Asian Fisheries Science **13**(2000): 1-12
Asian Fisheries Society, Manila, Philippines
https://doi.org/10.33997/j.afs.2000.13.1.001

Size Frequency Analysis by Averaged Shifted Histograms and Kernel Density Estimators

I.H. SALGADO-UGARTE[1], M. SHIMIZU, T. TANIUCHI and K. MATSUSHITA

[1]F.E.S. Zaragoza U.N.A.M. Biología
Guelatao 66, Ejército de Oriente
Iztapalapa 09230, A.P. 9-020, México D.F.
México

University of Tokyo
Faculty of Agriculture, Department of Fisheries
Yayoi 1-1-1, Bunkyo-ku
Tokyo 113, Japan

Abstract

Size frequency analysis in fisheries is commonly carried out through histograms and frequency polygons. However, these procedures present several drawbacks including dependency on the interval width and grid origin, discontinuity, and use of fixed width intervals. These problems prompted the authors to focus their interest in alternative, more efficient, computationally intensive methods. In this study we used kernel density estimators (KDE) computed by computationally efficient algorithms (averaged shifted histograms) to analyze published size data of coral trout (*Plectropomus leopardus*). The KDE's do not depend on the grid origin and are continuous estimators. We also discussed several methods in choosing the interval width (smoothing parameter or bandwidth). These nonparametric estimators provide smoother results, that allow characteristics such as skewness, outliers, and multimodality to be easily recognized. Using the variable bandwidth KDE in the latter case, the definition and separation of the modes were improved, and led to more precise and objective mixed components determination. The estimations for the individual components (mean, standard deviation and size from Bhattacharya's procedure) can be employed as initial values in any method for mixed distribution analysis or can be used directly to estimate the parameters of the von Bertalanffy growth function. Our experiences in this study suggest that KDE´s are valuable tools in length frequency analysis and related methods such as modal progression analysis.

Introduction

Traditionally, histograms and frequency polygons are employed to analyze size frequency data. Most of the time, in these graphical procedures, the *y*-axis represents the number (frequency) of observations falling in the intervals (bins); however fractions or percentage scales are employed, too. Another less common manner of representing the ordinate axis is by using a density scale defined as the frequency of the bin divided by the product of the total number of observations multiplied by the binwidth. Thus, histograms and frequency polygons are estimates of the density distribution of the data set.

In spite of their wide usage, these density estimators may be too crude for many purposes (Tarter and Kronmal 1976). Four problems are encountered when using histograms (Fox 1990):

1. Dependency on the origin. The investigator must choose the position of the origin of the bins (very often by using convenient “round” numbers). This subjectivity can result to misleading estimations because a change in the origin can change the number of modes in the density estimation (Silverman 1986, Fox 1990, Scott 1992). Figure 1 shows the well-known standard length (mm) data of coral trout (*Plectropomus leopardus*, Serranidae) reported by Goeden (1978). Each histogram from a) to e), uses the same binwidth (h = 38) but from a different origin (193.8, 201.4, 209, 216 and 224.2, respectively). There are bimodal, trimodal and tetramodal histograms. It would be arbitrary to choose any of these histograms to represent the length distribution. Unfortunately, this exercise of drawing several origin shifted histograms may lead the analyst to select (intentionally or not) the one best suited for his purposes.

