References

Appeldoorn, R. 1987. Modification of a seasonally oscillating growth function for use with mark-recapture data. J. Cons. Int. Explor. Mer 43:194-198.

Cloern, J.E. and F.H. Nichols. 1978. A von Bertalanffy growth model with a seasonally varying coefficient. J. Fish. Res. Bd. Can. 35:1479-1482.

Shepherd, S.A. and W.S. Hearn. 1983. Studies on southern Australian abalone (Genus Haliotus). IV Growth of H. laevigata and H. ruber. Aust. J. Mar. Freshw. Res. 34:461-475.

Pauly, D. and N. David. 1981. ELEFAN I, a basic program for the objective extraction of growth parameters from length-frequency data. Meeeresforsch. 28(4):205-211.

Pauly, D. and G. Gaschütz. 1979. A simple method for fitting oscillating length growth data, with a program for pocket calculators. ICES CM 1979/G:24:26 p.

Pitcher, T.J. and P.D.M. Macdonald. 1973. Two models for seasonal growth in fishes. J. App. Ecol. 10:599-606.

Splitting Length Distributions into Peaks and the Clean Class Concept

HANS LASSEN

Danish Institute of Fisheries and Marine Research

Charlottenlund Slot, DK2920 Charlottenlund, Denmark

The problem which we shall address in this contribution is that of splitting a composite length sample into its component distributions, each of which should be identified as an age-group.

The splitting is often done using methods which require that only one age-group contribute to the length distribution in a certain length range. A length range where all fish have the same age is called a "clean class".

It is here assumed that the growth of fish follows some growth function, e.g., a von Bertalanffy curve and that, for a given age, the lengths around the growth curve are normally distributed.

The basic model is

$$n(L) = N_1 \frac{dL}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_1}{\sigma_1}\right)^2\right] + N_2 \frac{dL}{\sqrt{2\pi}\,\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_2}{\sigma_2}\right)^2\right] + \ldots\ldots$$

for a total of K-terms, where

n(L)	:	the number of fish in length class 1 (midlength)
d L	:	range of length group
$N_1, N_2,..$	:	total number of fish belonging to group 1, 2, .. (e.g. age-group)
$L_1, L_2,..$	:	mean length of group 1, 2, ...
$\sigma 1, \sigma 2,..$	:	width of group 1, 2, ...

The estimation problem, i.e. finding $N_1, N_2, ..., L^{-1}, L^{-2},...,$ and $\sigma_1, \sigma_2, ...,$ in total 3 x k parameters, can be addressed in several ways.

Hasselblad (1966) and several subsequent authors discussed a least square solution, i.e.

$$\sum \left\{ n(L) - N_1 \frac{dL}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_1}{\sigma_1}\right)^2\right] \right\}^2 = \min$$

The minimum is found with respect to the parameters N_1, N_2, a.s.f. 3 x k parameters. This approach does not require any clean classes but is insensitive to small peaks. This may partly be overcome by dividing the squared difference by the theoretical (i.e., expected) number of fish for each term in the sum. Then this estimator is a chi-square estimator. Note that the number of length groups, K, is not part of the estimation problem; K must be known from elsewhere.

The Maximum Likelihood Approach (MLA), which has been in vogue recently, ends up in solving a similar set of equations as does the least square approach or the chi-square estimator. The actual form of the equations depends on the random noise in the sampling.

One model is

$$n(L) = N_1 \frac{dL}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_1}{\sigma_1}\right)^2\right] + N_2 \frac{dL}{\sqrt{2\pi}\,\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_2}{\sigma_2}\right)^2\right] + \ldots\ldots + \epsilon$$

where ϵ is an explicit form of the stochastic term.

The noise may originate from sampling of the population. Sampling is usually a two-stage process, first the fishery samples the fish population and secondly the fishery catch is sampled. If sampling errors are caused mainly by market sampling, then a multinomial distribution should be found and one can write the ML estimator equations and solve those for a particular data set. No clean class problem occurs, and k is not estimated.

Integrated Methods

Another class of methods tries to reduce the number of parameters by introducing:

- identification of length groups as age-groups
- linking age-groups by a growth equation, usually the von Bertalanffy growth function.

This class of models includes the ELEFAN I program (Pauly and David 1981) and the ML analysis proposed by Sparre (1987).

This changes the model to

$$n(L) = N_1 \frac{dL}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_1}{\sigma_1}\right)^2\right] + \ldots$$

$$L_t = L_\infty (1 - \exp(-K(t - t_o)))$$

The growth equation may include a seasonal growth term and/or may be a generalization of the von Bertalanffy equation (Pauly 1984).

Now the number of parameters is reduced to N_1, N_2, ..., σ_1, σ_2, and L_∞, K and t_o, i.e., 2 x k + 3 parameters.

