Canadian Journal of Fisheries and Aquatic Sciences

Journal canadien des sciences halieutiques et aquatiques

Volume 37, No. 9, September 1980

Volume 37, nº 9, septembre 1980

A New Approach to Length–Frequency Analysis: Growth Structure

JON SCHNUTE AND DAVID FOURNIER

Department of Fisheries and Oceans, Resource Services Branch, Pacific Biological Station, Nanaimo, B.C. V9R 5K6

SCHNUTE, J., AND D. FOURNIER. 1980. A new approach to length–frequency analysis: growth structure. Can. J. Fish. Aquat. Sci. 37: 1337–1351.

This paper presents a new approach to length–frequency analysis which takes account of biological structure in the mean lengths and standard deviations in length for various age-classes of fish. The new methods help determine biologically meaningful solutions, even when earlier methods lead to an ambiguous set of competing solutions. The structure of the standard deviations turns out to be especially important. For describing the means, new parameters are defined for von Bertalanffy growth which prove to have greater biological meaning and numerical stability than L_∞, K, and t_0. These new parameters can often be estimated easily from the raw data in cases where the species experiences a slowing of growth with age. This paper also presents χ^2 methods which can be used to rank competing solutions, although the results are not definitive. All methods are illustrated using data previously published for pike and abalone. An appendix describes in detail the computer programs required for the analysis.

Key words: length–frequency analysis, aging of samples, von Bertalanffy growth, growth, maximum likelihood estimation, nonlinear estimation

SCHNUTE, J., AND D. FOURNIER. 1980. A new approach to length–frequency analysis: growth structure. Can. J. Fish. Aquat. Sci. 37: 1337–1351.

L'approche nouvelle de l'analyse de la fréquence des longueurs présentée dans cet article tient compte de la structure biologique dans les longueurs moyennes et dans les écarts types de longueur chez diverses classes d'âge de poissons. Ces nouvelles méthodes aident à trouver des solutions biologiquement significatives, même quand d'autres méthodes ont donné un ensemble ambigu de solutions opposées. Particulièrement importante est la structure des écarts types. De nouveaux paramètres de croissance de von Bertalanffy sont utilisés pour décrire les moyennes et s'avèrent biologiquement plus significatifs et numériquement plus stabes que L_∞, K et t_0. Souvent, ces nouveaux paramètres sont facilement estimés à partir des données brutes dans les cas où l'espèce subit un ralentissement de croissance avec l'âge. On trouvera également dans cet article un exposé de méthodes de χ^2 pouvant servir à classer des solutions opposées, sans toutefois qu'on ait de résultats définitifs. Toutes ces méthodes sont illustrées par des données déjà publiées sur le brochet et l'ormeau. Les programmes d'ordinateur nécessaires à l'analyses sont décrits en détail en appendice.

Received September 10, 1979
Accepted May 22, 1980

Reçu le 10 septembre 1979
Accepté le 22 mai 1980

Printed in Canada (J5772)
Imprimé au Canada (J5772)

In studying a fish stock, the biologist frequently tries to determine three basic characteristics, namely, (1) the mean length at each age, (2) the distribution of lengths, and (3) the distribution of ages. Clearly these characteristics are related. For example, if the mean length at age and the length distribution are both known, then something can be said about the age distribution. This principle underlies the use of age–length keys.

Usually the distribution of fish lengths is the easiest characteristic of the three to ascertain. One need only take a suitably large random sample of the stock and measure the length of each fish in it. By contrast, the other two characteristics may be very difficult to determine, because they involve knowing the ages of fish in a sample. For this reason attempts have been made to use the length distribution alone to determine both the age distribution and the mean length at age.

Such methods date back to Petersen (1892), who used length–frequency modes in a population of *Zoarces viviparus* to identify distinct age-groups. An interesting reproduction of Petersen's key figure appears in Ricker (1975, p. 204); one can see immediately how distinct length clusters visually suggest distinct ages. About the same time, Pearson (1894) developed the first statistical treatment of the problem of overlapping component distributions, although his method applied to the case of two components only and did not easily generalize. For practical reasons, modern computers are necessary to implement a general statistical procedure for distinguishing an arbitrary number of overlapping component distributions. Hasselblad (1966) first described such a method in detail and developed the necessary computer program. However, workers in fisheries research have primarily favored the graphical methods of Harding (1949) and Cassie (1954). In a recent departure from this tradition, Macdonald and Pitcher (1979) show how some variations of Hasselblad's statistical technique can be applied to fish stocks. Also, McNew and Sommerfelt (1978) discuss difficulties which may occur when the distribution of lengths at each age is not normal as Hasselblad assumes.

The principal motive for length–frequency analysis has usually been to determine the distribution of ages of fish. For management purposes, it is important to know the age composition to predict the status of the stock in future years. However, as stated earlier, the analysis also gives information on the length at age. The interested worker may then go on to construct growth curves, say of the von Bertalanffy type. This has always been viewed as a separate problem. First, one might use the method of Cassie to find mean lengths at age; then, second, one might determine an appropriate growth curve. These have been considered separate issues, with a distinct procedure for each case.

This paper presents a new point of view in which these two previously separate issues become one. A single procedure applied to length–frequency data gives both the age composition of the stock and the parameters for growth. More explicitly, the procedure uses length–frequency data to determine the percentage of fish at each age, as well as parameters which define the mean length and standard deviation in length at each age. For example, three von Bertalanffy parameters (such as L_∞, K, t_0) might be determined which define a growth curve through the mean lengths.

Aside from giving growth data along with the usual age composition analysis, the new procedure has at least one further advantage over those used previously. It helps eliminate ambiguity. Most workers who try Cassie's method of plotting the cumulative frequency on probability paper soon discover that the location of inflection points (which serve as breaking points between age-groups) can be quite ambiguous. Also, each time one makes an ambiguous choice, that choice perpetuates itself into the determination of further inflection points. This difficulty is somewhat circumvented by Hasselblad's statistical procedure, but not entirely. Even Hasselblad's method may give rise to many competing solutions, particularly when the number of age-groups is large and older fish with different ages have similar lengths. Macdonald and Pitcher deal with such problems by invoking ad hoc constraints on some of the parameters, especially the standard deviations. In this paper, we explore different constraints, motivated biologically. Instead of allowing arbitrary mean lengths and standard deviations in length as Hasselblad does, our procedure requires that the means and/or standard deviations should conform to a growth pattern. For many species this is reasonable because, for example, the growth rate slows as the fish mature and, consequently, the mean lengths are related to each other in some way.

A general conclusion from this argument is that more biological structure leads to less ambiguity. Unconstrained methods may give many possible solutions to the same problem. By requiring that population characteristics (such as mean lengths at age) must bear some relationship to one another on biological grounds, one can perhaps eliminate a significant number of possible solutions and focus immediately on a restricted number which are of biological interest. Although emphasis is placed on von Bertalanffy growth, other relationships among population characteristics (for example, between standard deviations and means) will also be introduced. In all cases, the underlying principle remains the same, namely, introducing biological structure into the statistical model to locate solutions of biological interest. In each application the research biologist may wish to tailor the model according to his own view of the biological characteristics of the stock.

One final point should be clarified before the analysis begins. The terms "age distribution" and "length distribution" apply throughout to a stock which has been randomly sampled. That population may be a stock at sea, but more often it is the total population of caught fish. If random samples are taken from the commercial catch and the gear is selective for certain ages or sizes

of fish (as it almost always is), then, of course, the population at sea has not been randomly sampled. At best the results can speak for the hypothetical population of catchable fish, or, more simply, the entire catch itself. If inferences are to be made about a stock at sea, then the biologist may need to take samples with special noncommercial gear, or he may have to compensate his results from the commercial catch in some way to allow for gear selectivity.

von Bertalanffy Growth

Before introducing the structure of biological growth into the statistics of length–frequency, it is useful to take a fresh look at what is implied by von Bertalanffy growth. Suppose that the fish population being considered has M age-classes from age a_1 to a_M, where the ith age is

$$a_i = a_1 + i - 1; \qquad i = 1, \ldots, M. \tag{1}$$

Let μ_i be the mean length of fish at age a_i. Then, if these means lie on a curve of the type proposed by von Bertalanffy (1938),

$$\mu_i = L_\infty(1 - e^{-K(a_i - t_0)}); \qquad i = 1, \ldots, M. \tag{2}$$

In this equation, L_∞ is the theoretical length which fish approach as they grow older, K relates to the fraction by which the gap between current length and asymptotic length is reduced each year, and t_0 is the age at which fish length extrapolates back to zero along the curve.

As Ricker (1975) points out in connection with the work of Ford (1933) and Walford (1946), equation (2) implies that

$$\mu_{i+2} - \mu_{i+1} = e^{-K}(\mu_{i+1} - \mu_i). \tag{3}$$

Consequently, the distance between two successive mean lengths shrinks each year to the fraction

$$k = e^{-K} \tag{4}$$

of its former size. This reduction corresponds to a slowing in growth; as the fish get older, the change in mean length from one year to the next becomes smaller.

