Separation of a Mixture Distribution Into Its Gaussian Components*

MINA L. SORIANO**
International Center for Living Aquatic Resources Management (ICLARM)
MC P.O. Box 1501
Makati, Metro Manila
Philippines

Abstract

An application of an algorithm presented by J. Gregor in 1969 for decomposition of a mixture distribution into Gaussian components is presented here in relation to discrete mixture distributions of constant class interval. This graphical separation called Modified Gregor method (MG) uses the local modes to estimate the mean of components, the area around the mean for estimating the variance and subgroup population components. The Kolmogorov-Smirnov statistic is used for goodness of fit test of the composite distribution.

This deterministic method has proven to be comparable to and as efficient as the commonly used Bhattacharya method. For comparison, the MG and Bhattacharya methods are used in three examples related to fish population studies.

Introduction

A mixture distribution is a single composite function of two or more overlapping statistical distributions, referred to as the component distributions of the mixture. It is the weighted average of probability distributions with the sum of the weights equal to one. The generic formula for finite distribution, which is the primary consideration in this paper, is:

$$f(x) = \sum_i p_i f_i(x)$$

where $-\infty < x < +\infty$, $0 < p_i < 1$, $\Sigma p_i = 1$,
p_i is the mixing distribution and
$f_i(x)$ is the density function of component i.

Many complex mixtures can be found in biology, physics, medicine and other fields. Some of these are:

a. K. Pearson's study on separating trypanasoma (parasitic protozoan) frequency data, consisting of two overlapping normal distributions (Pearson 1914).

b. Gregor's algorithm (1969) for separation of mixtures applied to the DNA content in the nuclei of liver cell of rats.

c. In fishery biology, the mean sizes of different age groups of fish is very often estimated by the length-frequency distributions with two or more overlapping normal distributions (Harding 1949; Cassie 1954; Bhattacharya 1967; MacDonald and Pitcher 1979).

Identifiability of component distribution and estimation of parameters and mixing distribution are the two common problems regarding this type of distribution. This paper assumes that the mixture would have its components follow a normal distribution, and our focus of interest will then be on the estimation of parameters and mixing distribution.

Modified Gregor's Method (MG)

Gregor (1969) presented an algorithm, written in ALGOL-60, for the decomposition of mixture distributions into their normal components. This algorithm consists of three steps:

a) locating the means by using Fourier transform of the normal density function and decreasing standard deviation;
b) determining the standard deviation and frequency components using continued fraction approximation of the error function; and
c) testing the results using the Kolmogorov-Smirnov test statistic.

A modified approach of the method for grouped data uses steps b) and c) above. The means and variances are estimated using local modes and areas

*ICLARM Contribution No. 456.
**Present address: #83 San Rafael St., Mandaluyong, Metro Manila, Philippines.

around these modes which are apparent in the histogram. These modes may be said to be indicative of homogeneous subpopulations giving clues in the estimation of component means and standard deviations.

The expected frequency distribution or composite distribution for grouped data may then be defined as:

$$h(x) = \sum_{k=1}^{M} c * N_k * h_k(x; \mu_k, \sigma_k)$$

where c = class interval
N_k = the estimated number of cases belonging to the k^{th} component
$h_k(x; \mu_k, \sigma_k)$ = Gaussian density function of the k^{th} component
k = 1,2...M
and M = total number of single distribution in the mixture.

Note that the proportion of each component in the mixture may be estimated by:

$$P_k = N_k / \sum_{k=1}^{M} N_k$$

Computation of the Estimates of the Gaussian Density Parameters and the Mixing Distribution

a) The mean estimate for the k^{th} component (μ_k) is the weighted average of two or four adjacent classes on both sides of the local mode and the modal class.

b) In the same manner as the mean, the σ_k is computed using the area $2*c$ and $4*c$ on both sides of the mode (Gregor 1969 and references therein). Simpson's approximation for the area is used.

$$I_1 = \frac{c}{3}\,[g(x_{i-1})+4g(x_i)+g(x_{i+1})]$$

$$I_2 = \frac{c}{3}\,[g(x_{i-2})+g(_{i+2})+4\,\{g(x_{i+1})]$$

$$+g(x_{i-1})\} +2*g(x_i)]$$

where $g(x_i)$ = frequency distribution at class x_i.