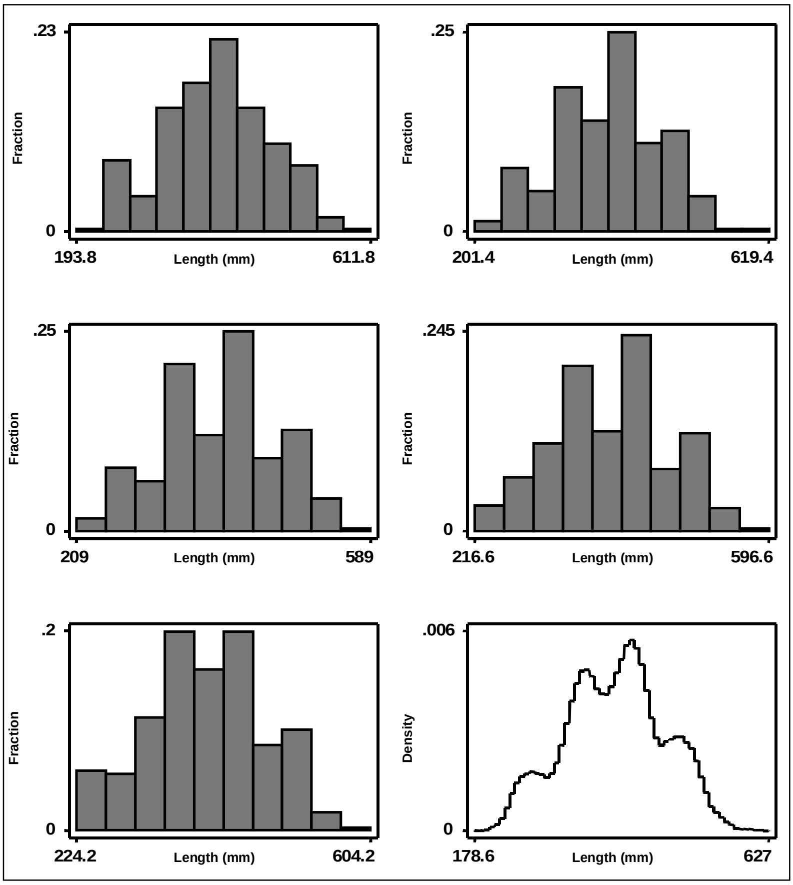

Fig. 1. Five histograms with different origins and the corresponding ASH for the coral trout length data.

2. Dependency on the width and number of bins. These parameters determine the smoothness of the frequency distribution (density estimation). Using few bins eliminates distribution details; a large number of bins produces a noisy estimation. Frequently, the number and width of the bins are determined arbitrarily despite their importance. As an example, consider Figures 2 and 3 for the coral trout data. The first histogram with five bins shows a smooth almost Gaussian distribution. The second, which has 50 bins displays a distribution with at least four modes.

3. Discontinuity. Histogram discontinuities are functions of the arbitrary bin locations and the discreetness of the data rather than of the population that is sampled. The local density is only computed at the midpoint of each bin and then the bars are drawn assuming a constant density throughout each bin (Chambers et al. 1983).

4. Fixed binwidth. If the bins are narrow enough to capture details where density is high, they may also be too narrow to avoid noise where density is low. This problem is often addressed by varying the binwidth, but the height of the bar is no longer proportional to its area, which may lead to misinterpretation.

Several methods attempting to overcome these problems have been proposed. The origin and discontinuity problems are attacked by calculating the local density at every data point. This is achieved, in essence, by constructing fixed-width bins around each data point and not only at the midpoints of intervals. From a formal point of view, the bin can be considered as a weight function that assigns positive weight to each observation within the interval and zero weight outside it. For example, in the traditional histogram, the weight is a constant value (uniform function) in the form of frequency, fraction, and percentage or density, assigned to each data point included in the interval. The individual values for the bin are added and the result can be represented geometrically through a rectangular shape (the classical bar) centered at the bin midpoint. In the alternative procedure to estimate the density, discontinuity is further addressed by considering gradually changing weight functions (like for example the Gaussian curve). In this manner, it is possible to employ a bell shaped figure centered at each data point and then add the individual curves to obtain the final result (for details see Chambers et al. 1983 or Härdle 1991). These notions led to the kernel density estimator (KDE), first proposed by Rosenblat (1956) which is defined as:

$$\hat{f}(x)=\frac{1}{hn}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right)$$

where $\hat{f}(x)$ is the density estimation of the variable x, n is the number of observations, h is the bandwidth or smoothing parameter, and K(Ÿ) is the smooth, symmetric kernel function integrating to unity. Table 1, adapted from Härdle (1991), lists some kernel functions.

These KDE´s employ fixed bandwidths. This feature makes the estimates sensible to noise in the tails or any other low count interval of the distribution. To attack this problem, there are procedures which reduce the

Table 1. Some common kernel functions.

Kernel	K(z)
Uniform	½ I(|z| ≤ 1)
Triangle (ASH)	(1 - |z|) I(|z| ≤ 1)
Epanechnikov	¾(1 - z^2) I(|z| ≤ 1)
Quartic	(15/16)(1 - z^2)^2 I(|z| ≤ 1)
Triweight	(35/32)(1 - z^2)^3 I(|z| ≤ 1)
Cosinus	(π/4)cos(π/2)z) I(|z| ≤ 1)
Gaussian	(1/)exp((-1/2)z^2