The idea behind equation (3) can be used to formulate a new version of the von Bertalanffy growth equation. Suppose that the first and final mean lengths μ_1 and μ_M are known. Call them l and L, respectively. Then from (1)–(2)

$$l = L_\infty(1 - e^{-K(a_1 - t_0)}), \tag{5}$$

$$L = L_\infty(1 - e^{-K(a_1 + M - 1 - t_0)}). \tag{6}$$

According to (3)–(4), the remaining means, from μ_2 to μ_{M-1}, must lie between l and L so that each year the incremental growth $\mu_{i+1} - \mu_i$ shrinks to the fraction k of its previous value. There is only one way to arrange the remaining means in this fashion. For example, if $k = 1$, the means must be equally spaced between l and L. If $k \neq 1$, it turns out that

$$\mu_i = l + (L - l)\frac{1 - k^{i-1}}{1 - k^{M-1}}; \qquad i = 1, \ldots, M. \tag{7}$$

This result is proved in Appendix A.

Equations (2) and (7) are merely two expressions for the same curve with different choices of parameters; (2) uses L_∞, K, t_0, while (7) uses l, L, k. Also, (4)–(6) gives the transformation from the first set of parameters to the second. The reverse transformation is given by

$$L_\infty = \frac{L - lk^{M-1}}{1 - k^{M-1}}, \tag{8}$$

$$K = -\log k, \tag{9}$$

$$t_0 = a_1 - \frac{1}{\log k}\log\left(\frac{L - l}{L - lk^{M-1}}\right). \tag{10}$$

The parameter set (l, L, k) is much more appropriate to length–frequency analysis than (L_∞, K, t_0). The biologist actually knows both the shortest and longest observed lengths. In practice l, the mean length of the youngest fish, lies somewhat above the shortest observed length; similarly L lies somewhat below the longest observation. Typically, the mean lengths cluster near L for older fish, while mean lengths for younger fish are more spread out, leaving noticeable peaks. Suppose that the biologist can formulate an opinion on the first and last means, l and L, as well as at least one further mean μ_i for some age a_i ($i \neq 1$ or M). Then these values of l, L, and μ_i can be substituted into (7), and the resulting equation can (in principle) be solved for k. Unfortunately, it is not possible to express analytically a general solution for k in equation (7); however, a few trial values of k will suggest a first estimate. In practice, the biologist may have an opinion about several mean lengths other than l and L. Again, the principle of trying a few k values usually leads to a reasonable first estimate.

In summary, it is not too difficult in many cases to obtain first estimates of l, L, and k which are biologically meaningful. The parameters l and L are based on observations near the shortest and longest lengths in the sample. Information on k then comes from an estimate of one or more means other than the first and last. Conceptually, k represents the fixed fraction by which the annual growth increment is multiplied each year. A choice of k near 1 implies almost uniform growth so that the means are spaced almost evenly between l and L. On the other hand, a choice of k near 0 implies that the annual growth decreases each year to a small fraction of its former size. Consequently, the first increment in mean length is relatively large, while the mean lengths of older fish tend to cluster near L.

By contrast, the significance of L_∞, K, and t_0, is not always obvious and sometimes even deceptive. For example, it is tempting to suppose that L_∞ corresponds roughly to the longest observed length, or a little beyond. However, if k is near 1 and the means are almost

evenly spaced, then L_∞ may lie far beyond the observed range. If k is near 0, then L_∞ is approximately equal to L by (8). Since L is the highest *mean* length, it lies below the longest observed length, and, consequently, so might L_∞ in this case. There have also been difficulties with the interpretation of K and t_0. As Ricker (1975, p. 221) points out, "it is misleading to refer to K as a growth rate" because K actually measures the exponential rate of approach to asymptotic size. The parameter K does not involve the units of length; a species may grow rapidly in cm/yr and still be characterized by a small value of K. Also, since t_0 is theoretically the age at which fish have length 0, it is difficult to understand biologically why t_0 is not always exactly 0. This problem, of course, is merely an artifact which results from extending the curve beyond the range of the data; however, it is still difficult to grasp intuitively what a reasonable value of t_0 might be for a particular data set.

Some of the problems just cited have led to controversy in the literature. For example, Knight (1968) shows that the interpretation of L_∞ as an asymptotic length may lead to complete nonsense. The point is that the nonsense stems not from the mathematics, but from the biological interpretation commonly placed on the results. The parameters l, L, and k merely summarize observed facts about the data. By contrast, L_∞, K, and t_0 have been used as indicators of fundamental biological characteristics of the fish. Such interpretations are almost always purely speculative; they may, as Knight points out, be wrong.

In this description of von Bertalanffy growth, emphasis has so far been placed on the mean length for each age. Of course, fish of the same age do not always have the same length; if they did, length–frequency analysis would be trivial. Associated with each mean length μ_i is a standard deviation in length σ_i. Just as the means might conform to a growth relationship, so also the standard deviations might be prescribed by some rule. For example, the standard deviations might be a linear function of the means; that is,

$$(11)\quad \sigma_i = s + (S - s)\frac{\mu_i - l}{L - l}; \qquad i = 1, \ldots, M;$$

where s and S are the standard deviations σ_1 and σ_M for ages a_1 and a_M, respectively. Alternatively, the σ's might be a linear function of the ages; that is,

$$(12)\quad \sigma_i = s + (S - s)\frac{i - 1}{M - 1}; \qquad i = 1, \ldots, M.$$

A special case of both (11) and (12), obtained when $s = S$, is the situation of constant standard deviation

$$(13)\quad \sigma_i = s; \qquad i = 1, \ldots, M.$$

Notice that (7) and (11) taken together imply that

$$(14)\quad \sigma_i = s + (S - s)\frac{1 - k^{i-1}}{1 - k^{M-1}}; \qquad i = 1, \ldots, M.$$

Consequently, in this case, the parameters s, S, and k play a role for the σ's similar to that of l, L, and k for the μ's. In analogy with the means, the standard deviations lie between s and S in such a way that the gap between two successive σ's shrinks each year to the fraction k of its former size. If $k = 1$, the σ's are equally spaced between s and S as described by (12).

In summary, there is a logical organization to the possibilities (11) to (14) for the σ's. Equation (13) is a special case of (12), which, in turn, is a special case of (14). Also, (14) is equivalent to (11) when the means conform to von Bertalanffy growth.

The parameters s and S are not completely analogous to l and L in one significant respect. While it is always true that $l < L$, it may happen as in (13) that $s = S$, or even $s > S$. Many factors may contribute to size variation among fish of one age. If these factors accumulate so that shorter fish fall farther behind and longer fish tend to do better, then it is reasonable that $s < S$. On the other hand, it may happen that younger fish experience considerable variability in growth rate, while older fish tend to reach a limiting size. In this case, as the animals grow older, the small ones tend to catch up, rather like children approaching adulthood. Consequently, $s > S$ because an initially large size range narrows with age. The special case $s = S$ might occur when the tendencies to each extreme just balance.

Length–Frequency Statistics

To describe the statistics of length–frequency sampling, it is necessary to extend the notation of the previous section. In a population with M age classes, let μ_i, σ_i, and π_i be, respectively, the mean length, standard deviation in length, and percentage (or fraction) of fish at age a_i ($i = 1, \ldots, M$). Suppose that fish are sampled randomly from this population and that the length of each fish is determined to lie in one of the N intervals

$$(x_j - w/2, x_j + w/2); \qquad j = 1, \ldots, N;$$

where

$$x_j = x_1 + (j - 1)w$$

is the midpoint of the jth interval and every interval has width w.

Assume that fish lengths are distributed normally in each age-group. Then, given that a fish has age a_i, the probability that its length lies in the jth interval is

$$(15)\quad q_{ij} = \frac{1}{\sigma_i\sqrt{2\pi}}\int_{x_j - w/2}^{x_j + w/2} \exp\left[-\tfrac{1}{2}\left(\frac{x - \mu_i}{\sigma_i}\right)^2\right]dx.$$

Here the integral (15) defines a section of area under a normal curve. Details for its numerical calculation are given in Appendix C. Suppose that a total frequency f of fish are sampled. Then the expected frequency of fish having length in the jth interval is

$$(16)\quad f_j = f\sum_{i=1}^{M} \pi_i q_{ij}.$$

TABLE 1. Notation summary in logical order.

Classification	Notation	Description
Age descriptors	i	Index for age
	a_i	ith age
	M	Number of age-classes
Length descriptors	j	Index for length
	x_j	Midpoint of jth interval
	w	Width of each interval
	N	Number of length intervals
Population parameters	μ_i	Mean length at age a_i
	σ_i	Standard deviation in length at age a_i
	π_i	Fraction (%) of fish at age a_i
	l, L, k	von Bertalanffy parameters for μ's
	s, S	Parameters for σ's
	P	Total number of population parameters
Sample size	f	Number of fish sampled
Observations	$\hat{f}_j$	Observed frequency in jth interval
	r	Small number of observations (integer)
	N_r	Number of intervals with $\hat{f}_j \geq r$
	$\hat{g}_r$	Observed total from intervals with $\hat{f}_j \geq r$
Expectations	q_{ij}	Probability of jth length, given age a_i
	f_j	Expected frequency in jth interval
	g_r	Expected total from intervals with $\hat{f}_j \geq r$
Measure of closeness	A	Separation statistic
Measure of fit	B_r	χ^2 statistic
	D_r	df

In this sum the product π_i times q_{ij} represents the overall probability that a fish has both age a_i (probability π_i) and a length in the jth interval (probability q_{ij}). Since a fish with specified length must come from one of the age-groups, these probabilities are then summed over the ages to give the expected probability for each length interval. In other words, (16) merely states that the expected number of fish in each length interval is the sum of expected numbers broken down by age.