Supposing $h(x) \sum_{k=1}^{M} = A_k\, h_k(x; \mu_k, \sigma_k)$

where $A_k = c*N_k$

is an adequate model of h(x), over the interval (μ_k-2c; μ_k+2c)

I_j can be written approximately as

$$I_j = \frac{A_k}{\sigma_k \sqrt{2}} \sum_{\mu_k - jc}^{\mu_k + jc} \exp\left[\frac{-(x-\mu_k)^2}{2\sigma_k^2}\, dx\right]$$

Setting $t = (x-\mu_k)/(\sigma_k \sqrt{2})$, it follows

$$I_j = \frac{2A_k}{\sqrt{\pi}} \int_0^{j*c} \sigma K \sqrt{2}\, \exp(-t^2)\, dt$$

$$= \frac{2A_k}{\sqrt{\pi}}\, \mathrm{erf}\left[\frac{j*c}{\sigma K \sqrt{2}}\right]$$

where erf denotes the error function and can be expressed as an approximation to a mathematical expansion of $\int \exp(-t^2)\, dt$. This approximation is expressed as:

$$\mathrm{erf}(x) = (30*x-x^3)/(30+9*x^2)$$

thus,

$$I_j = \frac{2A_k}{\sqrt{\pi}} * \frac{60*\sigma_k^2 - (j*c)^2}{60*\sigma_k^2 + (3\cdot j\cdot c)^2}$$

$$\left[\frac{j*c}{\sigma_k \sqrt{2}}\right]$$

$$\sigma_k^2 = -(7*I_2-2*I_1)/(12*(I_2-2*I_1)*c^2$$

$$\sigma_k^2 = \frac{7*\{g(x_{i-2})+g(x_{i+2})\}}{12*\{g(x_{i-2})+g(x_{i+2})\}} + \frac{26*\{g(x_{i-1})+g(x_{i+1})\}+6*g(x_i)}{24*\{g(x_{i-1})+g(x_{i+1})\}-72*g(x_i)} * c^2$$

c) The estimated number of cases in the subpopulation is obtained from the two areas described above, where A'_k and A''_k are as follows:

$$A_k' = \frac{\{g(x_{I+1})+4*g(x_I)+g(x_{I-1})\}*\{60fc\sigma^2_k+9*c^2\}}{6*\{60\sigma^2_k - c^2\}} * \hat{\sigma}_k\sqrt{2\pi}$$

$$A_k'' = \frac{\{g(x_{I+2})+g(x_{I-2})+4[g(x_{I-1})+g(x_{I+1})]+2g(x_I)\}\{60\sigma^2_k+36c\}}{12*\{60\sigma_k - 4c^2\}} * \hat{\sigma}_k\sqrt{2\pi}$$

For the interval $(\hat{\mu}_k-\hat{\sigma}_k, \hat{\mu}_k+\hat{\sigma}_k)$,

$$d_j = \max_x\ g(x) - A_k^{(j)}h_k(x; \hat{\mu}_k, \hat{\sigma}_k);\ j = 1 \text{ or } 2 \text{ prime(s)}$$

is computed and the A(j) value which is the smaller of the two maxima is used to compute the total frequency of the k^{th} component.

$$N_k = A(j)/c$$

For each local mode in the g(x) frequency distribution the three Gaussian parameters (μ, σ and N) are estimated and the expected frequency for that component is subtracted from g(x). Say, after the parameters μ_1, σ_1 and N_i have been estimated for component 1, the $h_1(x)$ is subtracted from g(x). This process goes on as follows:

$$\varphi_1(x) = g(x) - c*\hat{N}_1*h_1(x; \hat{\mu}_1, \hat{\sigma}_1)$$
$$\varphi_2(x) = g(x) - c*\hat{N}_1*h_1(x; \hat{\mu}_1, \hat{\sigma}_1) - c*\hat{N}_2*h_2(x; \hat{\mu}_2, \hat{\sigma}_2)$$
$$\vdots$$
$$\varphi_m(x) = g(x) - c*\hat{N}_k*h_k(x; \hat{\mu}_k, \hat{\sigma}_k)$$

Since, $h(x) = \sum c*\hat{N}_k*h_k(x; \hat{\mu}_k, \hat{\sigma}_k)$, subtraction will only terminate when the Kolmogorov-Smirnov (Ks) test is no longer significant at a specified alpha level (e.g., 5% level).