In all cases $z = (x - X_i) / h$

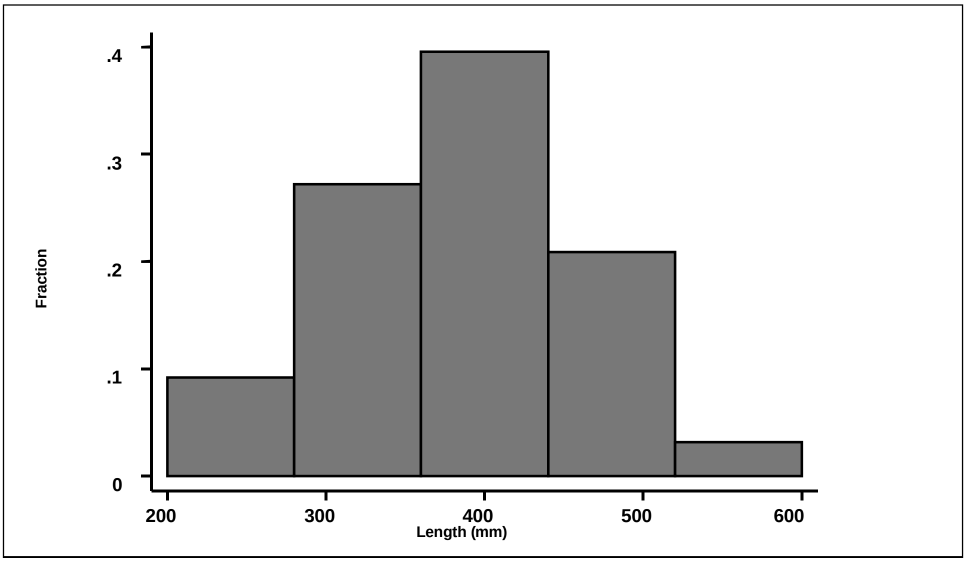

Fig. 2. Histogram with five bins for the coral trout length data.

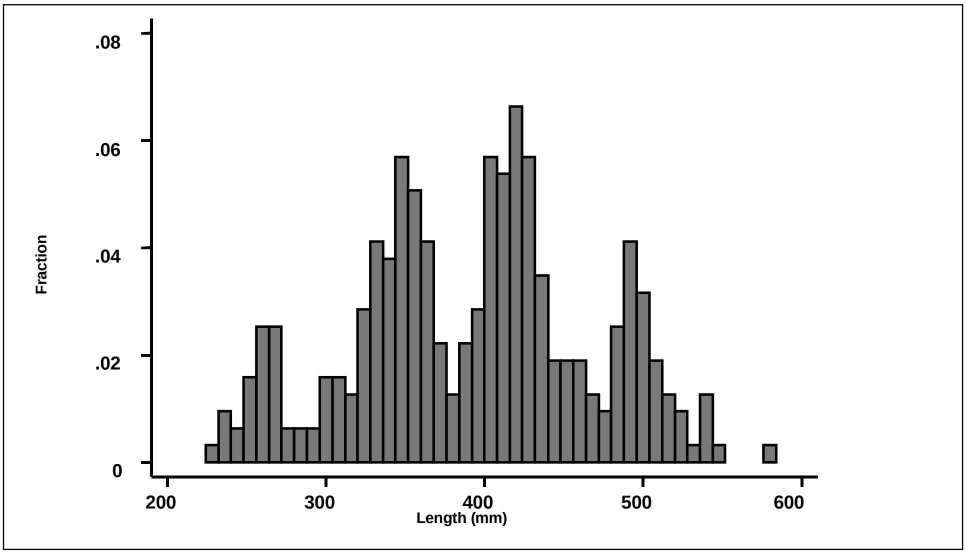

Fig. 3. Histogram with 50 bins for the coral trout length data.

bandwidth in regions of high data concentration and increase it where concentration is low. These varying KDE procedures (Jones 1990) retain details where observations concentrate and eliminate noise fluctuations where data are sparse.

The problem of choosing the width of the interval (bandwidth, *h*) remains. One approach to choose the bandwidth suggested by Tarter and Kronmal (1976) is to vary *h* until a "pleasing" smooth figure results. This procedure relies on the subjective assessment of the researcher, but may be adequate for exploratory purposes (Silverman 1986) since density features "appear" and "disappear" as the bandwidth changes (Silverman 1981a).

Statistical theory provides some guidance in the selection of an optimal bandwidth. Unfortunately it is generally not possible to optimize the bandwidth without previous knowledge of the shape of the true density distribution. Following Tukey (1977), Scott (1979) and Silverman (1978, 1986), the Gaussian distribution can be employed as a reference standard in choosing *h*. Applying a Gaussian kernel and minimizing the mean integrated squared error (MISE), the following scale estimate can be calculated:

$$s = \min \left[\left(\frac{\sum (x_i - \bar{x})^2}{n-1} \right)^{1/2}, \frac{H\text{spread}}{1.349} \right]$$

where H_{spread} is the "Hinge spread" or upper Fourth (F_U) minus lower Fourth (F_L), a resistant dispersion measure approximately equivalent to the interquartile range (Tukey, 1977). Then *h* can be chosen as:

$$h = \frac{0.9s}{n^{1/5}}$$

where *s* is the smaller of the two estimates of the Gaussian distribution spread parameter: σ, the standard deviation or the robust *F-pseudosigma*, as H_{spread} /1.349 is named (Hoaglin 1983; Fox 1990). This adjustment provides resistance to heavy tails and will work well for a wide range of densities but, it tends to oversmooth highly skewed and multimodal distributions (Silverman 1986). If this is the case, the "optimal" bandwidth can be considered as a starting point for subsequent fine tuning.