When a sample is taken, of course, the observed frequency of fish with lengths in the jth interval (call it $\hat{f}_j$) does not, in general, equal the expected frequency (f_j). This is true because the observation $\hat{f}_j$ includes statistical sampling error. In practice, since the population parameters (such as μ's, σ's, and π's) are not known, the expected frequencies are also unknown. The problem is to determine a set of population parameters which would lead to expected frequencies as close as possible to those observed. This requires a criterion for "closeness." In this paper, closeness is measured by the separation statistic

$$(17) \quad A = 2 \sum_{j=1}^{N} \hat{f}_j \log (\hat{f}_j / f_j).$$

The expected frequencies are said to be as close as possible to the observations when A is minimal. The significance of this particular criterion will be discussed shortly. It has been suggested previously for this application by Macdonald and Pitcher (1979), who refer to A as "twice the Kullback minimum discrimination information statistic."[1] For reference, the notation used here is summarized in Table 1.

To understand exactly how this criterion might be applied, notice from (15) that q_{ij} depends on μ_i and σ_i; in symbols,

$$q_{ij} = q_{ij}(\mu_i, \sigma_i).$$

Consequently, the expected frequency f_j depends on the μ's and σ's, as well as the π's. In symbols, from (16),

$$f_j(\mu\text{'s}, \sigma\text{'s}, \pi\text{'s}) = f \sum_{i=1}^{M} \pi_i q_{ij}(\mu_i, \sigma_i).$$

Substituting this result in (17) gives

$$A(\mu\text{'s}, \sigma\text{'s}, \pi\text{'s}) = 2 \sum_{j=1}^{N} \hat{f}_j \log [\hat{f}_j / f_j(\mu\text{'s}, \sigma\text{'s}, \pi\text{'s})].$$

In short, if a set of observed frequencies f_j which total f are given, that is,

$$(18) \quad \sum_{j=1}^{N} \hat{f}_j = f,$$

[1]There is no clear precedent for a name for A. Rao (1973; p. 352) refers to a similar statistic as the "Kullback–Liebler separator." Macdonald and Pitcher get their reference to $A/2$ from Kullback (1959). The reason for relating A to "separation" is given in Property 3 later. Alternatively one can speak of "information," or even "entropy." The factor of 2 in (17) makes A approximately a χ^2 statistic. See Appendix B.

then the equations (15)–(17) show explicitly how to calculate A from a given set of population parameters. The problem is to locate a particular set of parameters which minimizes A.

So far, A has been described as a function of the μ's, σ's, and π's. If, however, the means lie on a von Bertalanffy curve (7) determined by l, L, and k, then A can be considered a function of the von Bertalanffy parameters, that is,

$$A = A(l, L, k, \sigma\text{'s}, \pi\text{'s}).$$

Similarly, if the standard deviations depend linearly on the means, following (11) with parameters s and S, then

$$A = A(\mu\text{'s}, s, S, \pi\text{'s}).$$

If both the von Bertalanffy curve (7) and the linearity relation (11) apply, then

$$A = A(l, L, k, s, S, \pi\text{'s}).$$

In short, A is always a function of the population parameters, but the parameters of interest depend on whether or not the means and standard deviations conform to some growth law. If so, then the parameters of the growth law become fundamental. In this way, as stated in the introduction, the previously separate issues of growth and length–frequency analysis become one.

In a typical length–frequency data set, there are length intervals where the observed frequency is small (say, 0, 1, 2, 3, or 4). Some notation is needed to discuss these intervals in a systematic way. For a small number r, let N_r be the number of intervals on which $\hat{f}_j \geq r$. Also let $\hat{g}_r$ be the total number of observations associated with these N_r intervals. In symbols, define

$$(19) \quad g_r = \sum_{j=1}^{N}{}^{(r)} \hat{f}_j,$$

where $\sum^{(r)}$ refers to the sum over those values of j for which $\hat{f}_j > r$. Similarly, let

$$(20) \quad g_r = \sum_{j=1}^{N}{}^{(r)} f_j$$

represent the total expected frequency associated with intervals where $\hat{f}_j > r$.

Notice from (18) and (19) that $\hat{g}_1 = f$ because $\sum^{(1)} \hat{f}_j$ represents the sum of all nonzero frequencies, that is, the total number of observations. By contrast, from the definition (16), the expected frequencies are never zero, inside or outside the observed length range. Consequently, the total expected frequency on intervals where $\hat{f}_j > 1$ is smaller than f; that is, $g_1 < f$. This shows that

$$(21) \quad \sum_{j=1}^{N}{}^{(1)} f_j \leq \sum_{j=1}^{N}{}^{(1)} \hat{f}_j = f,$$

where, ordinarily, the inequality is strict.

To understand the significance of the "minimum A" criterion, the user should be aware of three basic properties of A. These can now be described, with the aid of the notation just defined.

PROPERTY 1. The "minimum A" criterion leads to exactly the same answer as maximum likelihood. However, it is numerically preferable to use A as the criterion, rather than likelihood or log-likelihood.

This should reassure the reader unfamiliar with the separation statistic A. He is actually finding the more familiar maximum likelihood estimates. It turns out that A is closely related to log-likelihood, except for a constant which can be large. This large constant can sometimes mask the behavior of the log-likelihood function, especially when computer precision is limited. Consequently, A tends to be more sensitive to the parameters. Details are given in Appendix B.

PROPERTY 2. In calculating A, one can omit intervals for which there are no observations. In other words, (17) can be replaced by

$$(22) \quad A = 2 \sum_{j=1}^{N}{}^{(1)} \hat{f}_j \log (\hat{f}_j / f_j).$$

The reader may already have noticed (and even been disturbed by) the fact that (17) is not defined if $\hat{f}_j = 0$ for some j, since then the logarithm is not defined. However, in the limiting senses as $\hat{f}_j$ tends to zero, the jth term of A tends to zero. (Recall that f_j is never zero.) This justifies (22). In principle, A can be considered an infinite sum over all possible length intervals; however, the only nonzero terms in the sum are those with $\hat{f}_j > 0$. It follows that A does not depend on how the observed fish lengths are grouped, except through the original choice of intervals. This fact has practical significance, as discussed later in the examples.

PROPERTY 3. The separation statistic A is always positive or zero when condition (21) is true. In fact, A is zero only if

$$(23) \quad f_j = \hat{f}_j$$

for every j.

The fact that A is positive justifies calling it a "separation," particularly since the separation is zero only if expected and observed frequencies agree completely. The inequality $A \geq 0$, which follows from (21), is one of the basic inequalities of information theory. It is proved, for example, by Rao (1973, p. 59). This inequality is not immediately obvious because A contains terms both positive (when $\hat{f}_j > f_j$) and negative (when $\hat{f}_j < f_j$). Notice particularly that, if (23) holds, then $A = 0$ because $\log 1 = 0$. In practice, the predicted frequencies do not exactly equal the observations, and the minimum value of A is positive.

Criteria for Testing the Fit

In length–frequency analysis, the separation statistic A (or, for that matter, any reasonable fitting criterion) is a distinctly nonlinear function of the population parameters. Unlike linear estimation, such as ordinary

linear regression, nonlinear estimation problems do not always have the convenient property that there is only one solution. In fact, length–frequency analysis typically leads to many solutions. By analogy, one can think of the fact in elementary algebra that the simplest nonlinear equation, a quadratic, usually has two solutions, while a linear equation has only one. In practice, it is often possible to reject one quadratic solution on physical grounds. For example, one solution might be negative when the solution of interest is positive. However, there is no general theory about which solution is right; the choice depends on the situation. Similarly in length–frequency analysis, the user must select among competing solutions, according to his knowledge of the situation. No general theory will make that choice.

A related problem is that the user may be unsure of the correct model for his data. He may not know precisely how many age classes there are, or what sort of structure (if any) to impose on the means and standard deviations. Often a diversity of different models will fit the data rather well. This problem is typical for length–frequency analysis; the data can be almost too easy to fit.

A simple criterion for selecting among various possibilities would be to pick the one which gives the lowest minimum value of the statistic A. Unfortunately, as examples given later illustrate, this may result in a poor choice. Generally (although not necessarily), the greater the number of parameters in a model, the lower the minimum value of A. This consideration suggests the usual application of a χ^2 test in which the number of degrees of freedom decreases as the number of estimated parameters increases. Such a test is based on a theorem of statistics, given precisely by Cramér (1946, section 30.3). It states that a certain statistic, evaluated at the maximum likelihood estimates, has asymptotically (for large sample size) a χ^2 distribution.