For clarity, let us define the local modes apparent in the g(x) distribution from which components are estimated and call them "primary components". After the primary components are subtracted from g(x), new modes will emerge; these will be referred to as "secondary components". "Tertiary components" will be those modes that appear after the subtraction of the secondary components and so on until a good composite distribution is found.

The above steps can easily be automated*. This whole new deterministic method of separating components eliminates subjective bias, and works as efficiently as the other existing methods.

*Editor's note: they have been incorporated into the Compleat ELEFAN package (Gayanilo et al. 1988) as part of the routine to decompose recruitment patterns into their component "pulses".

Bhattacharya Method

The Bhattacharya method, a commonly used graphical method for separation of mixture distribution to which the MG method will be compared is described briefly below:

Given the frequency distribution f(x), with x as the class midpoint, the Bhattacharya method transforms the frequency data to:

$$b_i(x) = \log_e \frac{f(x_{i+1})}{f(x_i)}$$

and plots these points along the midpoints (x's). One determines from the plot points to which a straight regression line with negative slope should be fitted from which the mean and standard deviation estimates of one component are derived. Computation of these estimates are as follows:

$\hat{\mu}_i=(-a/b)$ and $\hat{\sigma}_1=(-1/b)$,
where a = intercept of the regression line
b = slope.

Each normal component is computed sequentially and the number of "population" per component is the sum of the frequencies up to and including one class length after the last selected point in the regression. Fig. 1 illustrates the Bhattacharya method using the Compleat ELEFAN computer software (Gayanilo et al.. 1988).

Example 1 Recruitment pattern with bimodal frequency distribution

Separation of component distribution has wide application in fish population dynamics. To illustrate the two methods described above let us begin with a recruitment pattern of *Hirundichthys affinis* in the Caribbean Sea with apparently bimodal frequency. Recruitment pattern is much used in fish population studies for it defines an observable event such as movement or migration (Gulland 1983). All data used in the three examples were kindly contributed by P. Dalzell, ICLARM.

Fig. 2 shows the separation of the recruitment pattern using the Bhattacharya and MG methods, respectively, and the two composite distributions identified by both methods are summarized in Table 1. The estimated means and variances of the two components show no significant difference at 5% level and the Kolmogorov-Smirnov test is not significant for the observed and composite distributions in both methods.

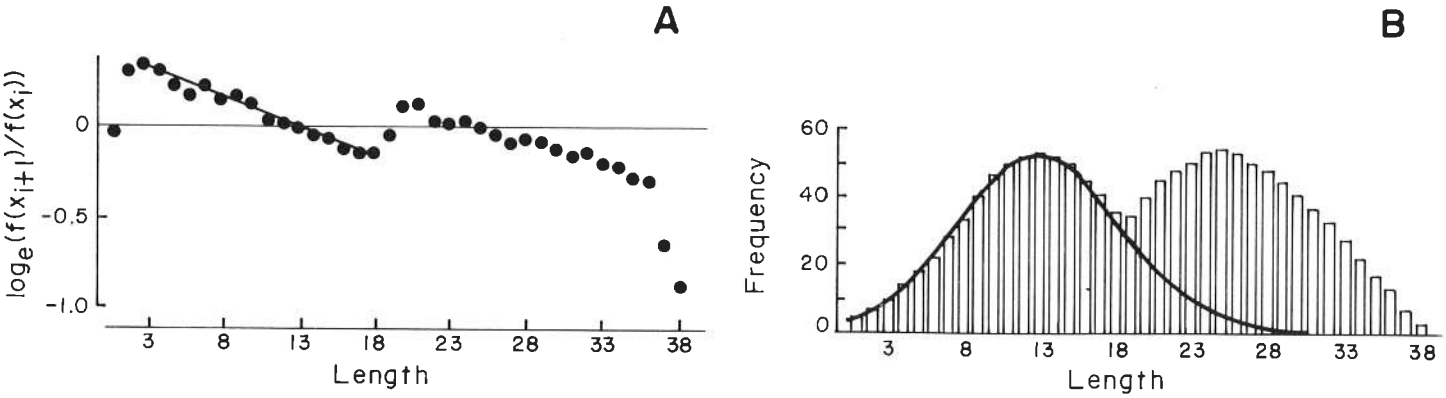

Fig. 1. Separation of mixture distribution of hypothetical data using the Bhattacharya method.

