

User's Guide to Program MIX:  
An Interactive Program for Fitting  
Mixtures of Distributions

Release 2.3  
January 1988

by  
P.D.M. Macdonald  
and  
P.E.J. Green

ICHTHUS DATA SYSTEMS  
59 Arkell Street  
Hamilton, Ontario  
Canada L8S 1N6

Copyright © 1988 ICHTHUS DATA SYSTEMS

ISBN 0-9692305-1-6

Printed in Canada by Guenther Printing, 66 Pleasant Avenue, Hamilton, Ontario, Canada L9C 4M7.

This publication is documentation for the computer program MIX.

MIX is proprietary software. ICHTHUS DATA SYSTEMS has the sole and exclusive right to distribute MIX and to grant licences. If you wish a licence to use MIX, please contact Peter Macdonald at ICHTHUS DATA SYSTEMS, 59 Arkell St, Hamilton, Ontario, Canada L8S 1N6, telephone (416) 527-5262.

A copy of the standard licence agreement form is shown on page 60. *Please respect the terms of the licence agreement.* Because users have paid for MIX, we are able to upgrade MIX and improve its documentation. Licensed users of MIX are offered upgrades to subsequent releases at a much reduced price.

The run-time library in the Apple Macintosh version of MIX 2.3 is  
Copyright © Absoft Corporation, 1987.

The run-time library in the IBM PC versions of MIX 2.3 is  
Copyright © Microsoft Corporation, 1982-1988.

The IBM PC versions of MIX 2.3 include graphics routines from the GRAFMATIC Library,  
Copyright © Microcompatibles, Inc., 1984.

IBM PC is a registered trademark of the International Business Machines Corporation. Macintosh is a trademark licensed to Apple Computer, Inc. MacDraw is a trademark of Apple Computer, Inc. VAX and VMS are trademarks of the Digital Equipment Corporation. UNIX is a trademark of Bell Laboratories.

## TABLE OF CONTENTS

1. Introduction .....	1
1.1 MIX: An interactive program for fitting mixtures of distributions .....	1
1.2 Special features for length-frequency analysis .....	4
1.3 Computer requirements .....	4
1.4 Screen graphics for the IBM PC .....	5
2. Statistical and numerical methods .....	6
2.1 Fitting a mixture distribution to grouped data by maximum likelihood .....	6
2.2 Constraints on the parameters .....	8
2.2.1 Constraints on proportions .....	8
0 (none) .....	8
1 (Specified proportions fixed) .....	8
2.2.2 Constraints on means .....	8
0 (none) .....	8
1 (Specified means fixed) .....	8
2 (Means equal) .....	9
3 (Equally spaced) .....	9
4 (Growth curve) .....	9
2.2.3 Constraints on sigmas .....	10
0 (None) .....	10
1 (Specified sigmas fixed) .....	10
2 (Fixed coefficient of variation) .....	10
3 (Constant coefficient of variation) .....	11
4 (Sigmas equal) .....	11
2.3 Numerical precision .....	11
3. How to run MIX .....	12
Option 0. List of options .....	12
Option 1. Read a new set of data .....	12
Option 2. Read a full set of parameter values .....	13
Option 3. Revise specified parameter values .....	13

Option 4. Estimate proportions for fixed means, sigmas.....	14
Option 5. Estimate means, sigmas for fixed proportions.....	14
Option 6. Estimate proportions, means, sigmas.....	16
Option 7. Restore parameters to values from previous step. ....	17
Option 8. Regroup data or restore to original grouping. ....	17
Option 9. Choose a distribution. ....	18
Option 10. Plot histogram.....	18
Option 11. Plot histogram and fitted components. ....	18
Option 12. Toggle to echo all I/O to I/O log.....	20
Option -1. STOP.....	20
4. Strategies for difficult cases.....	20
4.1 What to do when iterations will not converge.....	20
4.2 What to do when proportions go negative or do not sum to 1.....	23
4.3 What to do when there are small expected counts.....	23
5. The analysis of fisheries length-frequency distributions.....	24
6. Technical support for MIX.....	25
7. Licence agreement.....	25
8. Upgrades.....	26
References.....	26
Appendix.....	27
Example: An analysis of Heming Lake pike data.....	27
Standard licence agreement for MIX users.....	60

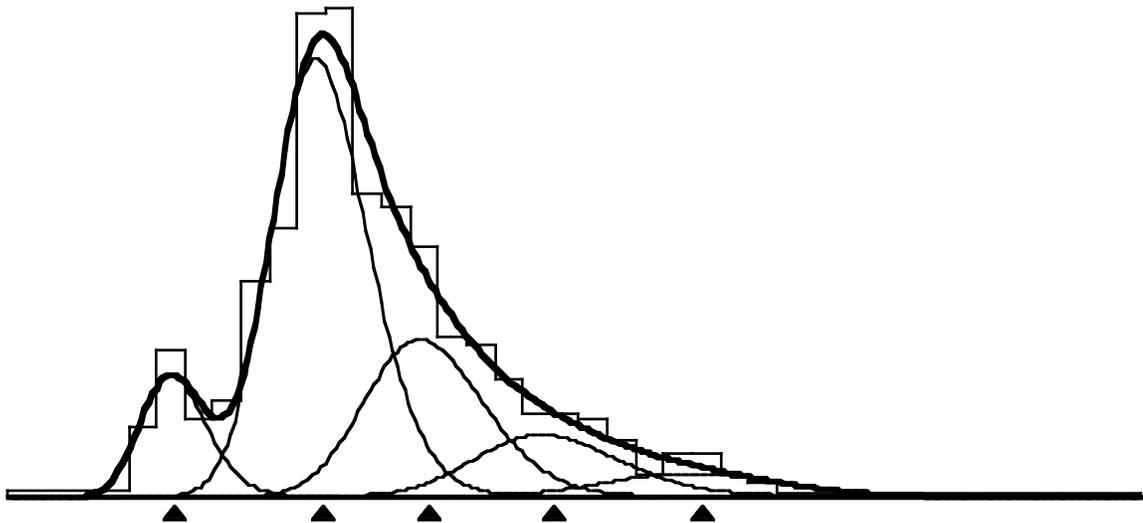
# User's Guide to Program MIX

## 1. INTRODUCTION

### 1.1 MIX: An interactive program for fitting mixtures of distributions

MIX analyzes histograms as mixtures of statistical distributions, that is, by finding a set of overlapping component distributions that gives the best fit to the histogram. The components can be normal, lognormal, exponential or gamma distributions. An example is shown in Figure 1; there are five component lognormal distributions with different weights, and their sum, shown as a thick line, matches the shape of the histogram as closely as possible. The statistical method used to fit the mixture distribution to the data is maximum-likelihood estimation for grouped data. MIX will fit up to fifteen components, with the data grouped over as many as eighty grouping intervals.

This is the best way to analyze samples from mixed populations. Size-frequency distributions in animal populations with distinct age-groups, times to failure in a mixture of good and defective items, and the distribution of some diagnostic measure in a mixed population of patients, some of whom have a given disease and some of whom do not, are all examples of mixed populations. MIX can also be used in a more general, descriptive, way to analyze multimodal and other irregularly-shaped histograms.



Plot #001 Data: Heming Lake Pike 1965 Components: Lognormal

Figure 1. An example of fisheries length-frequency analysis, shown with high-resolution graphics. The five components correspond to the five age-groups in the population, the thick line is their sum, the mixture distribution. The abscissa unit is length in cm. The triangles mark the mean lengths of the age-groups.

The prototype of MIX was developed by Macdonald and Pitcher (1979) for the analysis of fisheries length-frequency data, and this remains an important application (Macdonald 1987). Figures 1 and 2 show an example of length-frequency analysis.

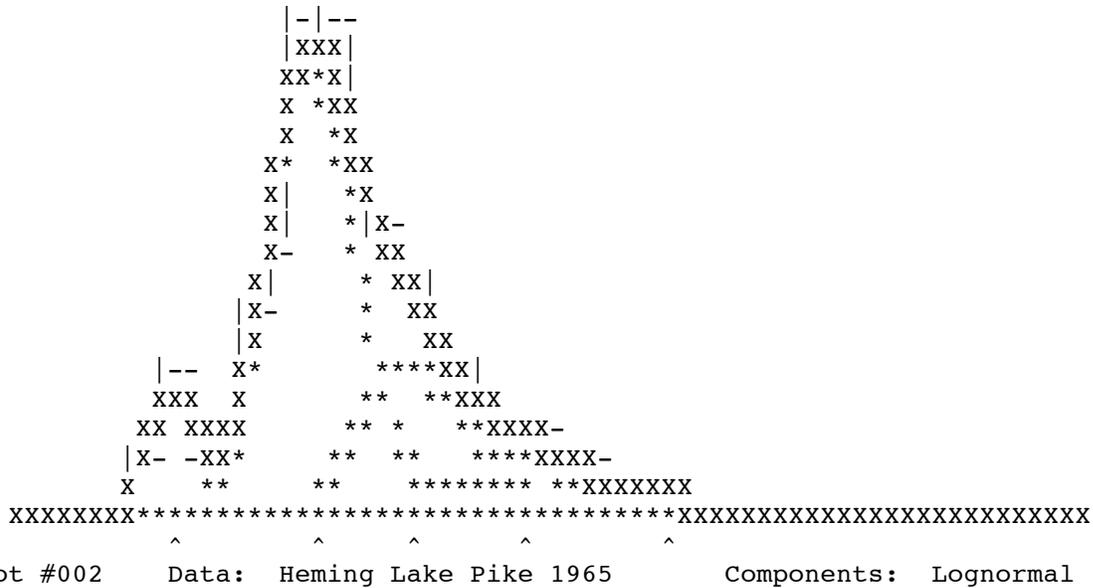


Figure 2. An example of text-mode graphics. This is the same fit as shown in Figure 1.

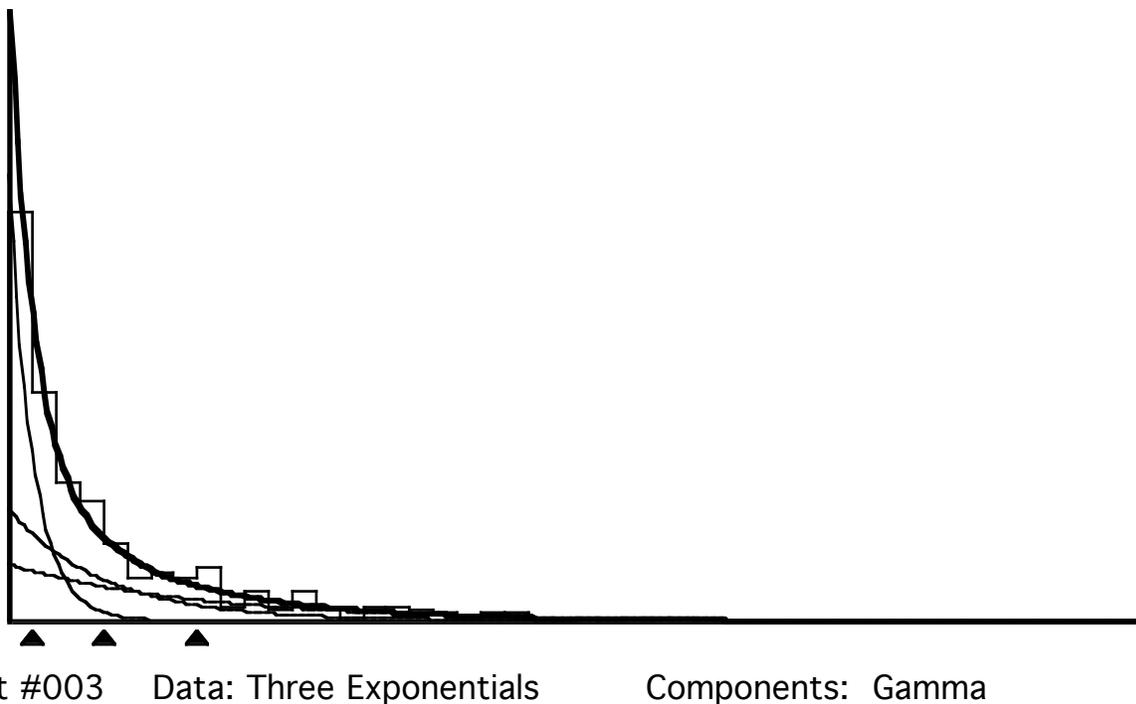
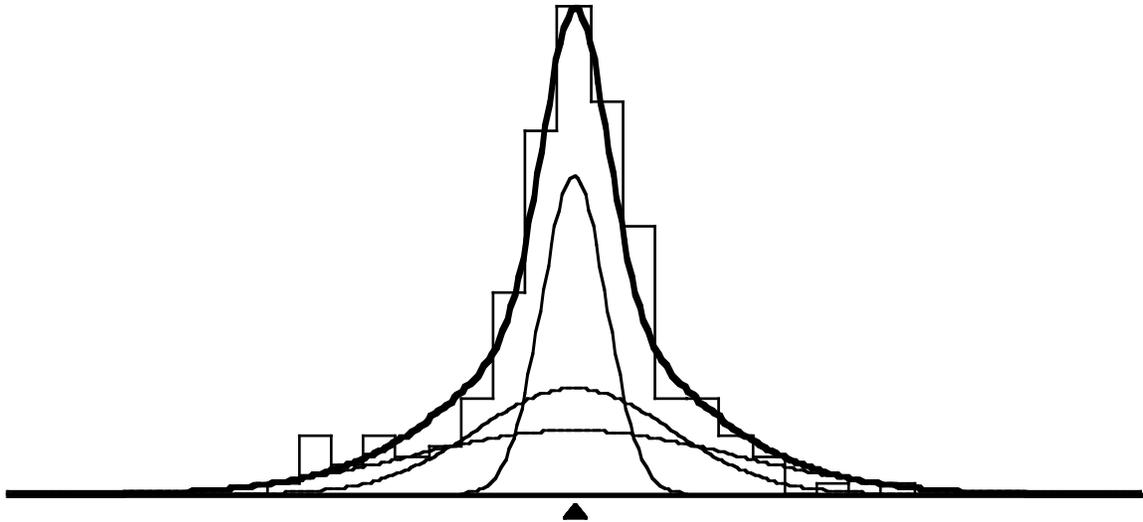