One way to construct a χ^2 statistic for this problem is to select a small value of r (say 1, 2, 3, or 4), and then lump together all those length intervals on which $\hat{f}_j < r$. The corresponding statistic, in the notation of (19)–(20) above, is then

$$(24)\quad B_r = \sum_{j=1}^{N(r)} \frac{(\hat{f}_j - f_j)^2}{f_j} + \frac{(g_r - \hat{g}_r)^2}{f - g_r}.$$

Here B_r is associated with $N_r + 1$ length groups, namely the N_r intervals where $\hat{f}_j \geq r$ and one further group of intervals with $f - \hat{g}_r$ observations. This suggests a number of degrees of freedom

$$(25)\quad D_r = N_r - P,$$

where P is the number of parameters estimated.

Unfortunately, Cramér's theorem cannot be applied directly to the statistic B_r for two reasons. First, except when $r = 1$, the groups of intervals used in calculating B_r differ from those used in obtaining the likelihood estimates (by minimizing A). Second, the method of selecting groups of intervals for B_r depends directly on the observed frequencies f_j.[2] In fact, the distribution of B_r is simply not known in general. It seems reasonable to conjecture that it might be approximately χ^2 distributed, at least for values of r large enough to guarantee several observations in every length group (say, $r = 3$ or 4). We do not attempt here to develop a complete mathematical theory for the use of B_r. Instead, we take a strictly pragmatic approach, as described in the next paragraph.

For a given model, we first compute the required parameter estimates by minimizing A. This avoids any arbitrary grouping of the data, other than the initial choice of length intervals, at least as far as the estimates are concerned. (See Property 2 for A, and the subsequent discussion, above.) Once the estimates are known, we then calculate B_r for several values of r and compute the corresponding χ^2 percentage levels for D_r degrees of freedom. (The necessary formulas are given in Appendix C). In comparing solutions obtained from various models, where each model may have more than one minimum point, we simply rank the results according to the χ^2 percentage levels. In this way, we can explore two questions: (1) are the rankings consistent for various values of r, and (2) do the highest ranked choices appear most reasonable biologically? The answers suggest that the statistic B_r can be a useful guide, but not a final criterion, for sorting out multiple solutions to a length–frequency estimation problem.

Appendix D gives details of computer methods used to implement all the procedures described in the previous paragraph.

Example 1. Northern Pike

The first data set considered here pertains to Northern pike (*Esox lucius*) from Heming Lake, Manitoba. It was originally presented by Macdonald (1969) and then reanalyzed a decade later by Macdonald and Pitcher (1979). The observed frequency data for this example and the next one are listed in Table 2. These data (previously published in graphical form) are needed for reference in the discussion here.

As Macdonald (1969) describes, the pike sample consists of 523 fish. The length of each is known, as well as an estimate of its age (between 1 and 5 yr) from scale analysis. This allows the results from length–frequency analysis to be compared with a standard. The first line in Table 3 (example 1.1) shows the population parameters, i.e. μ's, σ's, and π's, determined by scale reading. The next line (example 1.2) shows the final results from length–frequency analysis published by Macdonald and Pitcher (1979). Obviously, the agreement between examples 1.1 and 1.2 is quite good.

[2] In theory, the method of grouping should depend on the expected frequencies f_j. Since these frequencies are usually unknown, this requirement of Cramér's theorem is almost always violated in practice. The statistic B_r is at least based on a prescribed rule for grouping, not an ad hoc choice of the investigator.

TABLE 2. Observed frequency data for pike and abalone. Frequencies should be read first across and then down. (For example, $\hat{f}_1 = 4$, $\hat{f}_2 = 10$, $\hat{f}_3 = 21$ in the pike data.)

Pike frequencies. $N = 30$, $x_1 = 19$, $x_{30} = 77$, $w = 2$, $f = 523$									
4	10	21	11	14	31	39	70	71	44
42	36	23	22	17	12	12	11	8	3
6	6	3	2	1	1	1	0	1	1
Abalone frequencies. $N = 62$, $x_1 = 8$, $x_{62} = 130$, $w = 2$, $f = 431$									
2	7	7	4	0	0	0	0	0	1
1	4	3	5	7	3	5	1	0	3
0	3	10	3	2	5	9	8	8	15
8	11	13	10	15	13	11	12	14	12
17	14	17	18	20	10	11	10	11	9
7	8	10	5	6	4	4	7	3	4
0	1								

Macdonald and Pitcher arrive at the estimates 1.2 by assuming $\sigma_3 = 4$, $\sigma_4 = 5$, and $\sigma_5 = 6$, as shown. Before minimizing A, they also lump the last six frequencies (1, 1, 1, 0, 1, 1 in Table 2) to obtain a final group of five fish. They argue (p. 991) that this makes the model more robust to nonnormality. Nevertheless, it should be recognized that lumping constitutes an ad hoc decision with regard to the data. One could, instead, lump the last four or five frequencies into a single group. A consequence of this procedure is that the lumped data set does not distinguish between five pike with lengths 67 cm and five pike with respective lengths 67, 69, 71, 75, and 77 cm, as observed. Example 1.3 (Table 3) shows the estimates that are obtained without lumping the data. Not surprisingly, μ_5 turns out to be larger. In fact, because of the fixed standard deviations σ_3, σ_4, and σ_5, the last three means all turn out to be larger in example 1.3 than in 1.2. This shift results in an increase (from 49.5 to 55.6%) in the apparent proportion of age 2 fish.

Example 1.4 (Table 3) shows a minimum point for A obtained without lumping the data, but with the assumption that the standard deviations are linear on the means, as in (11). Notice that, like solution 1.2 of Macdonald and Pitcher (1979), example 1.4 also gives reasonable agreement with the results from scale analysis (example 1.1). The interesting difference is that 1.4 is obtained (i) without assuming explicit values for σ_3, σ_4, and σ_5 and (ii) without ad hoc lumping of the data. Only the linearity relation (11) is assumed. Example 1.5 shows a similar result based on the assumption that the standard deviations increase linearly with age, as in (12). Comparison between examples 1.4 and 1.5 shows that the two linearity assumptions (11) and (12) lead to very nearly the same result.

This discussion is not intended as criticism of solution 1.2 obtained by Macdonald and Pitcher (1979), but rather as an illustration of one more point of view. Part of their approach consists in lumping the data, as deemed appropriate, for robustness. Our approach is to avoid decisions in regard to the data as discussed earlier in connection with Property 2 for A. Instead, we put decisions on structure directly into hypotheses of the model. We utilize data lumping later, only as a tool for assessing the fit. A useful outgrowth of our approach in this case is that we need make no assumptions on particular values for the σ's. Our unknown standard deviations are s and S. In Macdonald and Pitcher's solution 1.2, the unknown deviations are σ_1 and σ_2, while σ_3, σ_4, and σ_5 are presumed known.[3]

Some assumption about standard deviations is certainly necessary for this problem. Examples 1.6 and 1.7 illustrate two minima for A in which there is no restriction on the σ's. (Incidentally, these examples show how A can have more than one local minimum.) Both solutions exhibit a peculiar feature: one of the standard deviations turns out extremely small. In each case an entire age-class is used to explain some minor amount of noise in the data. For instance, in example 1.6 the fifth age-class consists entirely of the two observations which comprise the small extreme right-hand mode of the histogram of observed frequencies shown in Fig. 1. From that point of view, example 1.6 would be an excellent fit if indeed the final mode comprised a whole age-group. In fact, both examples 1.6 and 1.7 suggest that, without some structure imposed on the standard deviations, the analysis points to a solution comprising only four groups with significant numbers of fish. Example 1.8 shows the solution obtained with this assumption, $M = 4$. Aside from components with minor numbers of fish, examples 1.6 to 1.8 look much the same. Biologically, example 1.8 is very reasonable if the combined group of age 4's and 5's is regarded as a single group of undistinguishable fish.

It is well known (see, for example, Macdonald and

[3]While the final draft of this paper was in revision, we discussed these results with Prof Macdonald and received a letter in reply showing another interesting result obtained with log-normal components. Certainly, no single approach can be considered definitive. Ordinarily, the practitioner will try several approaches to discover that which best suits his knowledge about the biological background for the data. In making a final decision, the χ^2 analysis given here may prove useful.

TABLE 3. Examples of means, standard deviations, and percentages for various fits to the pike data. All examples except numbers 1.1 and 1.2 correspond to minima for A.