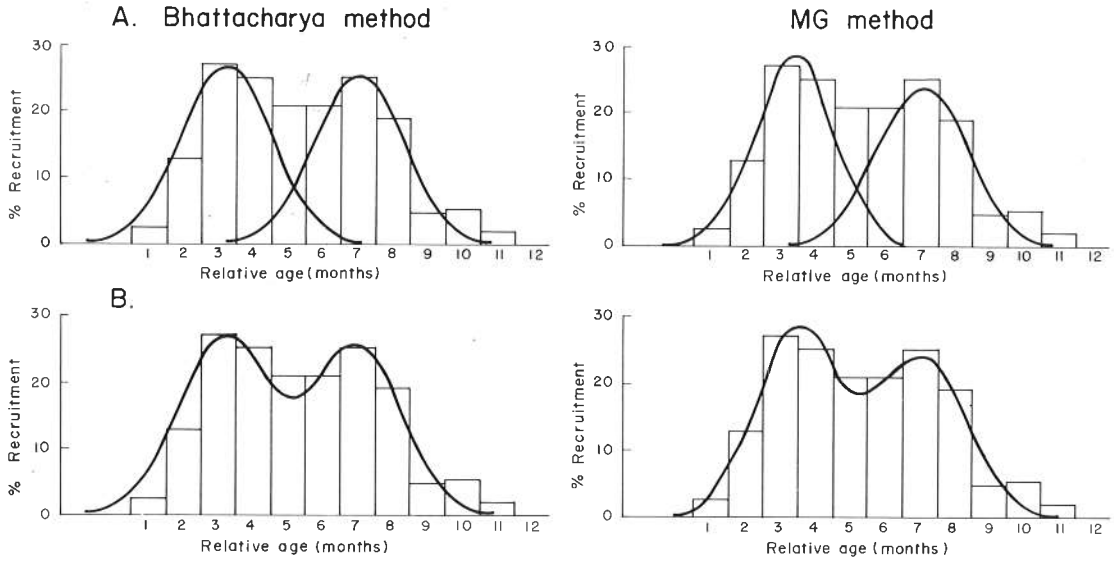

Fig. 2. Gaussian components (A) and composite distribution (B) of the recruitment pattern of *Hirundichthys affinis* identified using the Bhattacharya and MG methods.

Table 1. Estimated means, variances and subpopulations of two normal components identified using the Bhattacharya and MG methods for the recruitment pattern of *Hirundichthys affinis* with test for means and variances inequality and goodness of fit test for the composite distributions.

Comp. #	Bhattacharya method			MG method			Test of hypotheses	
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{N}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{N}$	H_o: $\mu_1 = \mu_2$	$\sigma_1 = \sigma_2$
1	4.41	1.110	48.51	3.19	1.319	53.52	ns	ns
2	7.99	1.344	48.33	7.01	1.280	49.17	ns	ns
	Ks[1] = .033 ns			Ks = .055 ns				

ns - not significant at 5% level
[1]Ks tabular value is 0.089

Example 2 Recruitment pattern with unimodal frequency distribution

Another recruitment pattern is that of *Oxyporhamphus convexus* in the Camotes Sea, Philippines, where the frequency distribution is unimodal but with an overlap of two normal distributions. Again the two methods were used to separate this mixture distribution (Fig. 3). Similar results as in example 1 are shown in Table 2. The estimated means and variances of the two components show no significant difference at 5% level and the Kolmogorov-Smirnov test is not significant for the observed and composite distributions in both methods.