Figure 3. A mixture of three exponential distributions fitted by MIX. Exponential distributions are fitted as gamma distributions with unit coefficient of variation; see §2.2.3.

MIX can also handle many other mixture distribution applications, such as mixtures of exponential distributions for time-to-failure studies (Figure 3) and scale mixtures with equal means for non-normal error analysis (Figure 4). Titterington *et al.* (1985) describe many applications of mixtures where the current version of MIX will give useful results.



Plot #004 Data: Means Equal Components: Normal

Figure 4. A scale mixture of three normal distributions fitted by MIX. A scale mixture has equal means, different standard deviations; see §2.2.2.

Estimating the parameters of a mixture distribution is difficult when the components are heavily overlapped because the overlapping obscures information about individual components. The mixture can only be resolved by bringing additional information to the problem. This information could be from additional samples, or from some form of prior information about the parameters and the relations between them. MIX allows the user to impose constraints on the parameters; for example, holding some parameters fixed, or constraining all the components to have the same coefficient of variation. The user can start with as many constraints on the parameters as necessary and work interactively towards a solution which has as few constraints as possible and makes sense in terms of the application.

A future release of MIX will allow the user to incorporate additional data in the analysis, in the form of stratified sub-samples: in length-frequency applications, sub-samples for age-determination would be taken at specific lengths, and analysed jointly with the overall length-frequency distribution.

MIX features a convenient interactive style; a choice between extremely rapid quasi-Newton optimization or slower but more fool-proof Nelder-Mead simplex optimization; extensive error checks; and excellent high-resolution screen graphics. With screen graphics, the user can often get very close to the optimal solution by simple visual steps, then use numerical optimization to finish off the fitting process. MIX computes standard errors for all estimates, and a goodness-of-fit test of the final fit.

## 1.2 Special features for length-frequency analysis

Most length-frequency applications can be handled by constraining either the component standard deviations or the component coefficients of variation to be equal (Macdonald 1987). However, in many applications there is an ill-defined ‘smear’ of older age-groups with relatively

small numbers in the right-hand tail of the distribution. These age-groups are sometimes best lumped into a single component, but its standard deviation may then be relatively large. When fitting three or more components, MIX allows you to estimate the standard deviation of the rightmost component separately or hold it fixed, while constraining the remaining components to have equal standard deviations or equal coefficients of variation.

When four or more components are being fitted the means can be constrained to lie along a von Bertalanffy growth curve (§2.2.2). The usual growth-curve parameters  $L_\infty$ ,  $k$  and  $t_1-t_0$  are computed. MIX computes standard errors for these parameters. Linear growth is permitted by constraining the means to be equally spaced. If the rightmost component represents all the oldest age-groups lumped together, you may choose to estimate its mean separately, or hold it fixed, while constraining the remaining means to lie on a growth curve or to be equally spaced.

### 1.3 Computer requirements

Versions of MIX are available for the IBM PC and PC compatibles, Apple Macintosh and mainframes. Some steps of the fitting process require heavy iterative calculation. On a mainframe, a Macintosh II, or an IBM PC-AT or COMPAQ® 386 with a floating-point coprocessor, most steps will be completed within a few seconds. A Macintosh Plus or an IBM PC-XT with an 8087 coprocessor will give quite acceptable execution speeds, but some steps may take a few minutes to complete. An IBM PC-XT without a coprocessor may take a few minutes to complete certain steps and may sometimes take an hour or more. All microcomputer versions display an iteration counter to show how quickly the iterations are progressing and beep when the iterations are completed.

- An IBM PC or a PC compatible should have at least 512K RAM and run MS-DOS 2.1 or higher. An 8087 floating-point coprocessor, while not required, is highly recommended as it speeds up calculation by about a factor of 10. High-resolution graphics require either a CGA, EGA or Hercules graphics card but if one of these is not available MIX will produce rough screen graphics in text mode (Figure 2). One disk drive is sufficient. MIX is supplied as an executable file.
- The Apple Macintosh version will work with a 512K Macintosh, but a Macintosh Plus, SE or II is preferred. One disk drive is sufficient. MIX is supplied as a stand-alone application in two versions. One will run on a Macintosh 512K, Plus, or SE. The other requires the MC68020 processor and MC68881 coprocessor on a Macintosh II or upgraded SE and gives incredibly high execution speeds.
- The mainframe version is supplied as ANSI Standard FORTRAN 77 source code. It has been compiled and tested on many systems, including VAX VMS, VAX UNIX, Pyramid UNIX, and Prime. Code to drive an off-line CALCOMP plotter is included, and this code can be adapted to other plotters. Screen graphics are in text mode (Figure 2).

---

## 1.4 Screen graphics for the IBM PC

MIX 2.3 will produce high-resolution monochrome screen graphics with either a CGA, EGA or Hercules graphics card. You must have the correct version of MIX 2.3; there is one version for CGA and EGA and another version for Hercules. The CGA card gives a resolution of 640×200 pixels, the EGA card gives either 640×200 pixels or 640×350 pixels, and the Hercules card gives 720×348 pixels.

The IBM PC versions of MIX 2.3 are linked with subroutines from the GRAFMATIC library, a product of Microcompatibles, Inc., 301 Prelude Drive, Silver Spring, MD 20901, U.S.A. It is an excellent collection of primitive and advanced graphics routines that can be linked with FORTRAN or PASCAL programs. ICHTHUS DATA SYSTEMS is licensed to distribute executable code linked with GRAFMATIC object modules.

Your ability to get a hard copy print-out of the screen in graphics mode will depend on what combination of graphics card, printer and operating system utilities you have. If you find that you are unable to print the screen, we recommend GRAFPLUS by Jewell Technologies. GRAFPLUS can be purchased from Microcompatibles for U.S.\$50.00 (1987 price, subject to change). When the GRAFPLUS or GRAFLASR command is executed from DOS you specify the graphics card and printer you are using. From then on, until the system is re-booted, the “print screen” function key, or an equivalent software command, will dump screen graphics to the printer. You can also use GRAFPLUS to save screen graphics to a file, to be retrieved and printed later.

If you do not have a graphics card, if your graphics card is not sufficiently compatible with CGA, EGA, or Hercules, or if you are running MIX 2.3 on a machine with the wrong graphics card, you have the option of text-mode graphics instead of high-resolution graphics; just respond with N at the prompt asking if the correct graphics card is installed. This prompt comes the first time you use Option 10 or Option 11 to draw a graph.

## 2. STATISTICAL AND NUMERICAL METHODS

### 2.1 Fitting a mixture distribution to grouped data by maximum likelihood

A finite mixture distribution arises when samples are drawn from a population that is a mixture of  $k$  component populations. Letting  $\pi_i$  represent the proportion of the total population that the  $i^{\text{th}}$  component population constitutes and letting  $f_i(x)$  represent the probability density function for some variable characteristic  $X$  within the  $i^{\text{th}}$  component population, then

$$g(x) = \pi_1 f_1(x) + \dots + \pi_k f_k(x)$$

is the probability density function for  $X$  in the mixed population.

MIX assumes that the components can be described by either normal, lognormal or gamma probability distributions. These are two-parameter distributions and without loss of generality the parameters can be taken to be the mean and standard deviation. Let  $\mu_i$  represent the mean and  $\sigma_i$  the standard deviation of the  $i^{\text{th}}$  component density  $f_i(x)$ . The objective of fitting the mixture to

data is to estimate as many as possible of the parameters  $\pi_1, \dots, \pi_k; \mu_1, \dots, \mu_k; \sigma_1, \dots, \sigma_k$ . The component standard deviations  $\sigma_1, \dots, \sigma_k$  are referred to as the “sigmas” in output from MIX.

For theoretical and practical reasons it will not always be possible to estimate all of the parameters, particularly when the components overlap and obscure one another. This is discussed by Macdonald and Pitcher (1979). Thus it is often desirable to reduce the number of parameters by assuming constraints. The proportions are, of course, already subject to the constraint  $\pi_1 + \dots + \pi_k = 1$ , so there are only  $k-1$  “free” proportions. Suitable constraints for the means and standard deviations will depend on the application. It may be that, for some component  $i$ ,  $\mu_i$  and  $\sigma_i$  are known from other data and can be held fixed at those given values. In some applications it may be reasonable to assume that the standard deviations are all equal,  $\sigma_1 = \dots = \sigma_k$  or that the coefficients of variation are all equal,  $(\sigma_1/\mu_1) = \dots = (\sigma_k/\mu_k)$ . These and other constraints allowed by MIX are discussed in §2.2.

MIX assumes that the data are grouped, in the form of numbers of observations over successive intervals. Data often come grouped (as a histogram) or can be grouped with very little loss of information. Grouping greatly simplifies the calculation of maximum likelihood estimates (Macdonald and Pitcher 1979). The grouping intervals are specified by their right-hand boundaries. The first (leftmost) and last (rightmost) intervals are open-ended; that is, if there are  $m$  intervals, the first interval includes everything up to the interval boundary  $x_1$ , the second everything from  $x_1$  to  $x_2$ , and so on to the  $m-1^{st}$  interval, which includes everything from  $x_{m-2}$  to  $x_{m-1}$ , and the  $m^{th}$ , which includes everything above  $x_{m-1}$ . Thus it is only necessary to specify  $m-1$  boundaries. The choice of boundaries is discussed in §5 and in Macdonald and Pitcher (1979).

MIX can be used if percent, mass, or something other than a sample count is given for each interval, but the standard errors of the estimates and the goodness-of-fit tests will not be valid in such cases, except in a relative sense within the analysis of a given data set.