Example no.	Mean lengths (cm)					Standard deviations (cm)					Percentages (as fractions)					A
	μ_1	μ_2	μ_3	μ_4	μ_5	σ_1	σ_2	σ_3	σ_4	σ_5	π_1	π_2	π_3	π_4	π_5	
1.1[a]	23.33	33.09	41.27	51.24	61.32	2.44	3.00	4.27	5.08	7.07	0.105	0.465	0.298	0.090	0.042	21.28
1.2[b]	22.50	33.10	40.04	48.57	60.02	1.92	3.40	4.00	5.00	6.00	0.085	0.495	0.238	0.128	0.054	19.70
1.3	22.66	33.73	42.13	52.58	65.15	1.80	3.63	4.00	5.00	6.00	0.082	0.559	0.225	0.110	0.024	17.14
1.4	22.67	33.44	40.73	51.27	62.77	1.80	3.53	4.70	6.40	8.24	0.082	0.498	0.265	0.128	0.027	16.20
1.5	22.68	33.29	39.77	50.58	62.10	1.81	3.44	5.07	6.70	8.33	0.083	0.448	0.306	0.134	0.029	16.16
1.6	22.65	33.46	40.48	52.90	76.02	1.79	3.54	5.51	7.78	0.49	0.082	0.469	0.310	0.136	0.003	13.29
1.7	22.74	33.46	40.11	41.77	50.98	1.84	3.37	0.65	5.90	10.11	0.084	0.484	0.034	0.241	0.157	14.98
1.8	22.61	33.65	41.04	51.59	—	1.76	3.67	5.92	9.85	—	0.080	0.490	0.279	0.151	—	16.84

[a]True results from scale reading, published by Macdonald (1969).
[b]Length–frequency solution published by Macdonald and Pitcher (1979).

Pitcher 1979, p. 991) that solutions like 1.6 and 1.7 with standard deviations less than the interval width may be meaningless. The question is, how can they be avoided? For all the examples in Table 3, the values of A are lowest by far in examples 1.6 and 1.7. Search algorithms to minimize A will try to reach these points. Ad hoc constraints, such as lower bounds on the σ's, may simply result in a solution tight on the constraints as the algorithm seeks the low point. In a biological context, it may be more appropriate to invoke a general hypothesis, such as linearity of the σ's, to avoid unreasonable minima. In other contexts, such as the analysis of physical X-ray spectra, this hypothesis may be completely inappropriate. These remarks illustrate a general point of this paper: the structure of the σ parameters often distinguishes biological applications of mixture analysis from applications in other fields. The pike example underscores the importance of the σ's. In practice, the biologist may be far more interested in the π's and μ's than he is in the σ's, yet the structure of the σ's may determine the solution he finds.

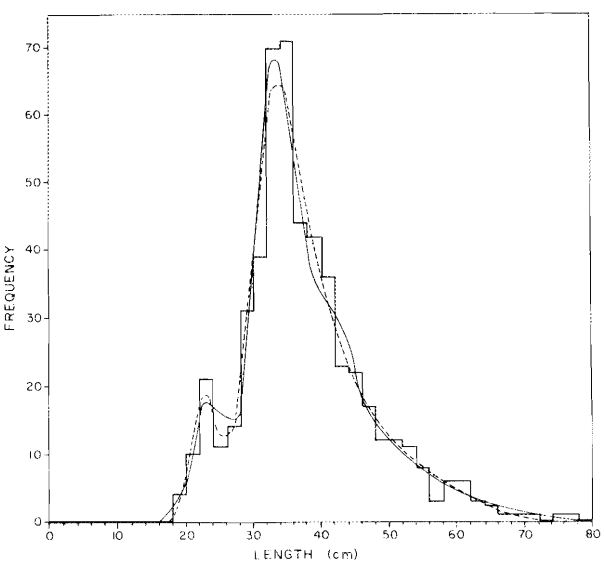

FIG. 1. Length–frequency histogram and two curves of expected frequencies for the pike data. The solid curve pertains to example 1.1, and the broken curve to example 1.4.

All examples in Table 3 lead to a good match between expected and observed frequencies. Figure 1 shows the expected frequency curves for examples 1.1 (solid curve) and 1.4 (broken curve) in relation to the observed frequency histogram. Visually, both curves are close to each other and to the histogram. Similarly, the other examples in Table 3 also correspond to reasonable fits from the point of view of matching the observations, even though the values of A vary considerably among the examples. This illustrates the multiplicity problem discussed earlier: the data are almost too easy to fit.

Table 4 shows the results of a χ^2 analysis applied to solutions 1.3 to 1.8 from Table 3. In each case, both the statistic B_r and the corresponding χ^2 level are calculated for $r = 1$, 2, 3, and 4. The resulting χ^2 levels are then used to rank the various solutions. Table 5 lists the rankings obtained in this way. As described earlier, the theoretical basis for this process may be incomplete, but the results are interesting. Rankings are identical with $r = 2$ or 3, and very similar with $r = 4$. For all three values, $r = 2$, 3, or 4, the highest ranked solutions are the most biologically reasonable ones, namely 1.4 and 1.5 in which the σ's increase linearly. The two lowest ranked solutions, 1.6 and 1.7, are those in which one of σ's is unacceptably small. Examples 1.8 (with only four distinguishable age-groups) and 1.3 (with three prescribed σ's) are ranked in the middle. In short, for $r = 2$, 3, or 4, the χ^2 rankings correspond well with biological validity in the results.

When $r = 1$, the χ^2 rankings are less meaningful biologically because the unacceptable solution 1.6 ranks essentially the same as the acceptable solutions 1.4 and 1.5. Table 4 shows χ^2 levels for these three cases which are almost identical (65.1, 65.4, 65.5%) when $r = 1$.

TABLE 4. Quantities for assessing the fit of examples in Table 3. Examples 1.1 and 1.2 are excluded because they do not correspond to minima for A.

Example no.	P	D_r				A	B_r				χ^2 level (%)			
		$r=1$	$r=2$	$r=3$	$r=4$		$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
1.3	11	18	13	12	10	17.14	16.65	12.24	12.22	11.08	54.8	50.8	42.8	35.1
1.4	11	18	13	12	10	16.20	15.10	11.39	11.38	10.38	65.5	57.8	49.7	40.8
1.5	11	18	13	12	10	16.16	15.16	11.34	11.34	10.43	65.1	58.2	50.0	40.4
1.6	14	15	10	9	7	13.29	12.33	11.16	11.16	10.18	65.4	34.5	26.5	17.8
1.7	14	15	10	9	7	14.98	14.72	9.81	9.81	9.26	47.1	45.7	36.6	23.4
1.8	11	18	13	12	10	16.84	16.73	11.70	11.70	11.20	54.2	55.2	47.0	34.4

This difficulty confirms expectations about problems with a χ^2 analysis when $r = 1$. The theory suggests that B_1 simply may not be χ^2 distributed because there may be too few observations in some length-groups.

A related comment applies to the statistic A. For theoretical reasons given in Appendix B, the statistic A is approximately equal to B_1. (See also footnote 1.) This approximation is evident in the examples of Table 4. However, in view of the problems with B_1, it may be wrong to suppose that A is χ^2 distributed. This difficulty underlies the approach taken here whereby the "minimum A" criterion is used to obtain parameter estimates and then the statistic B_r ($r \geq 2$) is used to rank various proposed solutions. In practice, it appears advisable to check several values of r to see how the rankings are affected.

Example 2. Northern Abalone

The second data set considered here pertains to Northern abalone (*Haliotis kamtschatkana*) from the Queen Charlotte Islands, British Columbia.

It appears, along with numerous other data, in the report by Breen and Adkins (1979). The particular abalone population considered here is a composite associated with all *Nereocystis* communities in the sampled region. (See Fig. 54 of the report.) There are two minor modifications of the published data. Three large animals, deemed by P. A. Breen (personal communication) to lie in a separate group from all the rest, are omitted. Also, to reduce computation, the frequencies are grouped in 2-mm intervals, rather than 1-mm as reported. This grouping probably reflects realistic limits in determining abalone size (P. A. Breen personal communication). The data, so revised, are listed in Table 2.

There is no known independent method of aging this species. Consequently, a standard, like example 1.1 for the pike data, is not available for comparison with results from length–frequency analysis. The precise number of age-classes is not even known, but it is believed to be much larger than 5. The length–frequency histogram in Fig. 2, with its many modes, suggests that this might be true. In spite of these problems, some independent information is available on growth. Abalone have been tagged and recovered 1 yr later to measure annual growth (Quayle 1971). Using a Walford plot of these size data, one can estimate L_∞ and k. (See Ricker 1975.) In this way Breen (1980) obtains a value of L_∞ equal to 128.9 mm and k equal to 0.766.

In view of this evidence, it is reasonable to investigate the results of length–frequency analysis based on von Bertalanffy growth. As a start, consider some possible age-groups suggested by modes in the histogram

TABLE 5. Example numbers placed in rank order by the χ^2 levels in Table 4. For each value of r, the six examples are ranked highest to lowest, left to right.

r	Rank order					
	1	2	3	4	5	6
1	1.4	1.6	1.5	1.3	1.8	1.7
2	1.5	1.4	1.8	1.3	1.7	1.6
3	1.5	1.4	1.8	1.3	1.7	1.6
4	1.4	1.5	1.3	1.8	1.7	1.6

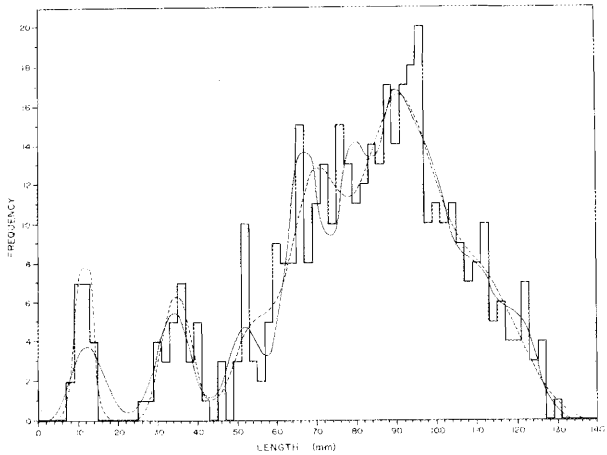

FIG. 2. Length–frequency histogram and two curves of expected frequencies for the abalone data. The solid curve pertains to example 2.1, and the broken curve to example 2.4.