One drawback presented by the KDE´s is the large number of calculations required to compute them. Scott (1985) suggested an alternative procedure to overcome this problem: To eliminate the influence of the chosen origin, he proposed to get the average of several histograms with different origins instead of choosing one among them. This is the averaged shifted histogram (ASH). Subsequently, Härdle and Scott (1988), developed the more general framework called WARP (weighted averaging of rounded points).

The calculation of an ASH-WARP estimate takes three steps (1) binning the data; (2) calculating the weights, and (3) weighing the bins. The last graph of Figure 1 displays the result of averaging the five shifted histograms included before; the trimodality of the data is evident as a result of a significantly improved signal-to-noise ratio obtained by averaging the origin. WARPing can be used to approximate a particular kernel density estimator by selecting the appropriate weight function. The WARP approaches the kernel function as the number of averaged histograms increases (Härdle 1991).

Materials and Methods

In order to illustrate the use of the KDE´s presented above and to analyze size frequency distributions, we used the coral trout (*Plectropomus leopardus*, Serranidae) length data set (*n* = 316), adapted from Goeden (1978). A more elaborate nonparametric approach for multimodality assessment using different fish length data is presented in Salgado-Ugarte (1995).

To calculate the kernel and ASH-WARP density estimations, the programs presented in Salgado-Ugarte et al. (1993, 1995a, 1995b) were used. These programs permit estimating the density distribution by using different kernel functions and procedures, employing discretized and ASH-WARP implementations. Besides, some of them permit to count and estimate the modes present in the density estimation. The variable bandwidth KDE was used in conjunction with a computerized version (Salgado-Ugarte et al. 1994) of the Bhattacharya´s method to exemplify the utility of smooth density estimations in mixed distribution's component identification and characterization. We used the optimal Gaussian bandwidth as the starting point and then decreased it to finally choose *h* = 5.

Results

The histogram employing the optimal Gaussian binwidth (Scott 1979) *h* = 38, and origin at 133 is presented in Figure 4. The estimation suggests the multimodality of the data but does not provide enough detail. The estimation resulting from the average of five shifted histograms (Fig. 5) clearly reveals the existence of at least four modes even when the same binwidth (*h* = 38) is employed. The origin of the grid used for calculation is no longer important. The Gaussian kernel density estimation using the bandwidth proposed by Silverman (1986) *h* = 20 provides a smooth estimate showing four somewhat oversmoothed modes (Fig. 6). Using this optimal value as an upper limit, we tried smaller values until we arrived at the final choice. The variable Gaussian kernel density estimation employing a bandwidth (geometric mean) of *h* = 5 is included in Figure 8. The Gaussian components estimated by the Bhattacharya´s method are also included. Bhattacharya's plot for the KDE is shown in Figure 7 with the Gaussian components clearly suggested by the negatively sloped segments in contrast with the noisy graph obtained by using the original grouping scheme. Table 2 presents the estimated parameters for the identified Gaussian components. The mean values for components 1 to 6 (excluding component 2a) were used to estimate the von Bertalanffy growth function through nonlinear regression:

Table 2. Estimated parameters for Gaussian components.

Component	Mean	Standard deviation	Component size
1	259.9	9.9	26
2a	303.4	12.2	16
2	350.5	16.0	87
3	415.6	15.5	103
4	458.6	11.2	17
5	494.3	10.8	43
6	530.7	14.7	12
		Total	304