MIX can also be used to test the goodness-of-fit of the model to the data and, in some cases, it can be used to test the validity of certain constraints. These tests depend on the chi-square approximation to the likelihood ratio statistic (Rao 1965) and will be valid as long as most of the intervals have expected counts of 5 or greater. The goodness-of-fit chi-square statistic is printed after each fitting step. The degrees of freedom are computed as the number of grouping intervals minus 1 minus the number of parameters estimated. Note that MIX does not count parameters that were held fixed during an estimation step as parameters estimated: if in fact they had been adjusted to fit the data at an earlier step in the session they have in a sense been estimated and the degrees of freedom computed by MIX should be reduced by at most 1 for each such parameter. After a successful fit, MIX will compute a significance level (P-value) for the goodness-of-fit test (see Option 6 in §3). In the situation just described, where some fixed parameters had been estimated at earlier steps, the P-value should be re-calculated from a table of chi-square, using the reduced degrees of freedom. If the counts in most intervals are small (most less than 5, say), then the P-value given should be considered as a poor approximation. If the data give percents, mass, or anything other than counts over the grouping intervals, then the P-value will have no meaning,

although a reduced “chi-square” value will still indicate an improved fit, relative to another fit to the same data.

If the data can be fitted with and without a certain constraint, the validity of that constraint can be tested. Removing the constraint will, in general, reduce the chi-square and the degrees of freedom; the reduction in chi-square is itself a chi-square statistic with degrees of freedom equal to the reduction in degrees of freedom (Rao, 1965, p.350). This is only valid if the data give actual counts over intervals and if most counts are 5 or greater. In this way, it is possible to test whether or not the proportions of the mixture are all equal, whether or not the means lie on a growth curve, or whether or not the data came from a mixture of exponential distributions, to give just a few examples. The test for exponential distributions is done by fitting gamma distributions, first with the constraint that the coefficient of variation be fixed at 1, then without that constraint.

In the Example in the Appendix, the hypothesis that the means lie on a growth curve (assuming lognormal distributions and a constant coefficient of variation) can be tested by a chi-square statistic of  $12.4566 - 11.9477 = 0.5089$  on  $16 - 14 = 2$  degrees of freedom. The fits used in this test are found on pages 50 and 44. Since  $P = 0.78$ , the hypothesis that the means lie on a growth curve cannot be rejected.

The goodness-of-fit test only indicates how well the mixture distribution  $g(x)$  fits the histogram overall. If the components overlap extensively the test is not very sensitive to features that are obscured by the overlapping, such as skewness of the component distributions. Hence we cannot conclude from the analyses in the Appendix whether the component distributions in the pike data are really normal, lognormal or gamma; each fit is about as good as the other. Similarly, the test shown above, to determine whether or not the means lie on a growth curve, has very low power.

## 2.2 Constraints on the parameters

The constraints on the parameters are explained below, under the headings that will appear on the screen as prompts.

### 2.2.1 Constraints on proportions

#### 0 (none)

Only the natural constraint  $\pi_1 + \dots + \pi_k = 1$  is imposed. MIX does not constrain the proportions to be non-negative. Negative values can occur in some pathological situations and suggestions for handling them are given in §4.2.

#### 1 (Specified proportions fixed)

In addition to the natural constraint  $\pi_1 + \dots + \pi_k = 1$ , any or all of the proportions may be held fixed while other parameters are being estimated. If  $a$  is the number of proportions held fixed in this way, the number of free proportions is  $k-a-1$ , where the  $-1$  accounts for the natural constraint. If, for example,  $k = 5$  and you want the third and fifth proportions to be held fixed, enter NNYNY at the prompt, without separators between the characters. MIX does not constrain the proportions to be non-negative. Negative values can occur in some pathological situations and suggestions for handling them are given in §4.2.

To constrain the proportions to be equal, hold each one fixed at  $1/k$ .

### 2.2.2 Constraints on means

#### 0 (none)

MIX will attempt to estimate all  $k$  means  $\mu_1, \dots, \mu_k$ .

#### 1 (Specified means fixed)

Specified means may be held fixed while MIX attempts to estimate the remaining means. If, for example,  $k = 5$  and you want the third and fifth means to be held fixed, enter NNYNY at the prompt, without separators between the characters.

#### 2 (Means equal)

This constraint assumes that  $\mu_1 = \mu_2 = \dots = \mu_k$ . MIX attempts to estimate their common value. The common value is initialized at  $\mu_1$ . This constraint is allowed if there are at least two components and the standard deviations are all different from each other; such a mixture is called a “scale mixture” (Figure 4).

#### 3 (Equally spaced)

This constraint assumes that  $(\mu_2 - \mu_1) = (\mu_3 - \mu_2) = \dots = (\mu_k - \mu_{k-1})$ . Only two means,  $\mu_1$  and  $\mu_2$ , are estimated directly. Subsequent means are computed from the relation

$$\mu_i = \mu_1 + (i - 1)(\mu_2 - \mu_1), \quad i = 3, \dots, k.$$

In size-frequency applications where the  $\mu_i$  are mean sizes in successive age-groups, this constraint corresponds to the assumption of linear growth. This constraint is allowed if there are at least three components.

If there are four or more components, MIX gives the option to let the  $k^{\text{th}}$  (rightmost) component be different while constraining  $\mu_1, \dots, \mu_{k-1}$  to be equally spaced;  $\mu_k$  can then be held fixed or estimated separately.

#### 4 (Growth curve)

This constraint forces the means to lie along a von Bertalanffy growth curve of the form

$$\mu_i = L_\infty \{1 - \exp[-\kappa(t_i - t_0)]\}$$

where the components are assumed to be age-groups spaced exactly one year apart,  $\mu_i$  is the mean size of individuals in the  $i^{\text{th}}$  age-group, age is measured in years,  $t_0$  is the hypothetical age at zero size,  $t_i$  is the actual age of the  $i^{\text{th}}$  age-group,  $L_\infty$  is the hypothetical ultimate mean size of individuals in that population and  $\kappa$  is the growth parameter. Only the first three means  $\mu_1, \mu_2, \mu_3$  are estimated. Subsequent means are computed from the relation

$$\mu_i = \mu_1 + \frac{(\mu_2 - \mu_1)^2}{(\mu_2 - \mu_1) - (\mu_3 - \mu_2)} \left\{ 1 - \left( \frac{\mu_3 - \mu_2}{\mu_2 - \mu_1} \right)^{i-1} \right\}, \quad i = 4, \dots, k.$$

It can be shown that

---

$$L_{\infty} = \mu_1 + \frac{(\mu_2 - \mu_1)^2}{(\mu_2 - \mu_1) - (\mu_3 - \mu_2)}$$

$$\kappa = -\log_e \{(\mu_3 - \mu_2)/(\mu_2 - \mu_1)\}$$

$$(t_1 - t_0) = -\kappa^{-1} \log_e \left( 1 - \frac{\mu_1}{L_{\infty}} \right)$$

MIX computes and displays these three values and their standard errors but it must be remembered that they are very unreliable when estimated from data (Schnute and Fournier 1980). The fitted values of  $\mu_1$ ,  $\mu_2$ ,  $\mu_3$  are much more interpretable.

The growth curve constraint is allowed if there are at least four components. It cannot be used unless  $(\mu_3 - \mu_2) < (\mu_2 - \mu_1)$ ; if this does not hold, Option 3 can be used to increase  $\mu_2$  until it does hold.

If there are five or more components, MIX gives the option to let the  $k^{th}$  (rightmost) component be different while constraining  $\mu_1, \dots, \mu_{k-1}$  to lie on a growth curve;  $\mu_k$  can then be held fixed or estimated separately.

### 2.2.3 Constraints on sigmas

#### 0 (None)

MIX will attempt to estimate all  $k$  standard deviations  $\sigma_1, \dots, \sigma_k$ . If all the proportions and all the means are also being estimated, this choice is not likely to work unless the  $k$  components show as  $k$  clear modes in the histogram.

#### 1 (Specified sigmas fixed)

Specified standard deviations will be held fixed while MIX attempts to estimate the remaining standard deviations. If, for example,  $k = 5$  and you want the third and fifth standard deviations to be held fixed, enter NNYNY at the prompt, without separators between the characters.

#### 2 (Fixed coefficient of variation)

The coefficients of variation  $(\sigma_1/\mu_1), \dots, (\sigma_k/\mu_k)$  will all be held at the same fixed value. MIX will display the current value of  $\sigma_1/\mu_1$  and give the option to use that value as the fixed value or input a new value. Because the coefficient of variation and the means completely determine the standard deviations, the standard deviations do not count as estimated parameters. This constraint is allowed if all of the means are positive and different from each other.

Note that if the components are gamma distributions, fixing the coefficient of variation at 1 will force them to be exponential distributions, since, for the gamma distribution,  $\sigma/\mu = p^{-1/2}$ , where  $p$  is the shape parameter (Rao 1965, p.133), and a gamma distribution with  $p = 1$  is an exponential distribution.

If there are three or more components, MIX gives the option to make the  $k^{th}$  (rightmost) component different while constraining components 1 to  $k-1$  to have a fixed coefficient of variation;  $\sigma_k$  can then be held fixed or estimated separately.

**3 (Constant coefficient of variation)**

This constraint assumes that  $(\sigma_1/\mu_1) = (\sigma_2/\mu_2) = \dots = (\sigma_k/\mu_k)$  and MIX attempts to estimate the common value. The common value is initialized at  $\sigma_1/\mu_1$ . MIX estimates  $\sigma_1$  and computes the other standard deviations from the relation

$$\sigma_i = (\mu_i / \mu_1) \sigma_1, \quad i = 2, \dots, k.$$

This constraint is allowed if there are at least two components and all of the means are positive and different from each other.

If there are three or more components, MIX gives the option to make the  $k^{th}$  (rightmost) component different while constraining components 1 to  $k-1$  to have a constant coefficient of variation;  $\sigma_k$  can then be held fixed or estimated separately.

**4 (Sigmas equal)**

This constraint assumes that  $\sigma_1 = \dots = \sigma_k$ . MIX attempts to estimate the common value. The common value is initialized at  $\sigma_1$ . This constraint is allowed if there are at least two components and the means are all different from each other.

If there are three or more components, MIX gives the option to make the  $k^{th}$  (rightmost) component different while imposing the constraint  $\sigma_1 = \dots = \sigma_{k-1}$ ; you can then hold  $\sigma_k$  fixed or estimate it separately.

**2.3 Numerical precision**

Accuracy of the final estimates to four significant digits is adequate for most practical applications. The estimates will be accurate to at least five significant digits because the normal and gamma probability integrals computed by MIX are generally accurate to at least seven digits. Iterations in Option 4 and Option 6 continue until the *absolute* difference from the previous iteration is less than 0.0000005 for each parameter. Absolute rather than relative difference was used on the assumption that measurement units would be chosen to keep the order of magnitude of the means and standard deviations more or less in the range of 1 to 100.

The Nelder-Mead optimization in Option 5 is the step most sensitive to imprecision because large changes in parameter values may only affect the least significant digits of the chi-square being minimized. For this reason, all versions of MIX use DOUBLE PRECISION arithmetic throughout.

The subroutine for computing the gamma probability integral includes code for computing the derivative with respect to the shape parameter. We have not seen this calculation in any other statistical software.

**3. HOW TO RUN MIX**

The method of opening MIX and initiating execution will depend upon your computer and operating system. If special instructions are needed for your version of MIX, special

---

documentation will be provided, either on a separate sheet of instructions or in a file called README on the program disk.

When execution starts, you will be prompted to respond with Y if you wish to see the List of Options displayed or N if you wish to proceed directly to the prompt for a choice of Option. If you type Y the following will appear on the screen:

LIST OF OPTIONS

0. List of options.
  1. Read a new set of data.
  2. Read a full set of parameter values.
  3. Revise specified parameter values.
  4. Estimate proportions for fixed means, sigmas.
  5. Estimate means, sigmas for fixed proportions by constrained search.
  6. Estimate proportions, means, sigmas with or without constraints and/or give diagnostic displays.
  7. Restore parameters to values from previous step.
  8. Regroup data or restore to original grouping.
  9. Choose a distribution.
  10. Plot histogram.
  11. Plot histogram and fitted components.
  12. Toggle to echo all I/O to I/O log.
- 1. STOP.

MIX is designed so that any option may be chosen at any step. Illogical choices, such as attempting to do a fit before data have been read or attempting to estimate proportions when only one component is being fitted, will be skipped over after an explanatory message is displayed.

The use of each Option is described in detail below.

**Option 0. List of options.**

Display the list of the Options on the screen.