ELEFAN I and the other methods using growth equations overcome the problem of finding the number of peaks. This is accomplished by introducing the growth equation and thereby accepting an infinite number of peaks near L_∞. The growth curve will suggest the number of peaks to be expected. Peaks not accounted for or peaks which are misplaced for some reason or another influence the goodness-of- fit. If many peaks are not accounted for or are missing then the model has a low degree of explanation, it fits badly. In ELEFAN I, the goodness-of-fit is measured by the ESP/ASP ratio.

If mortalities are to be estimated, it is a problem of estimating stock size, N_a. The obvious extension, $N_{t+1} = N_t e^{-Z}$, is seldom introduced in the splitting procedure, due to the variability in recruitment and in problems in raising samples to total population. This extension would reduce the number of parameters to σ_1, σ_2, ..., plus recruitment and growth. One further step along this line would be to postulate that $s_t^2 = a \times t$, which would further reduce the number of parameters. This latter step is sometimes suggested as a check on the calculations but I have not seen a fully integrated model with all these submodels. Someone may try it out some day.

The actual estimation of the parameters of any of this class of models can be approached using an ML approach, following specification of a sampling error function (Sparre 1987), using ELEFAN I or a Least Squares model. The clean class problem does not occur and k (i.e., the number of cohorts) is determined by the growth model.

The ELEFAN I program does not use the normal distribution model for any given length-group but instead requires more generally, that

$$n(L) = N_1 f_1(L; \bar{L}_1) + N_2 f_2(L; \bar{L}_2) + \ldots\ldots$$

where the f's are functions which have a maximum, or possibly several maxima, close by. The f's become zero far away from the maximum.

Graphical Methods

A third class of methods are the graphical approaches (Harding 1949; Cassie 1954; Tanaka 1966; Bhattacharya 1967). They analyze each sample independently, as opposed to the integrated methods discussed above. Therefore, one has to estimate the number of peaks (k) in each sample and this generally cannot be done based on the length frequencies alone. The number of peaks (k) in the length distribution is estimated by assuming that certain length intervals are only affected by one and only one peak. Such an interval is called *a clean class*.

All three methods analyze one model

$$n(L) = N_1 \frac{dL}{\sqrt{2\pi}\,\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_1}{\sigma_1}\right)^2\right] + N_2 \frac{dL}{\sqrt{2\pi}\,\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{L-\bar{L}_2}{\sigma_2}\right)^2\right]$$

The next step in these analysis to to find a set of "clean" classes i.e. length intervals where only one peak contributes significantly. Then one either:

- plots the data on probability paper (Cassie's method)
- takes logarithms and plots on graph paper (Tanaka's method)
- takes the differences of logarithms and plot on graph paper (Bhattacharya's method).

All these transformations provide graphs which can be *visually* (and thereby subjectively) evaluated for finding where the peaks are and *how many* occur. The crux of these methods is the concept of the "clean class" i.e., a length range where only one peak (i.e., age-group) contributes to the length distribution.

Since this problem is partially linked to the number of peaks, k, in the length distribution it is not surprising that no objective method can be devised which will automatically find the clean classes. However, Pauly and Caddy (1985) attempted to devise such a method, based on the identification of sets of three successive points with very high correlation coefficients on a Bhattacharya plot. This implies, among other things, that

- "clean" classes covers at least 3 length intervals (this puts an upward limit on the number of peaks which can be found.)
- too many peaks will tend to be found (unless an extremely high value of the critical correlation coefficient is used).

Also the investigation of whether other points in the Bhattacharya plot belong to the same clean class is a rather arbitrary process which requires assumptions about the structure of the stochastic variability of the sampling. The use of correlation coefficients is, moreover, not valid since the data points are not independent, since each point involves some information from the previous point.

This method hides a series of extra assumptions or subjective judgement behind "critical" correlation coefficients, choice of the number of points used in the computations, etc. It would be better, I believe, to take responsibility for your choice, make your choice and stand by it.

Concluding Remarks

Many discussions within the FAO/DANIDA group of lecturers have centered on how much one may rely on an integrated method and to what extent one should break down the analysis into smaller bits and let the user make the judgments about how to assemble the various details into a comprehensive assessment. To me it is obvious that integrated methods require a very elaborate set of output tables and graphs to enable the researcher to check the model, the goodness-of-fit figures usually provided are not adequate. Further, the newcomer to the field may overlook the basis of the methods when presented with an integrated model and may end up as a button-pusher. I guess that, in the long run, the integrated methods will be as transparent to the student as say a Ford-Walford plot is to those of to-day. But, personally I am not convinced we are there now and I feel that all the integrated analysis programs which I have seen are lacking in ability to indicate how well the model fits the data. ELEFAN I, for example, provides one with the ESP/ASP ratio and a graph showing the distributions and the fitted curve. However, it does not show which peaks are close to the growth curve and therefore only deviate from the model due to noise and which peaks deviate so much as to conflict with the data.