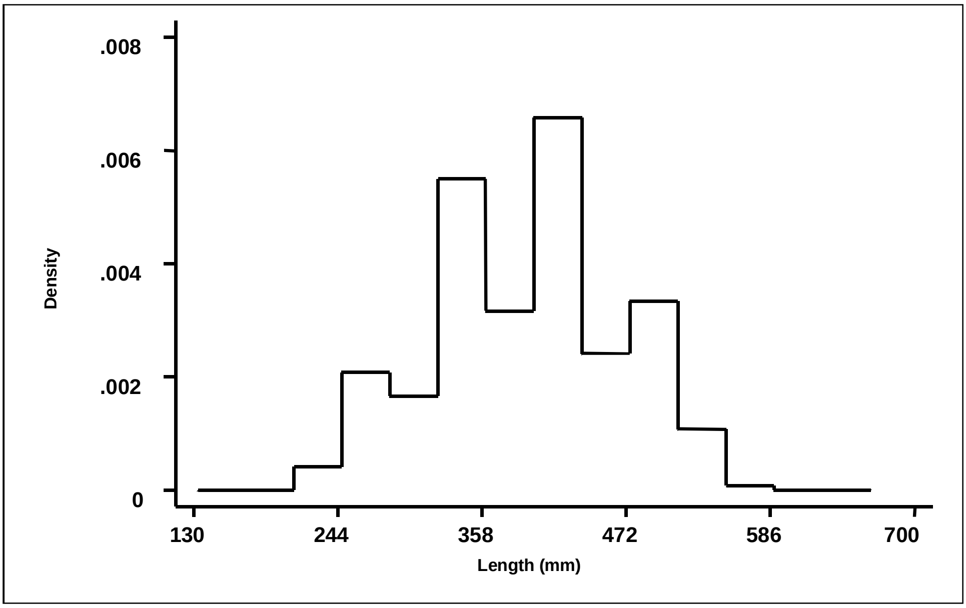

Fig. 4. Histogram with the optimal Gaussian binwidth (h = 38) and origin at 133 for the coral trout length data.

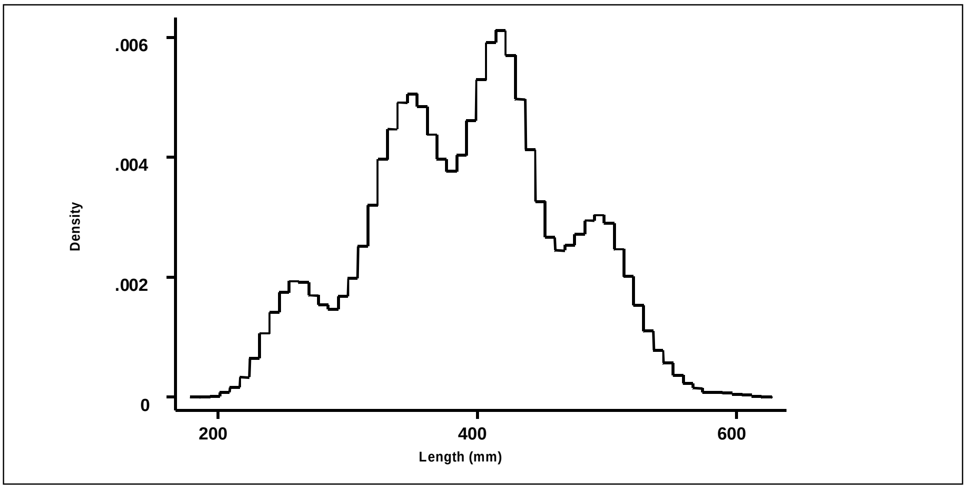

Fig. 5. Density estimation with five averaged shifted histograms using the optimal Gaussian binwidth (h = 38) for the coral trout length data.

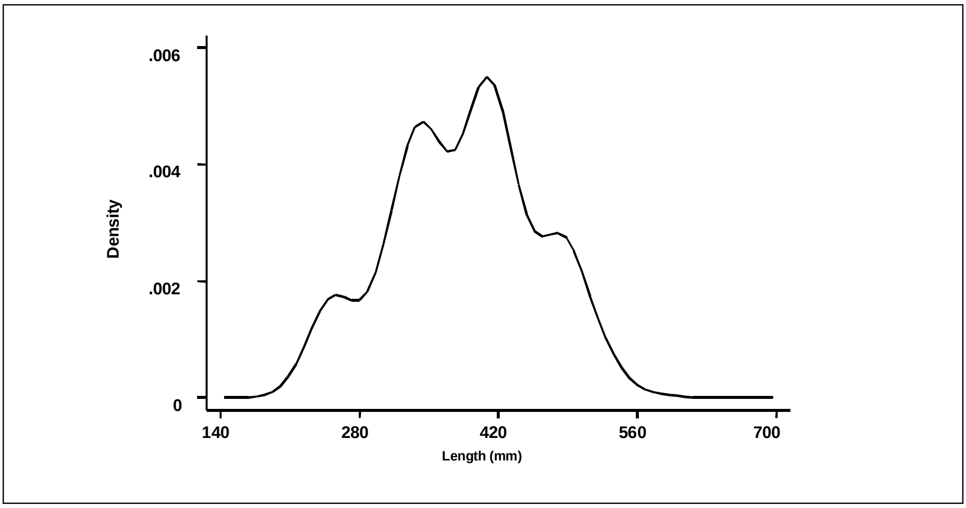

Fig. 6. Gaussian kernel density estimation using the optimal Gaussian bindwidth (*h* = 20) for the coral trout length data.