**Option 1. Read a new set of data.**

Input a new data set, either from a prepared file or from the keyboard. The data may then be edited and written onto a file.

If you are entering data from the keyboard, MIX will prompt for a title (1 to 25 characters) and the number of grouping intervals. A maximum of 80 grouping intervals is allowed. If there are  $m$  grouping intervals in the data there will be  $m-1$  right-hand boundaries to enter (see §2.1) so MIX will first prompt for  $m-1$  counts and right boundaries. The right boundaries must be in strictly ascending order. Enter each count and right boundary pair on a new line; enter the count first and separate it from the right boundary by a space or a comma. When all  $m-1$  pairs have been entered, MIX will prompt for the count in the last interval. After it has been entered, the data will then be displayed on the screen for verification. There is then provision to re-enter any count and right boundary (or, in the case of the rightmost interval, just the count) if required. MIX checks to see if

all the boundaries are in strictly ascending order and will not proceed to the next step until any exceptions have been corrected.

Data files can be created beforehand using a text editor. Write the title in columns 1 to 25 on the first line and write the number of intervals on the second line. Then write the pairs of counts and right boundaries, starting each pair on a new line and separating the count from the right boundary by a comma or space. End with the count from the rightmost interval, again on a new line. Place an empty line at the end of the file. Make sure that your text editor saves the file as an ASCII file with no extraneous control characters.

Boundaries must be in strictly ascending order. The provisions for editing and checking, described above, will be used when MIX reads the data file. There is a limit of 80 grouping intervals. If the data on a file exceed that limit, counts after the 80<sup>th</sup> interval will be added into the count for the 80<sup>th</sup> interval.

### **Option 2. Read a full set of parameter values.**

Read in the number of components in the mixture and a complete set of parameter values. The maximum number of components allowed is 15. The components must be indexed so that the means are in non-decreasing order,  $\mu_1 \leq \mu_2 \leq \dots \leq \mu_k$ . If any two consecutive means are equal the corresponding standard deviations must be in strictly ascending order. That is,  $\mu_i = \mu_{i+1}$  is allowed only if  $\sigma_i < \sigma_{i+1}$ . MIX will not accept values unless these requirements are satisfied.

If the proportions do not sum to 1, you will be given the option to re-scale them so that they do. If any proportion is negative, a warning will be displayed.

You can use MIX to fit a single normal, lognormal or gamma distribution, by specifying that there is just one component. This is not a mixture, so no proportion is entered.

### **Option 3. Revise specified parameter values.**

Change any of the parameters. If you change a proportion and the proportions no longer sum to 1 you will be given the option to re-scale them; however, you may not always wish to do so since re-scaling will change the value you just assigned to a proportion (unless the value was zero). Even if the proportions do not sum to 1, the iterations of Option 4 or Option 6 may converge to proportions that do sum to 1.

To display the current values of the parameters, use Option 3 and quit without making any changes. Note, however, that this will cause the values saved from the previous step to be overwritten by the current values, so they cannot be recovered by Option 7.

### **Option 4. Estimate proportions for fixed means, sigmas.**

Estimate all of the proportions while keeping all other parameters fixed. This step is very fast and usually converges. In any application where the proportions are being estimated, try Option 4 immediately after Option 2. If a negative proportion results, see §4.2.

You will be prompted for the number of iterations. Usually, 10 or 20 will be more than adequate. Entering 0 will abort this Option without changing any of the parameter values.

On the microcomputer versions of MIX, a counter displays the number of iterations. When the iterations finish, a short beep indicates convergence, a long beep indicates that the limit of iterations was reached, a double beep indicates that the iterations failed and the parameters have been restored to their values from the previous step.

### **Option 5. Estimate means, sigmas for fixed proportions by constrained search.**

While holding the proportions fixed, use Nelder-Mead direct search to fit the remaining parameters under the constraints chosen. The algorithm is based on that of O'Neill (1971); see Macdonald and Pitcher (1979) for additional references. You will have to specify upper and lower limits for the means, upper and lower limits for the sigmas, an initial step size for each parameter being estimated, the maximum number of function evaluations allowed, the frequency of convergence checks and an "accuracy index". The "accuracy index" is your required standard deviation of vertex values, that is, the square root of the variable REQMIN discussed by O'Neill (1971).

Upper and lower limits on the means and sigmas make Option 5 more efficient by keeping the search within reasonable bounds.

The initial step sizes should reflect how far you think the initial values are from the true values; if you think an initial value is within 2 units of the true value, for example, try a step size of 1 or 2. Note that if you are holding some or all of the values fixed, you must enter step sizes for all of the parameters, even though only those corresponding to free parameters will be used.

Direct-search optimization typically requires up to 100 function evaluations per parameter being estimated. However, experience has shown that a total of as few as 100 or 200 function evaluations will often suffice to get the values close enough for Option 6 to converge on the next step. Option 5 is expensive on mainframe computers and very time-consuming on microcomputers, especially if gamma distributions are being fitted, so avoid requesting more than 200 function evaluations. It is often faster to experiment with Option 6, adding constraints until convergence is achieved, then gradually lifting the constraints, than it is to wait for Option 5 to find a fit.

A convergence check frequency of 10 or 20 is recommended: this is the number of function evaluations done between checks to see if the accuracy index is satisfied.

Very roughly, if an accuracy index of 1 is attained, the value of the chi-square will have been minimized down to the units digit; if an accuracy index of 0.1 is attained, it will have been minimized down to the first decimal place. An accuracy index of 1 or 0.1 is recommended, but even if that accuracy is not attained before the limit of iterations is reached ("CONVERGENCE CRITERION NOT SATISFIED") the parameter estimates may still be good enough for Option 6 to work on the next step.

If the data give percents, mass, or anything other than counts over the grouping intervals, an accuracy index of 1 or 0.1 may not be suitable and you may have to experiment to find a better value, according to the relative magnitude of the “chi-square”.

The final prompt of Option 5 asks if you want to abort; this is the only provision for escape if you realize that your input is not appropriate. Response Y leads back to the prompt to choose an Option, response N begins the direct search.

The output is self-explanatory. The choice of constraints is indicated by acronyms **FIXED** (fixed), **MEQ** (means equal), **EQSP** (means equally spaced), **GCRV** (means on a growth curve), **FCOV** (fixed coefficient of variation), **CCOV** (constant coefficient of variation) and **SEQ** (standard deviations equal) under those parameters which, by reason of the constraints, are not being estimated directly.

It may be that the initial values lie outside the region of admissible values defined by the upper and lower limits on the means and standard deviations and the iterations never penetrate the admissible region. This will be flagged by an error message and the final value of chi-square will be .100000E+17. In extremely pathological cases, the means and standard deviations will lie within the upper and lower bounds specified but are inadmissible for some other reason. For example, they may specify a mixture that is nowhere near the observed histogram. This will be flagged by an error message and the final value of chi-square will be .100000E+16.

Option 5 should not be used if the proportions do not sum to 1 or if there is a negative proportion, as the likelihood-ratio chi-square being minimized will be meaningless; it may even be negative. If necessary, use Option 3 to prepare the proportions before entering Option 5.

On the microcomputer versions of MIX, during the direct search, a counter displays the number of function evaluations. When the iterations finish, a short beep indicates convergence, a long beep indicates that the limit of function evaluations was reached, a double beep indicates that the parameter values were inadmissible.

### **Option 6. Estimate proportions, means, sigmas with or without constraints.**

Use efficient “scoring” iterations (Macdonald and Pitcher 1979) to estimate the parameters under specific constraints. The variance-covariance matrix for the estimates is computed and standard deviations are given for the estimates. The observed and expected counts for each interval may be tabulated or graphed.

The final prompt asks if you want to abort; this is the only provision for escape if you realize that your input is not appropriate. Response Y leads back to the prompt to choose an Option, response N begins the iterations.

On the microcomputer versions of MIX, a counter displays the number of iterations. When the iterations finish, a short beep indicates convergence, a long beep indicates that the limit of iterations was reached, a double beep indicates that the iterations failed and the parameters have been restored to their values from the previous step.

The iterations will not always converge, especially if insufficient constraints are imposed or if the initial parameter values are not good. This is discussed in §4.1. For diagnostic purposes, the maximum number of iterations may be set to 0; in this case, the parameter values will not be changed but any of the tables or displays may be obtained.

In cases where the proportions are not being changed by Option 6, such as when the iteration limit is set to 0 or when all the proportions are held fixed, the proportions must be non-negative and sum to 1; otherwise, the goodness-of-fit chi-square computed by MIX will be meaningless; it may even be negative. The proportions can be prepared using Option 3.

The first prompt is for the maximum number of iterations to be allowed. In most cases, convergence will come after about 20 or 30 iterations if it comes at all, but some pathological cases will not converge until about 60 iterations. Enter 0 to get diagnostic displays without changing any of the parameter values.

The table of observed and expected counts is useful, especially to see where any small expected counts occur (§4.3). The graph of observed and expected counts is not as useful as the high-resolution graphs plotted by Options 10 and 11. It is a histogram if the grouping intervals are of equal width, but it is not re-scaled if they are not. It is useful only as a graphical representation of the table of observed and expected counts. The symbols used are:

- O Marks the observed count.
- E Marks the fitted, or expected, count, and used to shade in the column.
- X Used when an O and an E are superimposed.
- \* Used when the columns of E's goes off the page.
- I Used when an O and a \* are superimposed.

If the variance-covariance matrix for the estimates is requested, it will appear as a lower triangular matrix. The sequence of variables is: the free proportions in order; the directly estimated (or free) means in order; the directly estimated (or free) standard deviations in order.

Parameters are tabulated with their standard errors. Parameters that were held fixed are indicated by the word **FIXED** being displayed in place of a standard error. Parameters which, by reason of some other constraint, were not estimated directly, have no standard errors given.

If too many components have been assumed or too few constraints have been imposed or if the initial values are too far from the true values, either the information matrix will become singular or the parameters being estimated will iterate out of the admissible range. In either case, a message will be displayed and Option 7 will be called automatically to restore the parameters to values from the previous step. See §4.1 for a discussion of what to do next.

### **Option 7. Restore parameters to values from previous step.**

Restore parameters to their values from previous step. No input is required. The restored values are displayed.

**Option 8. Regroup data or restore to original grouping.**

Regroup the data or restore the original grouping. This option facilitates the removal or re-insertion of interval boundaries. Restoration of the original grouping is the only way to re-insert interval boundaries. Boundaries can be removed one at a time by entering a boundary at the prompt; the two intervals on either side will then become one and the two counts will be summed.

You can use Option 8 to write the data to a file. This is useful if you forgot to create a file in Option 1, or if you have regrouped the data and want to save it in its regrouped form.

To display the current data, use Option 8 without restoring to the original grouping or removing a boundary.

**Option 9. Choose a distribution.**

Select a distribution. The choice is between normal, lognormal or gamma distributions. By default, the normal distribution is selected when execution begins.

Because the lognormal and gamma distributions are defined only for positive-valued random variables, the distribution will be reset to the normal distribution and a message will be displayed if the first right boundary is negative or if a mean is negative when the lognormal or gamma distribution is chosen. This can happen during Option 9, or after any one of Options 1, 2, 3, 7 or 8.

**Option 10. Plot histogram.**

A high-resolution graph of the histogram of the current data will be displayed. Although the leftmost (first) and rightmost ( $m^{th}$ ) intervals are always open-ended (§2.1), on the histogram the first interval is shown as being twice the width of the second and the  $m^{th}$  is shown as being twice the width of the  $m-1^{st}$ . The first and  $m-1^{st}$  right boundaries are marked and labeled on the abscissa. MIX also looks for three boundaries in between them that are as close as possible to being equally-spaced, and marks and labels them on the abscissa. The plots done by Options 10 and 11 are numbered sequentially during the session, beginning at Plot #001.

Apple Macintosh users can copy the plot to the clipboard or save it as a MacDraw file. The graph from Option 10 shown on page 30 was produced in this manner.

IBM PC users can send the plot to a printer by pressing Y or y when the graph is displayed, although this may require additional software, as explained in §1.4. Pressing almost any other key will clear the screen and bring the next prompt for an Option number.