TABLE 6. Examples of means (mm), standard deviations (mm), and percentages (as fractions) for various fits to the abalone data. All examples correspond to minima for A.

Example no.	Quantity	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$	$i = 11$	A
	Modes[a]	11	36	52	66	76	86	96	102	112	?	122	
	Est. μ's[b]	11	31.7	49.4	64.3	77.1	87.9	97.1	104.9	111.5	117.2	122	
2.1	μ's	12.38	33.90	51.84	66.80	79.27	89.67	98.34	105.56	111.59	116.61	120.80	66.13
	σ's	4.25	4.25	4.25	4.25	4.25	4.25	4.25	4.25	4.25	4.25	4.25	
	π's	0.046	0.069	0.059	0.170	0.163	0.182	0.135	0.051	0.067	0	0.058	
2.2	μ's	11.29	34.76	53.17	67.61	78.93	87.81	94.78	100.25	104.53			51.33
	σ's	2.05	3.35	4.65	5.95	7.25	8.55	9.85	11.15	12.45			
	π's	0.047	0.068	0.058	0.178	0.071	0.241	0.015	0.032	0.290			
2.3	μ's	11.28	34.98	53.75	68.60	80.36	89.67	97.03	102.87	107.48	111.14		46.55
	σ's	1.84	3.66	5.10	6.24	7.14	7.86	8.42	8.87	9.22	9.50		
	π's	0.046	0.069	0.063	0.200	0.074	0.274	0.089	0	0	0.185		
2.4	μ's	11.29	34.87	53.87	69.19	81.53	91.48	110.64					46.26
	σ's	1.83	3.66	5.14	6.33	7.29	8.06	9.55					
	π's	0.046	0.069	0.065	0.209	0.084	0.329	0.198					

[a]From Fig. 2.
[b]From a von Bertalanffy curve with $l = 11$, $L = 122$, $k = 0.85$ (or $L_\infty = 149.2$, $K = 0.1625$, $t_0 = 0.5288$).

of Fig. 2. Table 6 lists a set of 10 major modes selected by eye. The nine gaps between them (25, 16, 14, 10, 10, 10, 6, 10, 10 mm, respectively) appear to shrink somewhat in von Bertalanffy fashion, but with a problem toward the end. One possible way to deal with this problem is to add an extra age-class, thus making two small gaps in place of one large one at the end. For instance, the parameters $l = 11$ mm, $L = 122$ mm, and $k = 0.85$ with 11 age-classes generate a set of means reasonably close to the observed modes, as shown in Table 6. Here l and L correspond to the first and last observed modal lengths, and k is found by trial and error to give reasonable approximations for the remaining modes. In this way, the histogram itself suggests first estimates for l, L, and k. Incidentally, the corresponding parameters, $L_\infty = 149.2$ mm, $K = 0.1625$, and $t_0 = 0.5288$ are not amenable to such simple intuitive motivation. (For example, L_∞ lies well beyond the range of data.) This illustrates the utility of the new von Bertalanffy parameters for length–frequency analysis.

When the eye scans the histogram in Fig. 2 and seeks to identify age-classes with the various modes, it tends to pick out clusters with roughly the same width. Example 2.1 in Table 6 reflects this perception. It is a solution with means on a von Bertalanffy curve, constant standard deviations, and 11 age-classes (in accord with the last paragraph). The solid curve in Fig. 2 represents the predicted frequencies for this case. The various age-groups show up distinctly as either modes or bends in the curve. Not surprisingly, in Table 6 the age-class next to the last registers zero percentage of the fish, since it appears as an artifact to accommodate von Bertalanffy growth. In Table 7, the parameters l, L, and k for example 2.1 turn out remarkably close to the initial estimates of 11, 122, and 0.85, respectively. In this case there are 14 parameters (l, L, k, s, and 10 π's), rather than 32 as would be required by an unstructured analysis of 11 age-classes. Unfortunately, in spite of the comparatively low number of parameters, Table 8 shows that the fit is still not good enough. Example 2.1 would be rejected by χ^2 analysis, for each value of r, even at the 5% cutoff level. Apparently, the requirement of constant standard deviations is too restrictive. In fact, Fig. 2 suggests this. For instance, the solid curve fits the first age-class poorly, because the estimated value of σ_1 is too large. This forces the first mode to be too wide and low compared with the corresponding mode on the histogram.

Example 2.2 (Table 6) seeks to overcome this problem by allowing the standard deviations to be linear on the age. The value of A drops, and the first age-class is accommodated much better. However, the σ's now become very large for the older age-classes, because they must increase by a fixed amount from each age to the next. In fact, σ_9 is so large that the ninth age-group accounts for all observations above 104 mm. This suggests a different perception of the histogram in Fig. 2. Perhaps the modes at higher lengths do not correspond to separate age-classes, but are just noise on the descending limb of a single normal curve.

One way to avoid the rapid growth in standard deviations of example 2.2 is to let them be linear on the means, so that the σ's reach a limiting size with the μ's. Example 2.3 illustrates a solution with this assumption. The results in Table 6 are similar to example 2.2, except for the higher age-groups. In particular, example 2.3 contains a final age-group quite isolated from all the rest. It appears that for the abalone population, with its numerous age-groups, the particular linearity assumption, (11) or (12), on the σ's may be important. This contrasts with the situation for pike, where either

TABLE 7. Parameters for the means and standard deviations in the fits to the abalone data cited in Table 6. The standard deviations are either constant (CO), linear on age (LA), or linear on the means (LM).

Example no.	Age-classes	l	L	k	L_∞	K	t_0	σ's	s	S
2.1	11	12.38	120.80	0.8337	141.78	0.1819	0.4978	CO	4.25	—
2.2	9	11.29	104.53	0.7843	120.12	0.2429	0.5936	LA	2.05	12.45
2.3	10	11.28	111.14	0.7916	125.02	0.2337	0.5954	LM	1.84	9.50
2.4	6	11.29	91.48	0.8059	132.78	0.2158	0.5882	LM	1.83	8.06

TABLE 8. Quantities for assessing the fit of examples in Table 6.

Example no.	P	D_r				A	B_r				χ^2 level (%)			
		$r=1$	$r=2$	$r=3$	$r=4$		$r=1$	$r=2$	$r=3$	$r=4$	$r=1$	$r=2$	$r=3$	$r=4$
2.1	14	40	36	34	28	66.13	57.69	54.19	53.02	49.14	3.5	2.6	2.0	0.8
2.2	13	41	37	35	29	51.33	49.02	43.44	42.93	37.83	18.2	21.6	16.8	12.6
2.3	14	40	36	34	28	46.55	43.65	39.46	38.73	34.26	31.9	31.8	26.5	19.2
2.4	12	42	38	36	30	46.26	43.33	39.37	38.62	34.17	41.4	40.8	35.2	27.4

assumption gave similar results. (See examples 1.4 and 1.5.)

Examples 2.2 and 2.3, especially the isolated final age-group of 2.3, suggest a new approach to the analysis. Perhaps the animals grow rapidly at first, forming identifiable age-classes, and then finally growth slows so much that all animals after a certain age appear as a single group. These ideas motivate example 2.4. Here the first six age-classes have means on a von Bertalanffy curve, while the last mean is independent of the rest. The standard deviations, including the last, are presumed linear on the means. There are 12 parameters: l, $L(=\mu_6)$, k, μ_7, s, S, and 6 π's. The predicted frequencies for this case are represented by the broken curve in Fig. 2. It is less undulating than the solid curve (for example 1.1), reflecting fewer age-classes. The general fit, especially for the first two age-classes, is distinctly better than in example 1.1. Notice also that the broken curve descends steadily through the higher modes of the histogram, which are regarded by the statistical model as noise.

Of the four length–frequency solutions given here for the abalone data, example 2.4 is certainly the most attractive. Table 8 shows that even though example 2.4 has the fewest parameters, it gives rise to the lowest minimum value of A. Not surprisingly, then, the χ^2 analysis ranks example 2.4 highest. Incidentally, in this case, unlike the earlier one for pike, the ranking of solutions is the same for $r = 1$, 2, 3, or 4. Finally, Table 7 shows that example 2.4 corresponds to values of L_∞ (132.78 mm) and k (0.8059) which agree reasonably well with the values ($L_\infty = 128.9$ mm, $k = 0.766$) obtained independently from other data by Breen (1980).