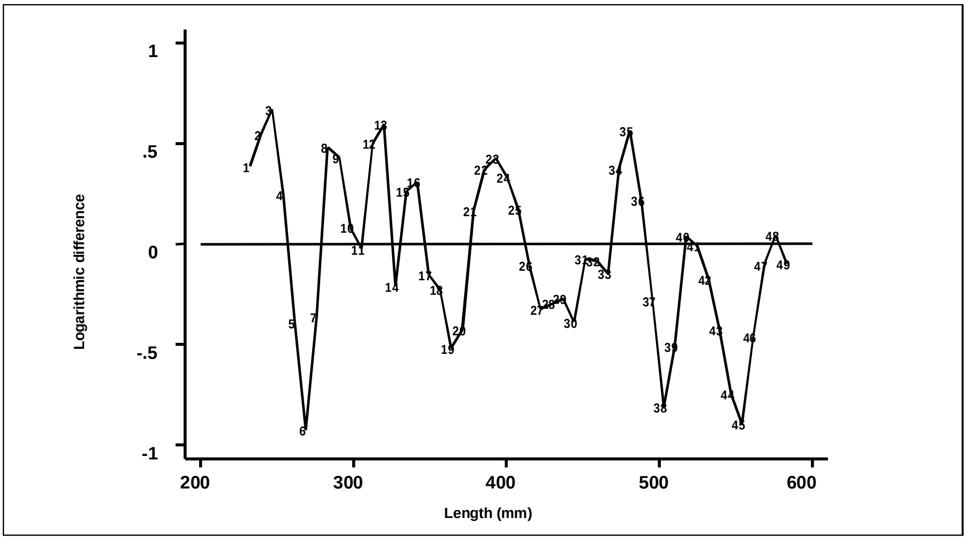

Fig. 7. Bhattacharya’s plot for the variable Gaussian kernel frequency for the coral trout length data.

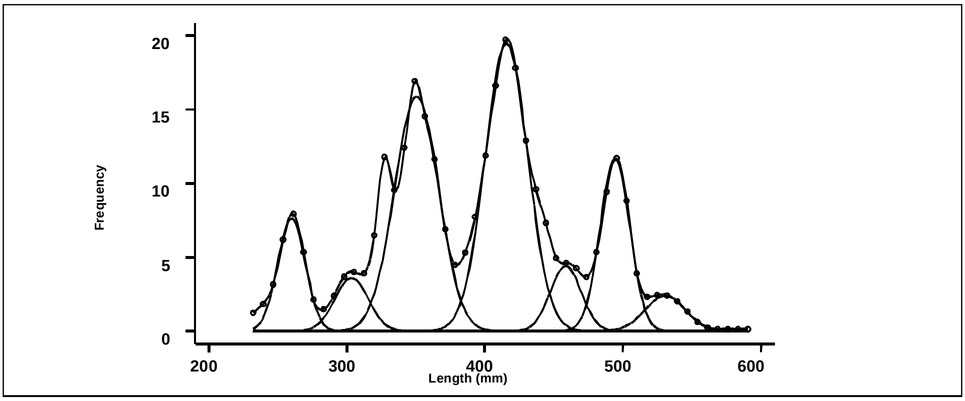

Fig. 8. Variable kernel density estimation (geometric mean *h* = 5) and estimated components for the coral trout length data.

$$L_t = 611.2513(1 - e^{-0.2836(t + 0.9687)})$$

Discussion and Conclusions

The use of size frequency to investigate the growth of animals dates back to the paper of Petersen (1892) in which he presented length measurements of fish and found that with temperate species breeding once a year it is relatively easy to define a cohort by a year-class (a mode in the histogram showing the frequency distribution). This cohort can be followed during the first part of its life by tracing the corresponding modes in the histograms from the samples, but when they approach their maximum size this is no longer possible, because by then, fish of different ages have reached approximately the same size (Sparre & Venema 1992).

The statistical procedure to show the distribution of the lengths (histogram) is a data smoother with the interval width being the smoother parameter (Härdle 1991). The number of modes depends on the interval width. In this way, in fisheries research a fundamental question must be "What is the best interval width?" In this respect, we can recall several suggestions from literature. Interval sizes of 0.5 cm for small species (< 30 cm) and 1.0 or 2.0 cm for larger species are widely used. Wolff (1989) proposed an empirically derived formula based on maximum observed size and estimated number of age classes in the sample for selection of the optimum interval size. It has been noted that in fisheries a compromise often has to be made between measuring a small number of fish slowly and accurately, and grouping larger numbers of measurements into wider intervals (Caddy 1986). This author suggests that the interval width should be small enough so that successive peaks were separated by five or six size class intervals. In his study Erzini (1990) argues that the optimum interval size for grouping length data is a function of sample size and biological characteristics, such as length at age variability, recruitment pattern, growth rate and maximum size, which affect how clearly defined the modes are in the distribution. Erzini's paper supports Caddy's suggestion quoted above finding that empirically based methods for determining the interval width may only be useful to provide rough estimates. Surprisingly, the effect of the origin of the intervals in the resulting histogram seems not to have received attention in fisheries research literature.

