(j)$ are estimated from both samples as c_j . This estimator can be shown to be of maximum likelihood (see Tenenbein 1970; Hochberg 1977; Jolayema 1990). It is also an example of stratified random sampling where the units are post-stratified by length (Swensen 1988). Related estimators, which are not fully efficient, are discussed by White and Castleman (1981) and Hand (1986).

In fisheries research, this approach is usually modified slightly. The number of fish aged from each length category, n_{ij} , is fixed by the investigator. The estimator remains unchanged and is still of maximum likelihood. The variance changes, however.

1.2 Approach 2 - the inverse key

It may happen that the cross-classified sample and the length sample are obtained from different populations. For example, in the first year of a study both the lengths and ages might be recorded for a sample of size m but in the second year only the length is recorded. The conditional probabilities $P(i|j)$ from the first year will not be applicable to the results in the second year if the population composition has changed (Kimura 1977; Westrheim and Ricker 1978). To see this, consider the probability that a fish is actually age 1 given that the length is 1. If, in the first year, all fish are age 1 then all of the fish of length 1 will in fact be age 1 ($P(i=1|j=1) = 1$). If, in the second year, none of the fish are age 1 then none of the fish of length 1 will in fact be age 1 ($P(i=1|j=1) = 0$).

There may be a way out of this dilemma. The probability that a fish is length j given that it is age i , $P(j|i)$, might not vary with the population composition. Thus, the age composition of the fish population will change each year as a variable number of young fish are recruited into the population and thus the probability that a fish is a certain age given its size, $P(\text{age}=\text{ilength}=j)$, will vary from year to year. On the other hand, the distribution of size about age, $P(\text{length}=j|\text{age}=i)$, should not change much as the population changes in composition except inasmuch as the growth may be somewhat dependent on environmental conditions. It thus may be entirely reasonable to suppose that $\text{Prob}(\text{length}=j|\text{age}=i)$ is constant from one year to the next or one area to the next.

Define the matrix P to have elements p_{ij} given by

$$p_{ij} = m_{ij}/m_i.$$

where the m_{ij} are the cell counts from the sample of size m and are defined as for the n_{ij} . Also, redefine the vector E to have elements

$$E = \begin{bmatrix} Y_1/N \\ Y_2/N \end{bmatrix},$$

that is, the vector E contains estimates of the marginal probabilities $P(j)$ obtained from just the length survey. Then, intuitively, the age composition should be related to the length composition by

$$E = P^T \hat{A}.$$

Thus, the age composition can be estimated by

$$\hat{A} = P^{T^{-1}} E \quad (1a)$$

or, more generally, by

$$\hat{A} = (PP^T)^{-1} PE \quad (1b)$$

Equation (1b) is the least squares estimator.

This estimator has appeared in the applied literature a number of times, e.g., as a hypothetical example of correcting deer age composition (Searle 1966 p. 93-4); as a method of correcting stock composition estimates for mixed fisheries - see Worlund and Fredin (1962), Fukuhara et al. (1962), Berggren and Lieberman (1978), Pella and Robertson (1978), and van Winkle et al. (1988); as a means of estimating prevalence of diseases - see Rogan and Gladen (1978), Greenland and Kleinbaum (1983), Hand (1986); as a correction for misclassification in a fourfold table relating disease status to risk factors (Kleinbaum et al. 1982 and references therein); as a method of correcting estimates of deer harvest composition obtained from hunter reports (D. Ingebrigtsen, MN Department of Natural Resources, pers. comm.); and for converting length-frequency distributions to age-frequency distributions (Clark 1981; Bartoo and Parker 1983; Kimura and Chikuni 1987; Hoenig and Heisey 1987). It can be seen that when the estimates from (1) are feasible, they are maximum likelihood estimates. Kimura and Chikuni (1987) and Hoenig and Heisey (1986) developed procedures for finding estimates when equation (1) produces infeasible results.

Thus, there are two approaches to using estimates of classification probabilities to convert a vector of length frequencies to an estimate of age composition. Method 1 is straightforward, is well known, and requires that the classification probabilities be estimated from a random sample of the population to which they will be applied. Method 2 is more complicated and involves a backward or inverse type of reasoning. For method 2, the classification rates are conditional on the age rather than on the length. The method has been repeatedly derived in the applied literature but does not appear to be well established in the statistical literature. In the next section, we show how the two methods can be combined.