These four examples certainly do not exhaust the possibilities for discussion of the abalone data, and they should not be regarded as final biological conclusions. They do, nevertheless, illustrate how growth structure aids the discussion of length–frequency analysis. Notice particularly how these methods assist in determining the number of age-classes. Indeed, part of the abalone analysis involves deciding how many age-classes are discernible. Growth structure appears to be a useful tool in making that decision. Possibilities for future analysis of the abalone data include a parametric description of the π's to reflect mortality and less restrictive descriptions of the μ's and σ's.

Conclusions

The examples in the previous two sections are intended to help the reader develop intuition on the use of growth structure in length–frequency analysis. Although the problem of multiple solutions makes a general theory impossible, the examples do suggest several basic conclusions or guidelines. These are presented here.

1. Because length–frequency analysis may lead to many solutions for the same data set, subjective decisions must often be made on biological grounds. One way to introduce biological opinion is to require that the means and standard deviations conform to an assumed growth model.

2. The χ^2 level associated with the statistic B_r in (22) can be helpful for choosing among competing solutions, although it should not be regarded as definitive. Solution rankings by χ^2 level tend to be consistent for low values of $r > 1$. On purely statistical grounds, the χ^2 analysis may suggest actually rejecting only a few of many competing solutions.

3. Even if the percentages and means are the parameters of greatest interest, the structure of the standard deviations may be the most important feature in determining which solution is obtained. If the σ's are assumed to have linear growth, it is generally more realistic to presume linearity on the means (11), rather than on the ages (12), particularly when the number of age-classes is large.

4. A von Bertalanffy structure for the means is sometimes too restrictive, but it can be useful for suggesting the number of discernible age-classes when that number is unknown. It can also be used to check the results of length–frequency analysis against an independent measure of growth from a tagging study over 1 yr.

5. Reasonable estimates for l, L, and k can often be obtained directly from the length–frequency histogram. By contrast, the corresponding values of L_∞, K, and t_0 may be much less apparent and even misleading.

6. In the minimization process, the parameters l, L, and k tend to be much more stable numerically than L_∞, K, and t_0.

7. Length–frequency analysis tends to lump the final age-classes together if they are in close proximity or contain small percentages of fish. In such cases it may be impossible to distinguish the final ages, and the best approach may be to assume that all fish beyond a certain age comprise a single group.

Acknowledgments

We are grateful to Prof Peter Macdonald for numerous helpful comments on this problem. He sent us an advance copy of his paper with Dr Pitcher, a listing of the programs used in their data analysis, and a sample run. We have been very fortunate to have their published work available as an example for investigating how growth-structured analysis might differ from earlier methods. We are also grateful to Dr Paul Breen for extensive advice regarding the abalone data which he and Bruce Adkins collected in 1978.

ANON. 1964. Handbook of mathematical functions. U.S. National Bureau of Standards, Washington, D.C.

BENYON, P. R. 1976. Remark AS R15, function minimization using a simplex procedure. Appl. Stat. 25: 97.

VON BERTALANFFY, L. 1938. A quantitative theory of organic growth. Hum. Biol. 10: 181–213.

BREEN, P. A. 1980. Measuring the fishing intensity and annual production in the abalone fishery of British Columbia. Can. Tech. Rep. Fish. Aquat. Sci. (In press)

BREEN, P. A., AND B. E. ADKINS. 1979. A survey of abalone populations on the east coast of the Queen Charlotte Islands, August 1978. Fish. Mar. Serv. MS Rep. 1490: 125 p.

CASSIE, R. M. 1954. Some uses of probability paper in the analysis of size frequency distributions. Aust. J. Mar. Freshwater Res. 5: 513–522.

CHAMBERS, J. M., AND J. E. ERTEL. 1974. Remark AS R11, a remark on Algorithm AS 47, function minimization using a simplex procedure. Appl. Stat. 23: 250–251.

CRAMÉR, H. 1946. Mathematical methods of statistics. Ninth printing, 1961. Princeton Mathematical Series, vol. 9, Princeton University Press, Princeton, NJ.

FORD, E. 1933. An account of the herring investigations conducted at Plymouth during the years from 1924–1933. J. Mar. Biol. Assoc. U.K. 19: 305–384.

HARDING, J. P. 1949. The use of probability paper for graphical analysis of polymodal frequency distributions. J. Mar. Biol. Assoc. U.K. 28: 141–153.

HASSELBLAD, V. 1966. Estimation of parameters for a mixture of normal distributions. Technometrics 8: 431–444.

HILL, I. D. 1978. Remark AS R28, a remark on Algorithm AS 47, function minimization using a simplex procedure. Appl. Stat. 27: 380–382.

KNIGHT, W. 1968. Asymptotic growth: an example of nonsense disguised as mathematics. J. Fish. Res. Board Can. 25: 1303–1307.

KULLBACK, S. 1959. Information theory and statistics. Wiley, New York. 395 p.

MACDONALD, P. D. M. 1969. FORTRAN programs for statistical estimation of distribution mixtures: some techniques for statistical analysis of length–frequency data. Fish. Res. Board Can. Tech. Rep. 129: 45 p.

MACDONALD, P. D. M., AND T. J. PITCHER. 1979. Age groups from size-frequency data: a versatile and efficient method of analysing distribution mixtures. J. Fish. Res. Board Can. 36: 987–1001.

MCNEW, R. W., AND R. C. SUMMERFELT. 1978. Evaluation of a maximum likelihood estimator for analysis of length–frequency distributions. Trans. Am. Fish. Soc. 107: 730–736.

MELSA, J. L., AND A. P. SAGE. 1973. An introduction to probability and stochastic processes. Prentice-Hall Elec. Eng. Series. Prentice-Hall, Englewood Cliffs, NJ.

NELDER, J. A., AND R. MEAD. 1965. A simplex method for function minimization. Comput. J. 7: 308–313.

O'NEILL, R. 1971. Algorithm AS 47, function minimization using a simplex procedure. Appl. Stat. 20: 338–345.

1974. Corrigendum, Algorithm AS 47, function minimization using a simplex procedure. Appl. Stat. 23: 252.

PEARSON, K. 1894. Contributions to the mathematical theory of evolution. Philos. Trans. R. Soc. London, Ser. A, 185: 71–110.

PETERSEN, C. G. J. 1892. Fiskensbiologiske forhold i Holboek Fjord, 1890–91. Beret. Dan. Biol. St. 1890(91), 1: 121–183.

QUAYLE, D. B. 1971. Growth, morphometry, and breeding in the British Columbia abalone (*Haliotis kamtschatkana* Jonas). Fish. Res. Board Can. Tech. Rep. 279: 84 p.

RAO, C. R. 1973. Linear statistical inference and its applications. 2nd ed. John Wiley and Sons. New York, London, Sydney, Toronto. 625 p.

RICKER, W. E. 1975. Computation and interpretation of biological statistics of fish populations. Fish. Res. Board Can. Bull. 191: 382 p.

VAN DER WAERDEN, B. L. 1969. Mathematical statistics. 2nd (English) ed. Springer-Verlag. New York, Heidelberg, Berlin. 367 p.

WALFORD, L. A. 1946. A new graphic method of describing the growth of animals. Biol. Bull. 90: 141–147.

Appendix A. Derivation of Equation (7)

Equation (7), which is an alternative form of the von Bertalanffy growth relationship (2), follows from three assumptions, namely,

(A1) $\mu_{t+2} - \mu_{t+1} = k(\mu_{t+1} - \mu_t)$,

(A2) $\mu_1 = l,$

(A3) $\mu_M = L.$

Here, (A1) follows from (3)–(4), and (A2) and (A3) are the definitions of l and L. When $k = 1$, (A1) shows that the means are evenly spaced, as stated in connection with (7). When $k \neq 1$, (A1) can be regarded as a second-order difference equation. It is linear with constant coefficients. The boundary conditions for (A1) are (A2)–(A3).

One possible solution to (A1) is

$$\mu_t = \alpha$$

for some constant α. Another possible solution, as the reader may verify is

$$\mu_t = \beta k^t$$

for some constant β. The theory of difference equations states that any solution to (A1) must be a combination of these two solutions, that is,

(A4) $\mu_t = \alpha + \beta k^t.$

Substituting (A4) in (A2) and (A3) gives

(A5) $l = \alpha + \beta k,$

(A6) $L = \alpha + \beta k^M.$

Equations (A5)–(A6) can be solved for α and β. Substituting the solutions in (A4) gives (7).

Another approach, independent of the one just given, involves showing algebraically that (4)–(6) and (8)–(10) are inverse transformations of each other. It can then be verified that (4)–(6) transforms (2) into (7), and (8)–(10) transforms (7) into (2).