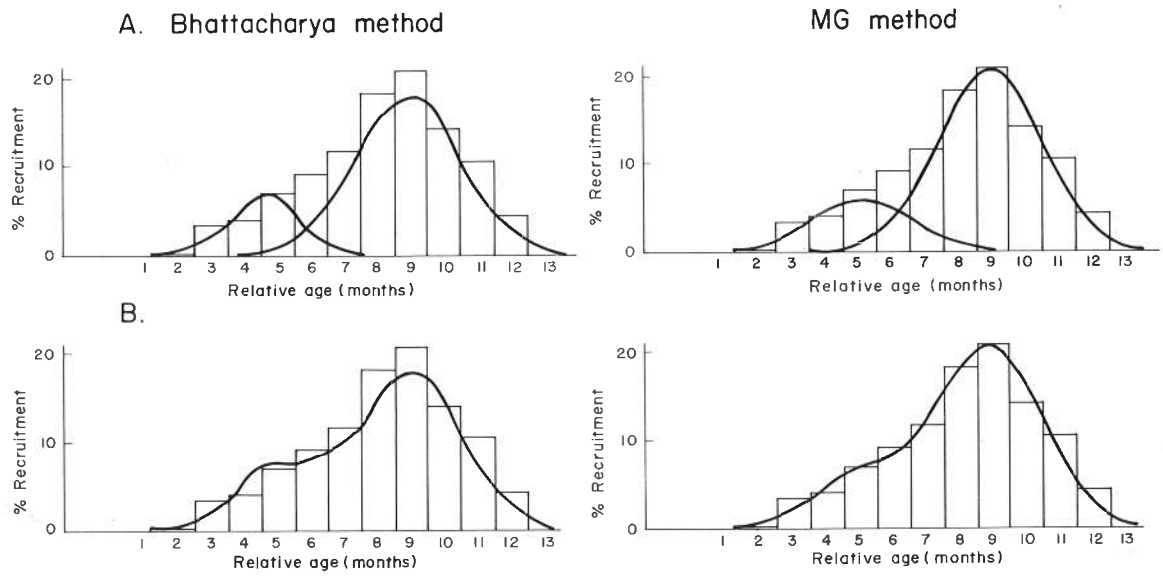

Fig. 3. Gaussian components (A) and composite distribution (B) of the recruitment pattern of *Oxyporhamphis convexus* identified using the Bhattacharya and MG methods.

Table 2. Estimated means, variances and subpopulations of the normal components of the recruitment pattern of *Oxyporhamphis convexus* identified using the Bhattacharya and MG methods with test for means and variances inequality and goodness of fit test for the composite distributions.

Comp. #	Bhattacharya method			MG method			Test of hypotheses	
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{N}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{N}$	H_o: $\mu_1 = \mu_2$	$\sigma_1 = \sigma_2$
1	4.65	1.058	17.78	5.09	1.523	22.15	ns	ns
2	8.90	1.771	75.97	8.92	1.562	78.02	ns	ns
	$Ks^1 = 0.079$ ns			$Ks = 0.021$				

ns - not significant at 5% level
* - significant at 1% level
[1]Ks tabular value is 0.089

Example 3. Multimodal length frequency-distribution

Another interesting application of the above techniques is in the analysis of length-frequency data (Pauly and Murphy 1982). Estimation of growth and mortality rates of fish populations are important for stock assessment purposes and these are estimated by knowing the age, length and abundance-length relationship of the fish. However, age-determination, usually done by otolith reading is time consuming and impractical for large samples. Generally the basic data on fish population come in the form of length frequencies. Under the assumption that the separate modes in length-frequency distribution could be interpreted as indicating age-groups (MacDonald and Pitcher 1979), the Bhattacharya and the MG methods can then be used for assessing fish stocks.

Length-frequency data of ring tailed surgeon fish, *Acanthurus xanthopterus*, were collected and separation of normal components was made using the Bhattacharya and MG methods (Fig. 4). Again for both methods, five normal components were identified. Table 2 shows the comparison of the estimated means and variances for both methods. The test shows no significant difference between corresponding means and variances. Table 3 for test of index of goodness of fit of the composite distributions using the Kolmogorov-Smirnov statistic shows that these are not significantly different from the observed frequency distributions at 5% level of significance.