2. COMBINED METHOD

2.1 Three samples

Assume that we have three samples of fixed size. Sample 1 is a random sample of size m collected during a prior time period or from a nearby location. All m fish were classified according to age and length. Sample 2 is a random sample of size n from the population of interest; all fish in this sample are classified by both variables. Sample 3 is a random sample of size N from the population of interest and all fish are classified according to just the length. We denote the count of fish with age classification i and length classification j in samples 1 and 2 by m_{ij} and n_{ij} , respectively. The count of fish in sample 3 with length classification j is denoted by Y_j . We assume that $P(j|i)$ for sample 1 is the same as for samples 2 and 3, and we denote this by $P(j|i)_{123}$. In general, subscripts are used to denote the sample or samples to which a probability applies.

Likelihood for method 1.

Approach 1 utilizes the information in samples 2 and 3. The likelihood for samples 2 and 3 is the product of independent multinomials and can be written

$$\Lambda_1 = \prod_{i=1}^2 \prod_{j=1}^2 [P(i|j)_{23} P(i)_{23}]^{n_{ij}} \prod_{j=1}^2 P(i)_{23}^{Y_j}$$

There are three parameters to be estimated: two conditional probabilities and one marginal probability $P(i)$. The goal is to estimate the proportion $P(i)_{23}$ that is age i and, by the invariance principle of maximum likelihood estimation, this can be accomplished by

$$\hat{P}(i)_{23} = \sum_{j=1}^2 \hat{P}(i|j)_{23} \hat{P}(j)_{23}$$

Likelihood for method 2.

Approach 2 utilizes the information in samples 1 and 3. The likelihood is again the product of two multinomials

$$\Lambda_2 = \prod_{i=1}^2 \prod_{j=1}^2 [P(j|i)_{123} P(i)_1]^{m_{ij}} \prod_{j=1}^2 \left[\sum_{i=1}^2 P(j|i)_{123} P(i)_3 \right]^{Y_j}$$

There are four parameters to be estimated: one $P(i)_1$ and one $P(i)_3$, and two conditional probabilities.

combined likelihoods.

The likelihood for all of the data can be written as

$$\Lambda_3 = \prod_{i=1}^2 \prod_{j=1}^2 [P(j|i)_{123} P(i)_1]^{m_{ij}} \prod_{i=1}^2 \prod_{j=1}^2 [P(j|i)_{123} P(i)_{23}]^{n_{ij}} \prod_{j=1}^2 \left[\sum_{i=1}^2 P(j|i)_{123} P(i)_{23} \right]^{Y_j}$$

Here, we have rewritten the likelihood for samples 2 and 3 in terms of $P(j|i)_{123}$. However, it should be noted that this likelihood reduces to Λ_1 (the classic key) when $m = 0$.

2.2 Combined method - generalizations

It may occur that four samples are available: the three samples discussed in section 1.3 plus a length-frequency sample from the same population as the population from which sample 1 was drawn. We can use this additional sample to obtain improved estimates of $P(i)_1$. This is of interest not only for its own sake but also because the $P(i)_1$ occur in the formulae for the variance-covariance matrix when this is calculated on the basis of expected information.

A general form of the likelihood is presented below which allows for I age classes, J length classes, and K surveys. Here, a survey refers to fish examined from the same time and place. In each survey, both variables are noted on a random sample and,

optionally, just the length is noted on another random sample. Denote the number classified as length j in the k th survey for the sample in which just the length is noted by Y_{jk} (note that Y_{jk} can be 0). Then, the general form of the likelihood is proportional to

$$\Lambda_g = \prod_{i=1}^I \prod_{j=1}^J \prod_{k=1}^K [P(j|i) P(i)_k]^{n_{ijk}} \prod_{j=1}^J \prod_{k=1}^K \left[\sum_{i=1}^I P(j|i) P(i)_k \right]^{Y_{jk}}$$

Here, $P(j|i)$ is assumed to pertain to all samples, $P(i)_k$ pertains to all samples within the k th survey, and n_{ijk} is the number of fish cross-classified as ij in the k th survey.