Mainframe users may choose to send the plot to an off-line plotter; this prompt comes before the plot is displayed on the screen. Mainframe screen graphics are in text mode.

**Option 11. Plot histogram and fitted components.**

A high-resolution graph of the histogram of the current data will be displayed. The weighted component distributions  $\pi_1 f_1(x), \dots, \pi_k f_k(x)$  and the mixture distribution  $g(x) = \pi_1 f_1(x) + \dots + \pi_k f_k(x)$  are computed from the current parameter values and superimposed on

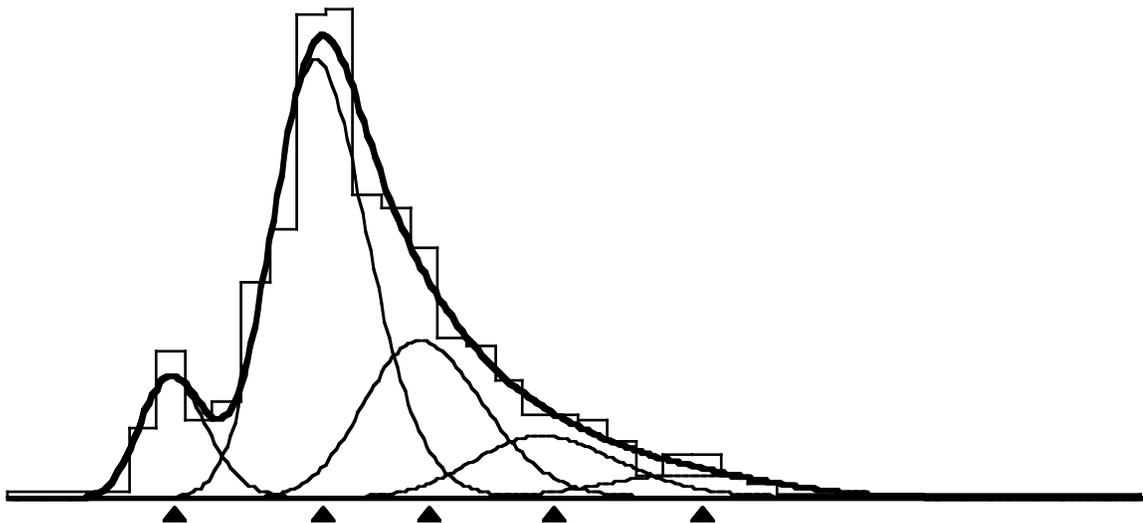
the histogram. The axes are not labeled, but the positions of the means  $\mu_1, \dots, \mu_k$  are indicated with triangles. The abscissa is scaled so that no component extends off either side of the graph. If lognormal or gamma distributions are being fitted the abscissa line begins at zero. The leftmost and rightmost grouping intervals are shown extending to their respective ends of the abscissa line. If the graph extends off the top of the screen you can have it re-drawn with reduced vertical scale.

The plots done by Options 10 and 11 are numbered sequentially during the session, beginning at Plot #001.

Apple Macintosh users can copy a plot to the clipboard or save it as a MacDraw file. The graphs in this *User's Guide* were produced in this manner. Macintosh users can also elect to create an ultra-high resolution plot with a 4 $\times$  magnification factor; this plot will appear the usual size on the screen but if it is saved on the clipboard or as a file and opened with a graphics program such as MacDraw it will be seen at its full size. It can be then be reduced when it is printed, to give publication-quality results. An example is shown in Figure 5.

IBM PC users can send the plot to a printer by pressing Y or y when the graph is displayed, although this may require additional software, as explained in §1.4. Pressing almost any other key will clear the screen and bring the next prompt for an Option number.

Mainframe users may choose to send the plot to an off-line plotter; this prompt comes before the plot is displayed on the screen. Mainframe screen graphics are in text mode.



Plot #001    Data: Heming Lake Pike 1965    Components: Lognormal

Figure 5. An example of ultra-high resolution graphics from the Apple Macintosh version. This is the same fit as shown in Figures 1 and 2.

### Option 12. Toggle to echo all I/O to I/O log.

The first time Option 12 is chosen, a file is opened to record all input and output. Plots from Options 10 and 11 are written to this file in text mode. Choosing Option 12 when the I/O file is open suspends the I/O log; choosing Option 12 when the I/O log is suspended re-opens it.

**Option -1. STOP.**

Terminate execution of MIX.

**4. STRATEGIES FOR DIFFICULT CASES****4.1 What to do when iterations will not converge**

Difficult cases arise when the components are extensively overlapped and the histogram does not show well-defined modes. The more the number of components exceeds the number of clear modes, the more difficult the data are to analyze.

If each component shows as a clear mode in the histogram, then starting values for the iterative calculations of Option 6 can easily be found by visual inspection of the histogram from Option 10, and these starting values will probably give convergence on the first attempt.

MIX uses scoring, a quasi-Newton iterative procedure, to compute the best-fitting parameter values in Options 4 and 6. Under the right conditions the iterations converge extremely quickly and standard errors of the estimates are computed in the process. If, however, the starting values and constraints are not well chosen, the iterations will soon diverge: an error message will be given and the parameters will automatically be restored to the values they had before the iterations began. For the inexperienced user, finding the right starting values and constraints to achieve convergence can be a frustrating experience if a good strategy is not adopted.

What is happening in these difficult cases is that there is a very broad range of parameter values giving more or less equally good fits to the data and there is no one set of values that is clearly a “best” fit. Option 6, attempting to find a maximum of the likelihood surface, fails because the surface is too flat. Alternative methods of calculation such as direct-search optimization (Option 5) or the EM algorithm will respond differently to this situation, wandering over the plateau of the likelihood surface for an excessively large number of iterations and eventually stopping at a point that may be rather arbitrarily chosen (Macdonald 1987). It would, of course, be more satisfying to summarize the data by defining a region of acceptable parameter values, but this is not easy to do when dealing with more than two or three parameters at a time.

The strategy recommended for MIX is to take advantage of the good features of scoring iterations while imposing enough constraints to prevent the iterations from diverging. As MIX is guided towards the solution, the constraints may be lifted gradually. In cases where all the components are not well defined in the histogram, it may not be possible to relax all of the constraints. If the constraints used for the final fit seem arbitrary, the fitting process can be repeated with an alternative choice of constraints to see how much the goodness-of-fit and the estimates depend on that choice.

Some users will routinely begin by using Option 5 to improve on the starting values of the means and standard deviations. Others, with experience, will prefer to avoid the rather long calculation time of Option 5 and begin by using Option 6, at first with lots of constraints (for example, holding all of the proportions and all of the standard deviations fixed).

If Option 6 fails and it is not clear what to do next, use Option 11 to plot a graph to see how well the starting values fit the histogram. Then, use Option 6 for diagnostic purposes, specifying a limit of 0 iterations: impose the same constraints as were imposed on the trial that failed. It will usually turn out that one or more of the parameters have exceedingly large standard errors associated with them, an indication that there is not enough information to estimate those parameters. The next step would be to try Option 6 again, holding those parameters fixed as well as imposing the constraints of the previous attempt. It may even turn out that the standard errors cannot be computed because the information matrix is singular (Macdonald and Pitcher 1979). This will happen if there is no information in the data for one or more of the parameters, an extreme case being where the user assigns a zero proportion to one component and then attempts to estimate mean and standard deviation of that component. This will also happen if the current parameter values are so far from their true values that the observed and fitted histograms bear no resemblance to each other. In either case, inspection of the plot from Option 11, inspection of the current parameter values and consideration of what the solution ought to be, should suggest a revision of the starting values and/or constraints that will be more successful on the next attempt. In the event that it is still not evident how to adjust the starting values, try Option 5.

It is always better to choose initial values for the standard deviations that are too small, rather than too large. Large standard deviations cause the components to overlap more than is necessary, obscuring the resolution of the means. It is often possible to get good estimates of the means while holding the standard deviations fixed at values slightly less than their true values.

The Example in the Appendix illustrates a strategy that will often work. In this Example we did have the advantage of knowing ahead of time that there were exactly five components present and that the coefficients of variation could be assumed to be constant. Ways to handle length-frequency distributions and other applications where the number of components is large and unknown are discussed in §5.

The main steps in the Example are as follows: the data were entered and displayed on a histogram, then starting values were given for the parameters. The starting values of the proportions did not have to be chosen carefully because Option 4 succeeded in finding good values. Option 5, restricted to about 200 iterations, was then used to improve the means while holding the standard deviations fixed. At this point, it is best to have some constraints on the standard deviations, holding most of them fixed if none of the other constraints offered seem to be applicable. Option 4 was used to revise the proportions in light of the new means and the final fit with constant coefficient of variation was then found by Option 6. It was not possible to relax the coefficient of variation constraint, but alternatives could be tried, such as equal standard deviations (Macdonald 1987) or holding some standard deviations fixed.

In this Example, the fits with normal, lognormal and gamma distributions are almost identical; the normal is the fastest to compute, so fits were first done using the normal distribution. The distribution was then switched to the gamma and the fit was adjusted by Option 6. The lognormal distribution was used for the rest of the Example.

Constraining the means to lie on a growth curve involves a major shift of the fit, so this was done by using Option 5 with all of the standard deviations held fixed before getting the final growth-curve fit with Option 6. This could also have been done by using Option 6 in two stages, first with the standard deviations and proportions fixed, then releasing the proportions and using the constant coefficient of variation constraint.

In the spirit of Cassie (1954) it has been suggested that, first, the parameters of the leftmost component be fitted while holding all others fixed; then all parameters of the two leftmost components, and so on until all have been fitted. This strategy is not recommended for MIX. It is preferable to adjust as many as possible of the components simultaneously on each step: holding means and standard deviations fixed while estimating proportions; then holding proportions fixed and constraining standard deviations while estimating means; and so on, until as many parameters as possible are estimated together.

Option 4, like Option 6, uses scoring iterations but is less likely to fail because the likelihood surface is quite well behaved when only the proportions are being estimated. If it does fail, it should be evident that very poor starting values were used or that too many components were assumed. It may, however, happen that Option 4 (or Option 6), while not actually failing, will return a negative value for a proportion. What to do if this happens is discussed in §4.2.

#### **4.2 What to do when proportions go negative or do not sum to 1**

It is possible to leave Option 2 or Option 3 with proportions that do not sum to 1. If the proportions are then revised by Option 4 or Option 6 the new values will sum to 1 and there is no problem. If they are not revised by Options 4 or 6 they should be re-scaled, either before leaving Option 2 or Option 3, or by choosing Option 3, quitting it, and accepting the offer to re-scale. If this is not done, any histograms or goodness-of-fit chi-square values will be nonsensical.

MIX will also, in some cases, tolerate “negative proportions” and Option 4 or Option 6 may give a negative estimate for a proportion. A warning is displayed when this happens.

If Option 4 or Option 6 gives a “negative proportion” it is probably an indication that you are trying to fit too many components. It is also possible that there really is a component there but the current value of its mean places it too far into one of the neighbouring components. There are then several strategies to choose from: use Option 11 to plot the current fit and see if any components are obviously misplaced; go back to Option 2 and re-enter the parameters, assuming fewer components; use Option 3 to set the offending proportion to a small positive value and hold it fixed at that value for at least the next few steps; use Option 3 to set the proportion to zero and hold it and the corresponding mean and standard deviation fixed (unless they are constrained in some other way) for at least the next few steps.

Remember that if there are, for example, 500 individuals in the sample and one component comprises 2% of the population, it will be represented by only about 10 individuals in the sample. If, furthermore, these individuals overlap with individuals from neighbouring component groups, it should be evident that there will be very little information from the mixed data to estimate anything

---

about that component. In some cases the only solution will be to say that the component is negligible and set its proportion to zero.

### 4.3 What to do when there are small expected counts

Zero and near-zero expected counts will arise in some of the grouping intervals when the initial parameter values are so inappropriate that the assumed distributions do not cover all of the data: for example, if the component standard deviations are all extremely small and the means lie nowhere near the observed histogram. In this case, the standard errors computed by Option 6 will be meaningless, because they are only valid conditionally upon the parameter values being close to their true values, and the iterations of Option 6 will generally diverge. This situation should be avoided by using Option 11 to plot a graph before you begin the fitting process, to make sure that the initial estimates are sensible.