The ASH gives a more detailed version of the density distribution (Fig. 5). This estimator in its interpolated version converges to a KDE as the number of shifted histograms increases using the appropriate weight function. In practice it is enough to average from 10 to 15 shifted histograms to arrive at the same results obtained using a KDE (Härdle 1991). This property is used to save a great number of individual calculations to estimate kernel density estimators.

The KDE's solve some of the problems of the histogram and are a suitable procedure for analyzing length frequency data. All the kernel functions listed in

Table 1 have an efficiency very close to that of the maximally efficient Epanechnikov (1969) kernel. As a consequence, a kernel function can be chosen on the basis of its computational effort (Silverman 1986).

In working with KDE's the effect of the interval origin is solved, but the problem of choosing the smoothing parameter (bandwidth) still remains. However, several statistical guidelines are available. The rule for bandwidth selection introduced above has been developed for a single underlying distribution. In mixed distributions, however, there are several components (Gaussian or of other kind), each with different parameters (such as mean and standard deviations in the Gaussian case). The ideal number and width of intervals may be different for each component. Dominant groups - components with many individuals - permit the use of a large number of small intervals; more sparsely populated components can support only a few, relatively wide intervals. The classical histogram uses a fixed bin width, hence it may do a poor job in portraying both the dominant and lesser components. Under these circumstances, it seems to be particularly appropriate to use the variable bandwidth KDE which adjusts the interval to provide details (decreasing h at high density regions) and to eliminate noise (increasing h where density is low). However, as with histograms it is important to measure lengths with the higher possible accuracy to be able to use a wide range of bandwidths.

We will stress here the use of the variable Gaussian kernel density estimation (Figure 8) as it does not depend on the placement of the origin, it adjusts the bandwidth according to the number of observations and in this way reveals more details and greater separation of the modes in comparison with fixed bandwidth estimations (including histograms and KDE's). The corresponding Bhattacharya´s plot (Fig. 7) reveals distinct, negatively sloped linear segments not distinguishable as easily working with minimally grouped histograms.

As a later step, the resulting parameter estimates for each component can be used as initial values in a subsequent maximum-likelihood estimation (Akamine 1985; Macdonald and Green 1988; Fournier et al. 1990). It is worth to note that the density estimation could recover the overlapped component around 460 mm of standard length, assumed by Sparre and Venema (1992) who applied a trial and error procedure to estimate growth from length frequency analysis with a histogram of the same data set. The estimated von Bertalanffy growth function parameters (Table 3) through nonlinear regression (adjusted $r^2 = 0.9999$, $P < 0.05$) with the variable bandwidth KDE do not differ significantly from the values estimated by Sparre and Venema (1992) who arrived at $L_\infty = 595$, $K = 0.34$ and $t_0 = -0.65$: These parameter values are included in the corresponding 95% confidence intervals from nonlinear regression of the KDE (Table 3).

Finally, the kernel density estimates provide several ways to test and evaluate multimodality (for details see Silverman 1981b, 1983, 1986). An example using the smoothed bootstrap method of Silverman (1981b) is presented in Salgado-Ugarte et al. (1997). Other approaches have been suggested by Cox (1966), Good and Gaskins (1980) and Wong (1985). The programs for kernel calculation and

Table 3. Nonlinear regression for von Bertalanffy growth function parameter estimation.

Parameter	Coefficient	Standard error	t value	P > | t |	95 % Confidence interval	
L_∞	611.2513	20.3498	30.037	0.000	546.489	676.0136
K	0.2836	0.0308	9.210	0.003	0.1856	0.3816
t_0	-0.9687	0.1598	6.062	0.009	-1.4771	-0.4601

Note: Standard errors, P values, CI's and correlations are asymptotic approximations.

Gaussian characterization can be obtained from the first author.

Acknowledgments

The first author is grateful to the Ministry of Education, Science and Culture of Japan and to the Universidad Nacional Autónoma de México (F.E.S. Zaragoza, D.G.A.P.A. and D.G.A.P.A.-P.A.P.I.I.T. project IN217596) for their support. We are grateful to the three anonymous referees for their valuable suggestions and corrections.

References

Akamine, T. 1985. Consideration of the BASIC programs to analyze polymodal frequency distribution into normal distributions. (in Japanese with English abstract), Bulletin of the Japan Sea Regional Fisheries Research Laboratory 35: 129-160.

Caddy, J.F., 1986. Size frequency analysis in stock assessment -some perspectives, approaches and problems. Proceedings of the 37th Annual gulf and Caribbean Fisheries Institute: 212-238.

Chambers, J.M., W.S. Cleveland, B. Kleiner and P.A. Tukey 1983. Graphical methods for data analysis. Belmont, CA: Wadsworth.