3. ALLOWING FOR STRATIFICATION BY LENGTH

Until now, we have assumed that the fish which are aged are a random sample of the fish which were measured for length. In practice, one is likely to consider the lengths when selecting the fish to age, e.g., one might choose equal numbers of fish from each length class. Indeed, there is good reason to consider the lengths when selecting the sample for aging: otherwise one might obtain by chance a sample in which none of the fish were classified in a particular length class and one would not be able to estimate some of the classification rates. Haitovsky and Rapp (1992) modified Approach 1 to allow for fixed numbers from each length category to be aged.

Here, we show how the general model can be modified to allow for stratification by length. Assume that at a previous time a sample of M fish was randomly selected and measured resulting in X_1 fish being classified as length 1 and X_2 fish as length 2, $X_1 + X_2 = M$. Suppose further that ages are determined for m_1 and m_2 fish of length 1 and 2, respectively. This results in a cross-classified table with fixed column totals of m_1 and m_2 with table entries of m_{ij} where the i indexes the age classification and the j indexes the length classification.

The likelihood for the M fish classified by just the length is simply a binomial

$$\begin{aligned} \Lambda_M &= \binom{M}{X_1} P(j=1)^{X_1} P(j=2)^{X_2} \\ &= \binom{M}{X_1} \left[\sum_{i=1}^2 P(j=1|i) P(i) \right]^{X_1} \left[\sum_{i=1}^2 P(j=2|i) P(i) \right]^{X_2} \end{aligned}$$

The likelihood for the cross-classified table is the product of two binomials, one for each column

$$\Lambda_{m_1, m_2} = \binom{m_1}{m_{11}} P(i=1|j=1)^{m_{11}} P(i=2|j=1)^{m_{21}} \binom{m_2}{m_{12}} P(i=1|j=2)^{m_{12}} P(i=2|j=2)^{m_{22}}$$

Now, in the population at large,

$$P(i|j) = \frac{P(j|i) P(i)}{\sum_i P(j|i) P(i)}$$

by Bayes rule. Substituting this into the product of the likelihoods Λ_M and Λ_{m_1, m_2} yields a likelihood in three unknowns: two conditional probabilities (e.g., $P(j=1|i=1)$ and $P(j=1|i=2)$) and one marginal probability (e.g., $P(i=1)$). Thus, when the prior data have been stratified by length, the likelihood still contains information on the conditional probabilities that are used to model the current data. Note that it is necessary to know the results of the length survey, i.e., the X_j , unlike for the estimators in Section 1 for which knowledge of the X_j is optional. The data for the current survey are handled in the exact same way when the cross-classified table is generated by fixing the numbers in each length category. These results generalize easily to the case where there are I age categories and J length categories.

Note that the estimator of $P(i)$ for Tenenbein's model is the same as the estimator based on stratification derived above for the case where the cross-classified data come from the same population as the sample for which just the lengths are measured.

4. DISCUSSION

The methods considered here combine two previously unrelated approaches. The general approach allows one to use previous information whose cost is essentially free. Often, the previous information on age will have been collected according to a scheme in which the population is (post-)stratified by the length categories. This presents no problem if the estimate of the population length composition is known. Hoening and Heisey (1986) did not stress this point; it appears that some people have used the inverse key approach without accounting for the fact that the number of fish aged from each length category was fixed by the investigator and was not necessarily proportional to the number of fish in the length category.

The use of previous data is based on the assumption that the classification rates ($P(\text{length}|\text{age})$) have not changed from sample to sample. This assumption can be tested using standard methods such as a likelihood ratio test.

We have not encountered any computational difficulties with the general approach. This may be due to the fact that good starting values can usually be obtained by using one of the traditional approaches in a preliminary analysis to obtain starting values.