Zero expected counts will also arise when the parameter values are appropriate but the data include a number of intervals with zero counts, at the extreme left or right of the histogram or between two well-separated peaks. This is undesirable for two reasons: the “empty” intervals increase computation time but add essentially no information to the data, and they render the chi-square goodness-of-fit test invalid by making the degrees of freedom, and hence the computed P-value, higher than is warranted. This situation can be avoided by combining with adjacent intervals any intervals that have small or zero observed counts, either while preparing the data file or, later, with Option 8.

MIX will calculate standard errors and compute the iterations in Option 6 even when there are intervals with expected counts of zero. Minimizing the likelihood-ratio chi-square is still a valid estimation procedure in these cases. However, the goodness-of-fit test will not be valid if there are too many intervals with small expected values. Most textbooks say that all expected counts must be  $> 5$ , or that a few expected counts can be as small as 2 if all others are  $> 5$ . MIX will print a warning message if more than 2 expected counts are  $< 1$ . In general, users are advised to inspect the Table of Observed and Expected Counts produced by Option 6 before they attempt to interpret the P-value computed for the chi-square test. If there are many intervals with very small expected values, or if there are intervals with zero expected values, then it might be advisable to re-group the data using Option 8 and repeat the fit by Option 6 before interpreting the goodness-of-fit test.

## 5. THE ANALYSIS OF FISHERIES LENGTH-FREQUENCY DISTRIBUTIONS

The pike data analyzed in the Appendix is an example of a length-frequency distribution. The five components correspond to groups of fish aged one to five, all older fish having been eliminated from the sample.

In many applications the data will be more difficult to analyze because an indeterminate number of age groups are present. A few fish will live much longer than most, growing more slowly as they age, so even if the fast-growing younger age-groups define clear modes at the left of the histogram, the right-hand tail will be a smear of many components with very small proportions.

One approach would be to determine the age of each of the older fish by reading rings on scales or otoliths and remove all fish beyond a certain age from the sample, as was done with the pike data.

Another approach would be to obtain samples from the older age groups, either stratifying by age and sampling lengths or stratifying by length and sampling ages. Determine the mean and standard deviation of the lengths in each age-group from these samples, and estimate only the proportions from the mixed sample. This approach is reviewed in Macdonald (1987). A future version of MIX will fit age-at-length data from length-stratified sub-samples simultaneously with the mixed length-frequency data.

If no age determination can be done, it may not be possible to estimate any parameters of the oldest age-groups or to eliminate all of the older fish from the mixed sample. You could try to use length-at-age data from another year or another location, if they are available, but the size distribution and population structure could be very different from that of the population you are trying to analyze. You could try to constrain the means to lie along a growth curve; in principle, if the first few age-groups show well-defined modes they should suffice to define the growth curve, but our own experience has been that there is too much variation in growth patterns between age-groups for this to be reliable.

If all the above suggestions fail, use Option 8 to move the boundary of the rightmost grouping interval so that most of the older fish are included in the rightmost interval. Treat all fish above a certain age as being in one component. Its mean should be set near the boundary of the rightmost grouping interval. Estimate the proportion and, if possible, the mean and standard deviation of this component, along with the parameters of the younger age groups. The mean and standard deviation will be artifacts of the grouping and hence will not have much biological significance, but the estimated proportion will be meaningful. The mean and standard deviation may be held fixed or estimated separately while imposing growth curve, linear growth, fixed coefficient of variation, constant coefficient of variation, or equal standard deviation constraints on the younger age-groups (§2.2.2, §2.2.3). It might be worthwhile to repeat the fit several times, changing the rightmost grouping interval, to see how sensitive the estimates are to this choice.

## **6. TECHNICAL SUPPORT FOR MIX**

MIX is special-purpose software intended to solve problems that are inherently difficult. Licensed users who encounter problems in applying MIX to their data may send ICHTHUS DATA SYSTEMS a disk containing a copy of the data file and a copy of the complete input/output log; we will do our best to find a solution. You may also telephone us at (416) 527-5262, 9:00 am to 9:00 pm Eastern Time.

## **7. LICENCE AGREEMENT**

We ask all users to sign a Licence Agreement, acknowledging that the Licence Fee gives them the right to run MIX on a single machine and make copies for back-up purposes only. We ask all users to respect the terms of this agreement: our capacity to improve MIX and its documentation depends on it. You may believe that you are doing your colleagues a favour by handing out copies

---

of MIX, or running MIX on several computers in your laboratory when you have only paid for a single-machine licence, but you are depriving ICHTHUS DATA SYSTEMS of revenue to which we are legally entitled and thereby impairing our ability to develop new software. A copy of the standard Licence Agreement is shown on page 60.

## 8. UPGRADES

Each time a new Release is announced, licensed users will be offered the upgrade for a nominal charge. Any licensed user who suggests a worthwhile improvement to MIX will be sent the next upgrade free of charge.

Any licensed user who succeeds in “crashing” MIX, so that control is involuntarily returned from MIX to the operating system, should send us details of the computer and operating system being used and a disk containing a copy of the data file and the complete input/output log for that session. In return, the user will receive the next upgrade free of charge.

## REFERENCES

- Cassie R. M. (1954). Some uses of probability paper in the analysis of size frequency distributions. *Australian Journal of Marine and Freshwater Research* **5**, 513-522.
- Everitt, B.S. and D.J. Hand (1981). *Finite Mixture Distributions*. Chapman and Hall, London. xi+143 pp.
- Macdonald, P.D.M. and T.J. Pitcher (1979). Age-groups from size-frequency data: a versatile and efficient method of analysing distribution mixtures. *Journal of the Fisheries Research Board of Canada* **36**, 987-1001.
- Macdonald, P.D.M. (1987). Analysis of length-frequency distributions. In R.C. Summerfelt and G.E. Hall [editors], *Age and Growth of Fish*, Iowa State University Press, Ames, Iowa. pp 371-384.
- McLachlan, G.J. and K.E. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York. xi+253 pp.
- O'Neill, R. (1971). Algorithm AS 47. Function minimization using a simplex procedure. *Applied Statistics* **20**, 338-345.
- Rao, C.R. (1965). *Linear statistical inference and its applications*. Wiley, New York. xviii+522 pp.
- Schnute, J., and D. Fournier (1980). A new approach to length-frequency analysis: growth structure. *Canadian Journal of Fisheries and Aquatic Sciences* **37**, 1337-1351.
- Titterton, D.M., A.F.M. Smith and U.E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York. x+243 pp.

**APPENDIX****Example: An analysis of Heming Lake pike data**

The data are described in Macdonald and Pitcher (1979). The mixture was known to consist of exactly five components. The five components correspond to the five age-groups present in the sample, all fish more than five years old having been removed from the sample. Results of other analyses of the same data are given in Macdonald (1987).

The following pages show the input/output log of an interactive session. Data entered by the user are shown in bold type. Explanatory remarks have been added in bold script, either in boxes or on the right-hand side of the page. All else is output from MIX.

---

---

MACDONALD &amp; PITCHER MIXTURE ANALYSIS

Reference: J. Fish. Res. Board Can. 36:987-1001

Program MIX copyright © 1985, 1986, 1987, 1988 by ICHTHUS DATA SYSTEMS.  
Release 2.3, January 1988.

Do you want to see a list of Options (Y/N) ?

**Y**

## LIST OF OPTIONS

0. List of options.
  1. Read a new set of data.
  2. Read a full set of parameter values.
  3. Revise specified parameter values.
  4. Estimate proportions for fixed means, sigmas.
  5. Estimate means, sigmas for fixed proportions by constrained search.
  6. Estimate proportions, means, sigmas with or without constraints and/or give diagnostic displays.
  7. Restore parameters to values from previous step.
  8. Regroup data or restore to original grouping.
  9. Choose a distribution.
  10. Plot histogram.
  11. Plot histogram and fitted components.
  12. Toggle to echo all I/O to I/O log.
- 1. STOP.

Option number? [0 for list of options, -1 to STOP]

**12***Open a disk file to keep a record of this session.*

OPENING FILE FOR I/O LOG

---

Enter file name, in single quotes:

'PIKE65.LOG'

Creating I/O file: PIKE65.LOG

Option number? [0 for list of options, -1 to STOP]

1

*Reading the 1965 Heming Lake Pike data from the keyboard. If the file PIKE65 is available, respond N to the next prompt to read the data from the file instead of from the keyboard; you will be prompted for the file name.*

READ A NEW SET OF DATA

Do you want to enter data from keyboard (Y/N) ?

Y

Enter title (1-25 characters):

**Heming Lake Pike 1965**

Enter the number of intervals

NOTE: Must be at least 2, at most 80:

25

Enter count and right boundary 24 times:

4 19.75

10 21.75

21 23.75

11 25.75

14 25.75

31 29.75

39 31.75

70 33.75

71 35.75

44 37.75

42 39.75

36 41.75

23 43.75

22 45.75

17 47.75

12 49.75

12 51.75

11 53.75

8 55.75

3 57.75

6 59.75

6 61.75

3 63.75

2 65.75

*Two errors will be put in deliberately, so that editing can be demonstrated later.*

*< This right boundary should have been 27.75.*

Enter count:

2

*< This count should have been 5.*

INTERVAL	OBSERVED	COUNT	RIGHT BOUNDARY	Heming Lake Pike 1965
----------	----------	-------	----------------	-----------------------

---

MIX 2.3

---

1	4.0000	19.7500
2	10.0000	21.7500
3	21.0000	23.7500
4	11.0000	25.7500
5	14.0000	25.7500
6	31.0000	29.7500
7	39.0000	31.7500
8	70.0000	33.7500
9	71.0000	35.7500
10	44.0000	37.7500
11	42.0000	39.7500
12	36.0000	41.7500
13	23.0000	43.7500
14	22.0000	45.7500
15	17.0000	47.7500
16	12.0000	49.7500
17	12.0000	51.7500
18	11.0000	53.7500
19	8.0000	55.7500
20	3.0000	57.7500
21	6.0000	59.7500
22	6.0000	61.7500
23	3.0000	63.7500
24	2.0000	65.7500
25	2.0000	

Any errors to correct (Y/N) ?

**Y**

Which interval is incorrect?

**25**

Enter correct count:

**5**

Any errors to correct (Y/N) ?

**N**

*Let MIX detect the incorrect boundary.*

\* ERROR \* INTERVAL BOUNDARIES OUT OF ORDER

INTERVAL	OBSERVED COUNT	RIGHT BOUNDARY	Heming Lake Pike 1965
3	21.0000	23.7500	
4	11.0000	25.7500	
5	14.0000	25.7500	

Do you want to abandon this data set (Y/N) ?

**N**

Which interval is incorrect?

**5**

Enter correct count and right boundary:

**14 27.75**

Any errors to correct (Y/N) ?

**N**

Do you want to store these data on a file (Y/N) ?

Y

Enter file name, in single quotes:

'PIKE65'

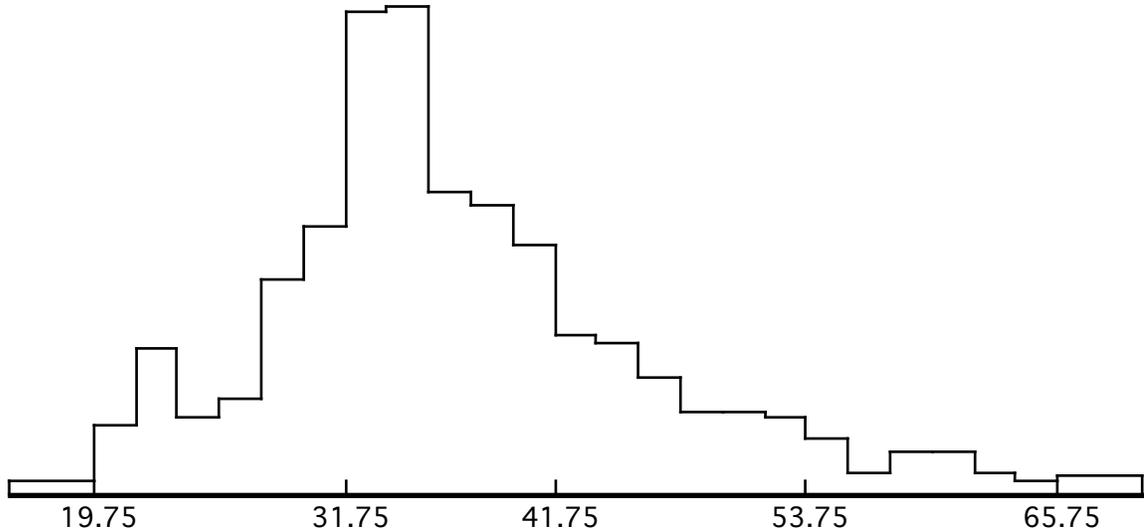
Writing to file: PIKE65

Do you want to display the data again (Y/N) ?