Discussion

From the examples shown above, the Bhattacharya and MG methods have proven to be

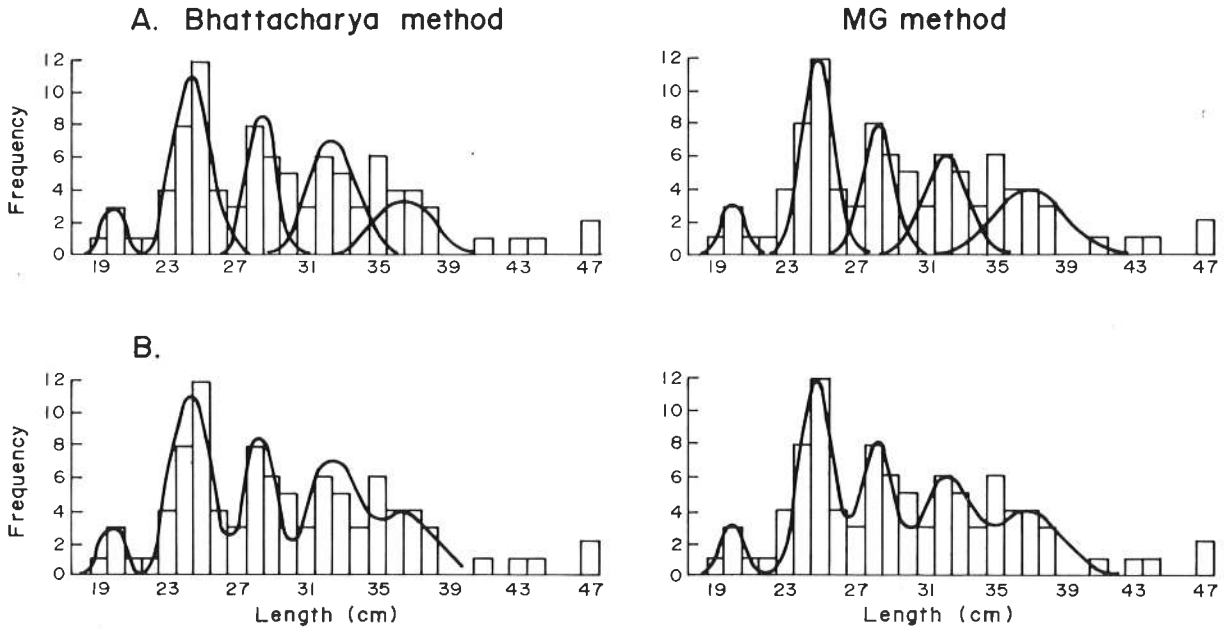

Fig. 4. Gaussian components (A) and composite distribution (B) of length-frequency data of *Acanthurus xanthopterus* identified using the Bhattacharya and MG methods.

Table 3. Estimated means, variances and subpopulations of the normal components of the *Acanthurus xanthopterus* length-frequency identified using the Bhattacharya and MG methods with test for means and variances inequality.

Comp. #	Bhattacharya method			MG method			Test of hypotheses	
	m	s	N	m	s	N	H_o: $m_1 = m_2$	$s_1 = s_2$
1	20.00	0.675	5.04	20.00	0.7026	5.55	ns	ns
2	24.45	1.128	30.12	24.83	0.9281	28.06	ns	ns
3	28.38	0.772	16.71	28.24	0.9182	18.52	ns	ns
4	32.50	1.285	22.18	32.14	1.3519	19.74	ns	ns
5	36.61	1.772	16.14	36.91	2.0954	20.90	ns	ns
	Ks^2 = 0.035 ns			Ks = 0.042 ns				

ns - not significantly different at 5% confidence level
[1]Ks tabular value is 0.091

comparable especially when overlapping between adjacent components is not too strong. The main advantage of the MG method is that it completely eliminates subjective bias in identifying the components of the mixture. The mean component using the MG method fits closely on the modes of the distribution. This is helpful especially in exploratory data analysis and/or when the user has no preconceived idea of what the component distributions are. But the user of both methods should also be forewarned about identification of more components than there should really be, just to get an acceptable composite distribution. Such a case of overfitting the model is dangerous to data analysis. The researcher should investigate more about the data to guard him or her against spurious results.

References

Bhattacharya, C.G. 1967. A simple method of resolution of a distribution into Gaussian components. Biometrics 23: 115-135.

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

Gayanilo, F.C., M.L. Soriano and D. Pauly. 1988. A draft guide to the Compleat ELEFAN. ICLARM Software 2, 65 p.

Gregor, J. 1969. An algorithm for the decomposition of a distribution into Gaussian components. Biometrics 25: 79-93.

Gulland, J.A. 1983. Fish stock assessment. FAO/Wiley Series on Food and Agriculture. Vol. 1. 223 p.

Harding, J.P. 1949. The use of probability paper for the graphical analysis of polymodal frequency distributions. J. Mar. Biol. Assoc. UK 28: 141-153.

MacDonald, P.D.M. and T.J. Pitcher. 1979. Age-groups from size-frequency data: a versatile and efficient method of analyzing distribution mixtures. J. Fish. Res. Board Can. 36: 987-1001.

Pauly, D. and G.I. Murphy, editors. 1982. Theory and management of tropical fisheries. ICLARM Conference Proceedings 9, 360 p.

Pearson, K. 1914. A study of trypanosome strains. Biometrika 10: 85-143.