REFERENCES

- Bartoo, N. and K. Parker. 1983. Stochastic age-frequency estimation using the von Bertalanffy growth equation. *Fish. Bull.*, U.S. 81:91-96.
- Berggren, T.J., and J.T. Lieberman. 1978. Relative Contribution of Hudson, Chesapeake, and Roanoke Striped Bass, *Morone saxatilis*, Stocks to the Atlantic Coast Fishery. *U.S. National Marine Fisheries Service Fisheries Bulletin*, 76, 335-345.
- Clark, W. 1981. Restricted least-squares estimates of age composition from length composition. *Can. J. Fish. Aquat. Sci.* 38:297-307.
- Fukuhara, F.M., S. Murai, J.J. LaLanne, and A. Sribhibhadh. 1962. Continental Origin of Red Salmon as Determined from Morphological Characters. *International North Pacific Fisheries Commission Bulletin* 8, 400-408.
- Greenland, S. and D.G. Kleinbaum. 1983. Correcting for Misclassification in Two-way and Matched-pair Studies. *International Journal of Epidemiology*, 12, 93-97.
- Haitovsky, Y. and J. Rapp. 1992. Conditional Resampling for Misclassified Multinomial Data with Applications to Sampling Inspection. *Technometrics* 34, 473-483.
- Hand, D.J. 1986. Estimating Class Sizes by Adjusting Fallible Classifier Results. *Computations & Mathematics with Applications*, 12A, 289-299.

- Hochberg, Y. 1977. On the Use of Double Sampling Schemes in Analyzing Categorical Data with Misclassification Errors. *Journal of the American Statistical Association*, 72, 914-921.
- Hoening, J.M. and D.M. Heisey. 1987. Use of a Loglinear Model with the EM Algorithm for Correcting Estimates of Stock Composition and Converting Length to Age. *Transactions of the American Fisheries Society*, 116, 232-243.
- Kimura, D. 1977. Statistical assessment of the age-length key. *J. Fish. Res. Bd. Can.* 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.
- Kleinbaum, D.G., L.L. Kupper and H. Morgenstern. 1982. *Epidemiologic Research: Principles and Quantitative Methods*, Belmont, CA: Wadsworth
- Jolayema, E.T. 1990. Relative Frequency Estimation in Multiple Outcome Measurement with Misclassification. *Biometrical Journal*, 6, 707-711.
- Nedelman, J. 1988. The Prevalence of Malaria in Garki, Nigeria: Double Sampling with a Fallible Expert. *Biometrics*, 44, 635-655.
- Pella, J.J. and T.L. Robertson. 1978. Assessment of Composition of Stock Mixtures. *U.S. National Marine Fisheries Service Fisheries Bulletin*, 76, 415-423.
- Rogan, W.J. and B. Gladen. 1978. Estimating Prevalence from the Results of a Screening Test. *American Journal of Epidemiology*, 107, 71-76.
- Searle, S. 1966. *Matrix Algebra for the Biological Sciences (Including Applications in Statistics)*, New York: Wiley.
- Swensen, A.R. 1988. Estimating Change in a Proportion by Combining Measurements on a True and a Fallible Classifier. *Scandinavian Journal of Statistics*, 15, 139-145.
- Tenenbein, A. 1970. A Double Sampling Scheme for Estimating from Binomial Data with Misclassification. *Journal of the American Statistical Association*, 65, 1350-1361.
- Van Winkle, W., K.D. Kumar and D.S. Vaughan (1988), "Relative Contributions of Hudson River and Chesapeake Bay Striped Bass Stocks to the Atlantic Coastal Population," *American Fisheries Society Monograph*, 4, 255-266.
- Westrheim, S.J. and W.E. Ricker. 1978. Bias in using an age-length key to estimate age-frequency distributions. *J. Fish. Res. Bd. Can.* 35:184-189.
- White, B.S. and K.R. Castleman (1981), "Estimating Cell Populations," *Pattern Recognition*, 13, 365-370.
- Worlund, D.D. and R.A. Fredin (1962), "Differentiation of Stocks," pp. 143-153 in *Symposium on Pink Salmon*, H.R. MacMillan Lectures in Fisheries, University of British Columbia, Vancouver.

Figure 1. Notation for the results of determining the length and age of n fish (left) and for determining the length of N additional fish (right).

		length		length			
		1	2	1	2		
age	1	n_{11}	n_{12}	$n_{1.}$	Y_1	Y_2	N
	2	n_{21}	n_{22}	$n_{2.}$			
		$n_{.1}$	$n_{.2}$	n			