Appendix B. Likelihood, Separation A, and χ^2

The log-likelihood function C associated with the observations $\hat{f}_j$ is given by

(B1) $C = \sum_{j=1}^{N} \hat{f}_j \log (f_j/f).$

(See, for example, van der Waerden 1969, p. 186). In the sum (B1), the jth term is 0 when $\hat{f}_j = 0$, consequently,

(B2) $C = \sum_{j=1}^{N}{}^{(1)} \hat{f}_j \log (f_j/f),$

where the special $\sum$ notation is defined in connection with (19). It follows from (B2) and (22) that

(B3) $A = -2C + 2 \sum_{j=1}^{N}{}^{(1)} \hat{f}_j \log (\hat{f}_j/f).$

In (B3) notice that, while A and C depend on the population parameters, the summation depends only on the observations. As functions of the parameters, then, the separation statistic A is minus twice the log-likelihood plus a constant. This shows that C is a maximum when A is a minimum, and vice versa, as stated in Property 1.

The constant term in (B3) is always negative, and it can be quite large. For example, with the pike data it is −2991.29. In Table 3, the values of A go from about 13 to 21, producing a change in the first decimal place from the smallest to the largest value. By contrast, the values of C for these examples would go roughly from −1506 to −1502, resulting only in a change of the fourth decimal place. This substantiates the claim in Property 1 that, if computer precision is limited, A makes a better objective function than C.

The value of A has the added advantage that it is more meaningful than C because it approximates the χ^2 statistic B_1. If x is near 1, then

(B4) $\log x \cong (x - 1) - \frac{1}{2}(x - 1)^2.$

When A is minimized, the expected frequency f_j is near the observed $\hat{f}_j$, so that the ratio $f_j/\hat{f}_j$ is near 1. It follows from (22) and (B4) that

$$A = -2 \sum_{j=1}^{N}{}^{(1)} \hat{f}_j \log (f_j/\hat{f}_j)$$

$$\cong -2 \sum_{j=1}^{N}{}^{(1)} \hat{f}_j \left[\left(\frac{f_j}{\hat{f}_j} - 1 \right) - \tfrac{1}{2} \left(\frac{f_j}{\hat{f}_j} - 1 \right)^2 \right]$$

$$= -2 \sum_{j=1}^{N}{}^{(1)} (\hat{f}_j - f_j) + \sum_{j=1}^{N}{}^{(1)} (f_j - \hat{f}_j)^2/\hat{f}_j$$

$$= 2(f - g_1) + \sum_{j=1}^{N}{}^{(1)} (\hat{f}_j - f_j)^2/\hat{f}_j.$$

In the final sum above, one can substitute f_j for $\hat{f}_j$ in the denominator with the same degree of approximation already used in applying (B4). This gives

(B5) $A \cong 2(f - g_1) + \sum_{j=1}^{N}{}^{(1)} (\hat{f}_j - f_j)^2/f_j.$

The definition (24) for B_1, taken with the fact that $g_1 = f$ (as discussed in connection with (19)), shows that

(B6) $B_1 = (f - g_1) + \sum_{j=1}^{N}{}^{(1)} (f_j - f_j)^2/f_j.$

It follows from (B5)–(B6) that

(B7) $A \cong (f - g_1) + B_1.$

Ordinarily, g_1 (the sum of expected frequencies on intervals with $\hat{f}_j \geq 1$) is close to f at minimum A. Since $f > g_1$, (B7) suggests that A should typically be slightly larger than B_1, as it is in all the examples of Tables 4 and 8.

Appendix C. The Normal Integral and χ^2 Levels

At biological laboratories, the library of mathematical literature is sometimes quite limited. As a result, some readers of this paper may have difficulty locating methods to calculate the normal integral (15) and to determine χ^2 levels associated with the statistic B_r. For convenience, suitable formulas are cited here.

Hasselblad (1966) suggests calculating q_{tj} in (15) by the approximation

(C1) $q_{tj} = \frac{w}{\sigma_t \sqrt{2\pi}} \exp \left[-\tfrac{1}{2} \left(\frac{x_j - \mu_t}{\sigma_t} \right)^2 \right].$

Unfortunately, (C1) can be a poor approximation when w is as large as one of the σ's. This is the case in all pike examples, where σ_1 is about 1.8 and w is 2. For the sake of improving (C1), define a function $F(z)$ sequentially as

follows:

$$t = (1 + b_0|z|)^{-1},$$

$$u = \sum_{i=1}^{5} b_i t^i,$$

$$v = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2),$$

$$F(z) = \begin{cases} 1 - uv, z > 0 \\ uv, z < 0, \end{cases}$$

where the constants b_0 to b_5 are, respectively, 0.2316419, 0.319381530, −0.356563782, 1.781477937, −1.821255978, and 1.330274429. Also define z_+ and z_- by

$$z_{+,-} = \frac{x_j - \mu_i}{\sigma_i} \pm \frac{w}{2\sigma_i}.$$

Then a good approximation to q_{ij} is

(C2) $q_{ij} = F(z_+) - F(z_-)$.

It turns out that the approximations (C1) and (C2) are close if w is small compared to σ_i because then the difference $z_+ - z_-$ is small compared to 1. The approximations leading to (C2) come from Anon. (1964).

The calculation of χ^2 levels involves two functions defined for $z > 0$ and an integer n, namely,

$$G_1(z, n) = (z/2)^{n/2} e^{-z/2}\left[1 + \sum_{i=1}^{\infty} \prod_{j=1}^{i} \left(\frac{z}{n + 2j}\right)\right],$$

$$G_2(n) = \begin{cases} \prod_{j=1}^{n/2} j & , n \text{ even} \\ \sqrt{\pi} \prod_{j=1}^{(n+1)/2} (j - \tfrac{1}{2}), & n \text{ odd} \end{cases}$$

Here the quantity $G_2(n)$ is, in fact, the gamma function calculated at $1 + n/2$. The χ^2 level, H, for the statistic B_r with D_r degrees of freedom, which corresponds to the probability that the statistic has a value B_r or larger, is

(C3) $H(B_r, D_r) = 1 - G_1(B_r, D_r)/G_2(D_r)$.

Formula (C3) is used to obtain the final four columns in Tables 4 and 8. The theoretical basis for (C3) is given by Melsa and Sage (1973).

Appendix D. Computer Methods

Numerous algorithms are available for performing a computer search to locate a function minimum. Some require the function's derivatives, and others, called direct search methods, do not. Although derivative-based algorithms are usually more efficient of machine time, direct search methods typically involve less human time because they do not require the user to calculate and program derivatives. In this application, where the function A can depend on the parameters in many different ways, it is particularly inconvenient to produce the appropriate derivatives for each case. Macdonald and Pitcher (1979) employ the Nelder–Mead algorithm, a remarkably simple direct search technique. The idea behind the method was originally formulated by Nelder and Mead (1965). Later, O'Neill (1971) translated it into a computer program. Various corrections have since been suggested by Chambers and Ertel (1974), O'Neill (1974), Benyon (1976), and Hill (1978). All minima reported here were located by this technique.

If a direct search method is used, this paper contains all formulas necessary for writing a computer program to generate results like those in Tables 3–8. For convenience in describing the calculations, let $z_1, \ldots, z_P$ be the P parameters used by the statistical model. Suppose that the first P_1 z's determine the μ's, the next P_2 z's determine the σ's, and the final P_3 z's determine the π's. (Thus $P_1 + P_2 + P_3 = P$.) For example, if the μ's are required to lie on a von Bertalanffy curve, then $P_1 = 3$ with $z_1 = l$, $z_2 = L$, $z_3 = k$. If the μ's are unrestricted, then $P_1 = M$ with $z_i = \mu_i$ for i from 1 to M. Similarly $P_2 = 2$ if the σ's are linear, and $P_2 = M$ if the σ's are unrestricted. In all examples here $P_3 = M - 1$, because the last percentage is determined by the previous ones.

In addition to the direct search algorithm, the user must supply program modules to do the following eight tasks:

1. Calculate the μ's from the first P_1 z's. This might involve (7).
2. Calculate the σ's from the next P_2 z's, as well as the μ's if required. This might involve one of (11)–(14).
3. Calculate the π's from the final P_3 z's.
4. Calculate the f's from the μ's, σ's, and π's. This involves (15)–(16), as well as (C2).
5. Calculate A from the f's and $\hat{f}$'s. This involves (22).
6. Calculate, and add to A, penalty functions which are large at forbidden parameter values, such as $\pi_i < 0$ or $\pi_i > 1$.
7. Given r, calculate B_r and D_r from the f's and $\hat{f}$'s. This involves (19)–(20), (24), and (25).
8. Calculate a χ^2 level H from B_r and D_r. This involves (C3).

These programs should be kept modular so that individual steps, especially 1 and 2 above, can readily be adapted to a particular model.

After these tasks have been programmed, they can be used by two master programs. In the first, the search algorithm calls steps 1–6 to compute A at the point $(z_1 \ldots, z_P)$. This allows the algorithm to explore P-dimensional space and locate a minimum point. The penalty functions (step 6) play an important role in forcing the algorithm to avoid unreasonable parameters. For example, $10^6\pi_i^2$ might be added to A if $\pi_i < 0$ and $10^6(1 - \pi_i)^2$ might be added if $\pi_i > 1$. Other parameters might also be restricted if the user finds it appropriate. The second master program, which is used after a minimum has been located, calls steps 1–4 and 7–8 to determine the relevant χ^2 levels for the minimum point.