N

Option number? [0 for list of options, -1 to STOP]

10



Plot #001 Data: Heming Lake Pi ke 1965

Option number? [0 for list of options, -1 to STOP]

2

*Read in starting values for all the parameters. Starting values for the means and sigmas should be as good as possible; they can be found by inspecting the histogram from Option 10, above, and from knowing something about the population being studied. Starting values for the proportions are less critical, since they can usually be improved, very efficiently, by Option 4.*

READ A FULL SET OF PARAMETER VALUES

How many components? [Must be at least 1, at most 15]

5

Enter the 5 proportions:

1 1 1 1 1

*The proportions need not sum to 1, since they can be re-scaled.*

Enter the 5 means:

20 30 40 50 60

MIX 2.3

---

Enter the 5 sigmas:

2 3 4 5 6

Proportions do not sum to 1. Do you want to re-scale (Y/N) ?

Y

Proportions

.20000 .20000 .20000 .20000 .20000

Means

20.0000 30.0000 40.0000 50.0000 60.0000

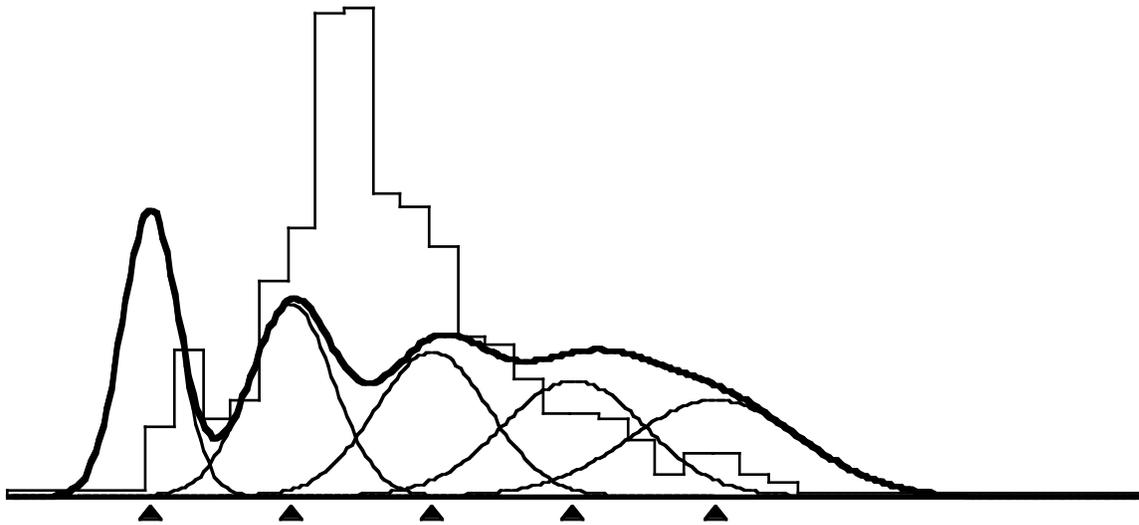
Sigmas

2.0000 3.0000 4.0000 5.0000 6.0000

Option number? [0 for list of options, -1 to STOP]

11

*Plot a graph now, to check that the means and sigmas are reasonable.  
Remember that the proportions have not yet been fitted.*



Plot #002 Data: Heming Lake Pi ke 1965 Components: Normal

Option number? [0 for list of options, -1 to STOP]

4

*Now improve on the starting values for the proportions.*

ESTIMATE PROPORTIONS FOR FIXED MEANS, SIGMAS

Distribution selected is Normal

Enter iteration limit:

20

Number of iterations = 9

Fitting Normal components

---

Proportions and their standard errors  
 .05057 .36771 .43737 .09155 .05279  
 .01024 .02413 .02743 .02083 .01313

Means (ALL HELD FIXED)  
 20.0000 30.0000 40.0000 50.0000 60.0000

Sigmas (ALL HELD FIXED)  
 2.0000 3.0000 4.0000 5.0000 6.0000

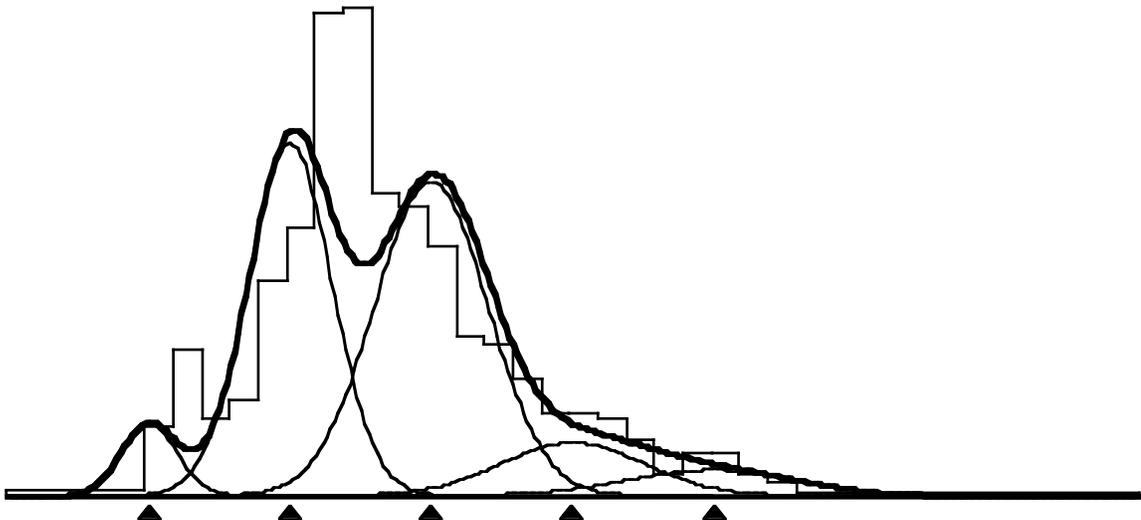
Degrees of freedom = 20

Chi-squared = 104.174

Option number? [0 for list of options, -1 to STOP]

11

*Plot a graph now, to see how well we have done. The means are not quite right, but the proportions and sigmas look good.*



Plot #003 Data: Heming Lake Pi ke 1965 Components: Normal

Option number? [0 for list of options, -1 to STOP]

6

*We will attempt to fit all proportions, all means, and a constant coefficient of variation, using Option 6. The attempt fails because we are trying to estimate too many parameters at once when the initial values of the parameters are too far from the true fit. This should have been evident from Plot #003. In fact, if we had looked carefully at the plots and then used Option 3 to revise the first three means to 23, 33, and 43, respectively, Option 6 would converge now.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
 AND/OR GIVE DIAGNOSTIC DISPLAYS

## MIX 2.3

---

Distribution selected is Normal

Enter iteration limit [0 gives displays with current parameters]:  
**30**

Display observed and expected counts as a table (Y/N) ?  
**N**

Display observed and expected counts as a graph (Y/N) ?  
**N**

Display variance-covariance matrix (Y/N) ?  
**N**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:  
**0**

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),  
2 (MEANS EQUAL), 3 (EQUALLY SPACED),  
4 (GROWTH CURVE).

Enter choice:  
**0**

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),  
2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),  
4 (SIGMAS EQUAL).

Enter choice:  
**3**

Is Kth sigma different (Y/N) ?  
**N**

Do you want to abort (Y/N) ?  
**N**

PARAMETERS OUT OF RANGE AFTER 5 ITERATIONS

RESTORE PARAMETERS TO VALUES FROM PREVIOUS STEP

Proportions, Means, Sigmas:

.05057	.36771	.43737	.09155	.05279
20.0000	30.0000	40.0000	50.0000	60.0000
2.0000	3.0000	4.0000	5.0000	6.0000

Option number? [0 for list of options, -1 to STOP]  
**5**

*We must improve the means now. We could inspect the plots, get better values "by eye", and revise the means using Option 3. We could try Option 6 with all proportions and all sigmas held fixed. Here we demonstrate the slowest but safest method, using Option 5 to adjust the means by direct search while holding the proportions fixed. We hold all the sigmas fixed, but we could have fitted a constant coefficient of variation and obtained equally good results.*

ESTIMATE SPECIFIED MEANS, SIGMAS FOR FIXED PROPORTIONS

Distribution selected is Normal

Constraints on means:

0 (NONE),                    1 (SPECIFIED MEANS FIXED),  
 2 (MEANS EQUAL),         3 (EQUALLY SPACED),  
 4 (GROWTH CURVE).

Enter choice:

**0**

Constraints on sigmas:

0 (NONE),                    1 (SPECIFIED SIGMAS FIXED),  
 2 (FIXED COEF. OF VARIATION),    3 (CONSTANT COEF. OF VARIATION),  
 4 (SIGMAS EQUAL).

Enter choice:

**1**

Which of the 5 sigmas are fixed?

For each, in order, enter Y (FIXED) or N (FREE):

**YYYYY**

Enter lower and upper limits for mean:

**19 65**

Enter lower and upper limits for sigma:

**1 10**

Enter initial steps for the 5 means:

**1 1 1 1 1**

Enter initial steps for the 5 sigmas:

**.5 .5 .5 .5 .5**

Enter iteration limit:

**200**

Enter convergence check frequency:

**10**

Enter accuracy index:

**1**

Do you want to abort (Y/N) ?

**N**

Number of function evaluations = 222

MIX 2.3

---

Number of restarts = 8  
Required standard deviation of vertex values = 0.100E+01

CONVERGENCE CRITERION NOT SATISFIED

Fitting Normal components

Proportions (ALL HELD FIXED)  
.05057 .36771 .43737 .09155 .05279

Means  
22.3050 31.8944 38.8973 49.7120 59.1185

Sigmas  
2.0000 3.0000 4.0000 5.0000 6.0000  
FIXED FIXED FIXED FIXED FIXED

Degrees of freedom = 19

Chi-squared = 29.8516

Option number? [0 for list of options, -1 to STOP]

4

*Now revise the proportions again, to adjust for the new means. Note that the iterations of Option 4 are very fast and they usually converge. Here, the fit is substantially improved by Option 4.*

ESTIMATE PROPORTIONS FOR FIXED MEANS, SIGMAS

Distribution selected is Normal

Enter iteration limit:

20

Number of iterations = 6

Fitting Normal components

Proportions and their standard errors  
.08376 .37744 .36832 .11748 .05299  
.01292 .02833 .03091 .02159 .01383

Means (ALL HELD FIXED)  
22.3050 31.8944 38.8973 49.7120 59.1185

Sigmas (ALL HELD FIXED)  
2.0000 3.0000 4.0000 5.0000 6.0000

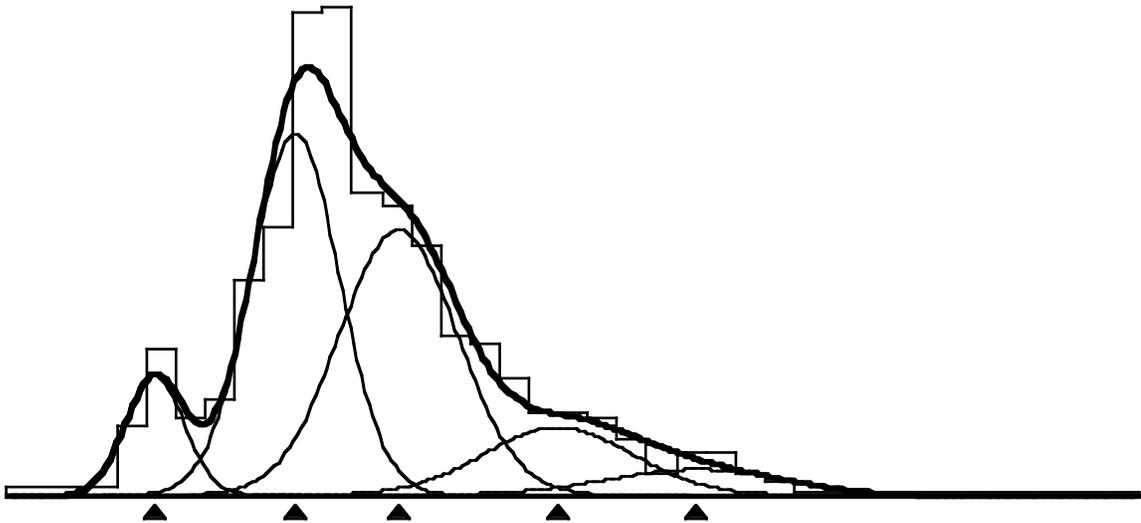
Degrees of freedom = 20

Chi-squared = 16.9645

Option number? [0 for list of options, -1 to STOP]

11

*Even though we have not started to fit the sigmas, the fit is looking very good.*



Plot #004 Data: Heming Lake Pi ke 1965 Components: Normal

Option number? [0 for list of options, -1 to STOP]

6

*We now go for the final fit in this sequence, using Option 6 with all proportions free, all means free, and a constant coefficient of variation.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Normal

Enter iteration limit [0 gives displays with current parameters]:

30

Display observed and expected counts as a table (Y/N) ?