Cox, D.R. 1966. Notes on the analysis of mixed frequency distributions. The British Journal of Mathematical and Statistical Psychology, 19:39-47.

Epanechnikov, V.A. 1969. Nonparametric estimation of a multidimensional probability density. Theory of Probability and Its Applications, 14: 153-158.

Erzini, K. 1990. Sample size and grouping of data for length-frequency analysis. Fisheries Research 9: 355-366.

Fournier, D.A., J.R. Sibert, J. Majkowski and J. Hampton, 1990. MULTIFAN, a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for Southern bluefin tuna (Thunnus maccoyii). Canadian Journal of Fisheries and Aquatic Sciences 47: 301-317.

Fox, J. 1990. Describing univariate distributions. *In* J. Fox and J.S. Long (eds.) Modern Methods of Data Analysis, pp. 58-125. Newbury Park, CA: Sage publications.

Goeden, G.B. 1978. A monograph of the coral trout, Plectropomus leopardus (Lacépède). Research Bulletin of the Fisheries Service Queensland, 1: 42 p.

Good, I.J. and R.A. Gaskins. 1980. Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteroite data. Jouranl of the American Statistical Association, 75:89-103.

Härdle, W. 1991. Smoothing Techniques. With Implementations in S. Springer-Verlag. New York.

Härdle, W. and D.W. Scott. 1988. Smoothing in low and high dimensions by weighted averaging using rounded points. Technical report 88-16, Rice University.

Hoaglin, D.C. 1983. Letter values: A set of selected order statistics. *In* Hoaglin, D.C., F. Mosteller and J.W. Tukey (eds.) Understanding robust and exploratory data analysis, pp. 33-57. New York, John Wiley & Sons.

Jones, M.C. 1990. Variable kernel density estimates and variable kernel density estimates. Australian Journal of Statistics, 32(3)

Macdonald P.D.M. and P.E.J. Green 1988. User's guide to program MIX: an interactive program for fitting mixtures of distributions. Ichthus Data Systems, Hamilton, Ontario, Canada.

Petersen, C.G.J., 1892. Fiskenes biologiske forhold i Holbaek Fjord, 1890-91. Beret. Danske Biol. St., 1890(1)1: 121-183 (in Danish).

Rosenblatt, M. 1956. Remarks on some nonparametric estimates of a density function. Annals of Mathematical Statistics 27: 832-837.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1993. Exploring the shape of univariate data using kernel density estimators. Stata Technical Bulletin 16: 8-19.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi. 1994. Semi-graphical determination of Gaussian components in mixed distributions. Stata Technical Bulletin 18: 15-27.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi, 1995a. ASH, WARPing, and kernel density estimation for univariate data. Stata Technical Bulletin 26: 2-10.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi, 1995b. Practical rules for bandwidth selection in univariate density estimation. Stata Technical Bulletin 27: 5-19.

Salgado-Ugarte, I.H., M. Shimizu, and T. Taniuchi, 1997. Nonparametric assessment of multimodality for univariate data. Stata Technical Bulletin 32: 27-35.

Scott, D.W. 1979. On optimal and data-based histograms. Biometrika, 66: 605-610.

Scott, D.W. 1985. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. Annals of Statistics, 13: 1024-1040.

Scott, D.W. 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. New York: John Wiley & Sons.

Silverman, B.W. 1978. Choosing the window width when estimating a density. Biometrika, 65: 1-11.

Silverman, B.W. 1981a. Density estimation for univariate and bivariate data. In Interpreting Multivariate Data, ed. V. Barnett, 37-53, Chichester, John Wiley and Sons.

Silverman, B.W. 1981b. Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society, B, 43: 97-99.

Silverman, B.W. 1983. Some properties of a test for multimodality based on kernel density estimates. *In* J.F.C. Kingman and G.E.H. Reuter (ed.) Probability, Statistics and Analysis. Cambridge, pp. 248-259. Cambridge University Press.

Silverman, B.W. 1986. Density estimation for statistics and data analysis. London: Chapman & Hall.

Sparre, P. and S.C. Venema, 1992. Introduction to tropical fish stock assessment. Part I. Manual. FAO Fisheries Technical Paper No. 306.1. Rev. 1. Rome, FAO, 376 p.

Tarter, M.E. and R.A. Kronmal 1976. An introduction to the implementation and theory of nonparametric density estimation. The American Statistician, 30: 105-112.

Tukey, J.W. 1977. Exploratory data analysis. Reading, MA. Addison-Wesley.

Wolff, M., 1989. A proposed method for standardization of the selection of class intervals for length frequency analysis. Fishbyte, 7: 5.

Wong, M.A., 1985. A bootstrap testing procedure for investigating the number of subpopulations. Journal of Statistical Computation and Simulation. 22: 99-112.

Manuscript received 12 July 1999; Accepted 02 December 1999.