Y

Display observed and expected counts as a graph (Y/N) ?

Y

Display variance-covariance matrix (Y/N) ?

Y

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

0

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

MIX 2.3

---

Enter choice:

0

Constraints on sigmas:

- 0 (NONE),
- 1 (SPECIFIED SIGMAS FIXED),
- 2 (FIXED COEF. OF VARIATION),
- 3 (CONSTANT COEF. OF VARIATION),
- 4 (SIGMAS EQUAL).

Enter choice:

3

Is Kth sigma different (Y/N) ?

N

Do you want to abort (Y/N) ?

N

Number of iterations = 11

INTERVAL	EXPECTED COUNT	OBSERVED COUNT	RIGHT BOUNDARY
1	4.0295	4.0000	19.7500
2	11.5477	10.0000	21.7500
3	17.4449	21.0000	23.7500
4	13.7595	11.0000	25.7500
5	12.8211	14.0000	27.7500
6	26.5010	31.0000	29.7500
7	49.7208	39.0000	31.7500
8	66.0656	70.0000	33.7500
9	64.2171	71.0000	35.7500
10	51.3314	44.0000	37.7500
11	40.2333	42.0000	39.7500
12	33.1477	36.0000	41.7500
13	26.6248	23.0000	43.7500
14	20.4391	22.0000	45.7500
15	16.2169	17.0000	47.7500
16	13.8090	12.0000	49.7500
17	11.8220	12.0000	51.7500
18	9.5720	11.0000	53.7500
19	7.4102	8.0000	55.7500
20	5.8451	3.0000	57.7500
21	4.9222	6.0000	59.7500
22	4.3083	6.0000	61.7500
23	3.6669	3.0000	63.7500
24	2.8831	2.0000	65.7500
25	4.6610	5.0000	

```

19.75 EEX
21.75 EEEEEEOEE
23.75 EEEEEEEEEEEEEEE O
25.75 EEEEEEOEEE
27.75 EEEEEEEEEEO
29.75 EEEEEEEEEEEEEEEEEEE O
31.75 EEEEEEEEEEEEEEEEEEEEOEEEEEEEE
33.75 EEEEEEEEEEEEEEEEEEEEOEEEEEEEEEE O
35.75 EEEEEEEEEEEEEEEEEEEEOEEEEEEEEEE O
37.75 EEEEEEEEEEEEEEEEEEEEOEEEEEE

```

```

39.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
41.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE O
43.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
45.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE O
47.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE X
49.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE OEE
51.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE X
53.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE O
55.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE O
57.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE OEE
59.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE O
61.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE O
63.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EX
65.75 EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE EX
      EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE

```

*The variance-covariance matrix is seldom useful, but it can be displayed by Option 6 if required. Here, the 10 rows (and, similarly, the 10 columns) correspond to the 10 free parameters:  $\sigma_1, \sigma_2, \sigma_3, \sigma_4; \mu_1, \mu_2, \mu_3, \mu_4, \mu_5; \sigma_1$ .*

```

Variance-covariance matrix for parameter estimates;
pi( 5) and all fixed parameters are excluded
 0.2069E-03
-0.4251E-04 0.5013E-02
-0.9522E-04-0.2793E-02 0.3022E-02
-0.5008E-04-0.1763E-02 0.6488E-04 0.1593E-02
 0.1840E-02 0.2142E-02-0.2438E-02-0.1199E-02 0.1864E+00
 0.1103E-02 0.3452E-01-0.2208E-01-0.1128E-01 0.4927E-01 0.3388E+00
 0.2002E-02 0.1009E+00-0.3493E-01-0.5324E-01 0.8489E-01 0.7211E+00 0.2872E+01
>>
 0.1833E-02 0.7862E-01 0.1913E-01-0.6068E-01 0.6932E-01 0.5055E+00 0.2923E+01
>> 0.5895E+01
 0.1278E-02 0.2249E-01 0.2166E-01-0.8162E-02 0.3875E-01 0.1136E+00 0.1140E+01
>> 0.3486E+01 0.5131E+01
-0.1277E-03 0.5636E-02-0.4071E-02-0.1417E-02 0.1156E-01 0.5456E-01 0.8525E-01
>> 0.2467E-01-0.4775E-01 0.3276E-01

```

Fitting Normal components

```

Proportions and their standard errors
 .09200   .46711   .25858   .12793   .05438
 .01439   .07080   .05497   .03991   .02184

```

```

Means and their standard errors
 22.7487  32.9675  39.7800  48.6285  60.1262
  .4317   .5821   1.6946   2.4280   2.2651

```

```

Sigmas (CONSTANT COEF. OF VAR. = .0955) and standard error
 2.1718   3.1474   3.7978   4.6425   5.7402
 .1810

```

Degrees of freedom = 14

Chi-squared = 11.2852 (P = .6635)

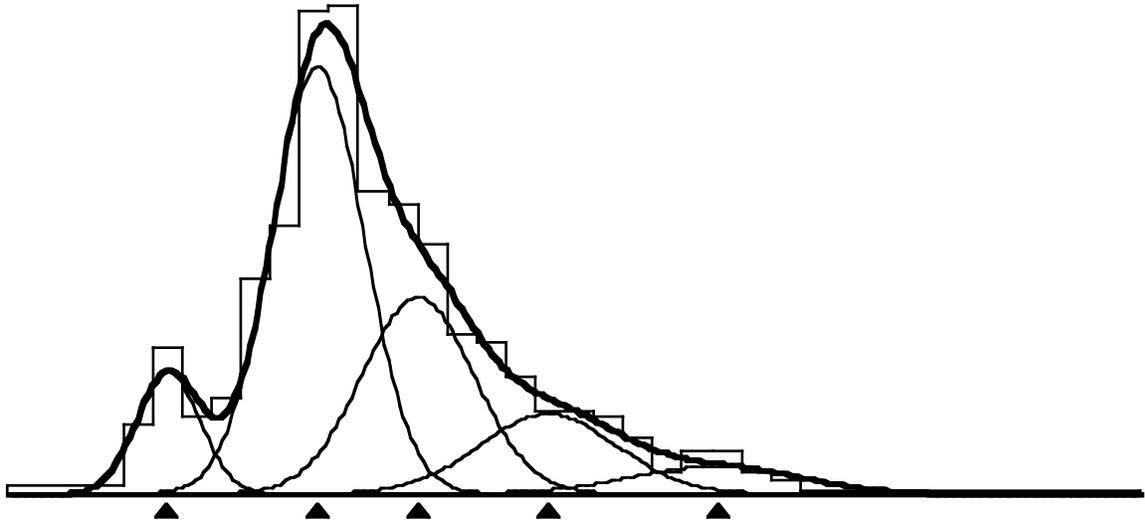
MIX 2.3

---

Option number? [0 for list of options, -1 to STOP]

11

*The chi-square test above, and the plot below, both indicate an excellent fit. We will not get a significantly better fit to these data.*



Plot #005 Data: Heming Lake Pi ke 1965 Components: Normal

Option number? [0 for list of options, -1 to STOP]

9

*Change from fitting mixtures of Normal distributions to fitting mixtures of Gamma distributions. Fitting Gamma distributions can take a lot of computer time.*

SELECT A DISTRIBUTION

Enter 1 (Normal), 2 (Lognormal), or 3 (Gamma):

3

Distribution selected is Gamma

Option number? [0 for list of options, -1 to STOP]

6

*Repeat the previous fit, assuming Gamma distributions for the components.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Gamma

Enter iteration limit [0 gives displays with current parameters]:

30

Display observed and expected counts as a table (Y/N) ?

---

**Y**

Display observed and expected counts as a graph (Y/N) ?

**N**

Display variance-covariance matrix (Y/N) ?

**Y**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

**0**

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

Enter choice:

**0**

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),

2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),

4 (SIGMAS EQUAL).

Enter choice:

**3**

Is Kth sigma different (Y/N) ?

**N**

Do you want to abort (Y/N) ?

**N**

Number of iterations = 17

INTERVAL	EXPECTED COUNT	OBSERVED COUNT	RIGHT BOUNDARY
1	3.7829	4.0000	19.7500
2	11.8123	10.0000	21.7500
3	17.4242	21.0000	23.7500
4	13.7666	11.0000	25.7500
5	12.9711	14.0000	27.7500
6	26.6764	31.0000	29.7500
7	49.8157	39.0000	31.7500
8	65.4037	70.0000	33.7500
9	63.8074	71.0000	35.7500
10	52.0518	44.0000	37.7500
11	40.7189	42.0000	39.7500
12	32.6402	36.0000	41.7500
13	26.1791	23.0000	43.7500
14	20.6984	22.0000	45.7500
15	16.6000	17.0000	47.7500
16	13.7440	12.0000	49.7500
17	11.4795	12.0000	51.7500
18	9.4116	11.0000	53.7500
19	7.5849	8.0000	55.7500
20	6.1474	3.0000	57.7500
21	5.0830	6.0000	59.7500
22	4.2318	6.0000	61.7500

MIX 2.3

23	3.4404	3.0000	63.7500
24	2.6553	2.0000	65.7500
25	4.8736	5.0000	

Variance-covariance matrix for parameter estimates;  
 pi( 5) and all fixed parameters are excluded

```

0.2182E-03
0.7915E-04 0.9921E-02
-0.1501E-03-0.4258E-02 0.4598E-02
-0.1111E-03-0.4480E-02 0.3267E-03 0.3545E-02
0.1978E-02 0.9162E-02-0.5659E-02-0.4348E-02 0.1957E+00
0.2214E-02 0.7246E-01-0.3583E-01-0.3096E-01 0.1089E+00 0.6440E+00
0.5781E-02 0.2607E+00-0.6190E-01-0.1527E+00 0.2766E+00 0.1898E+01 0.8454E+01
>>
0.5517E-02 0.2552E+00 0.5088E-01-0.1892E+00 0.2449E+00 0.1715E+01 0.1009E+02
>> 0.1915E+02
0.2286E-02 0.8024E-01 0.6137E-01-0.4548E-01 0.7506E-01 0.4586E+00 0.3987E+01
>> 0.1093E+02 0.1176E+02
0.2113E-03 0.1467E-01-0.8373E-02-0.5545E-02 0.3175E-01 0.1329E+00 0.3358E+00
>> 0.2464E+00-0.7937E-02 0.5278E-01
  
```

Fitting Gamma components

Proportions and their standard errors

.09684	.49464	.24084	.11729	.05039
.01477	.09961	.06781	.05954	.03311

Means and their standard errors

22.9483	33.3271	40.4566	49.2933	60.3671
.4424	.8025	2.9075	4.3763	3.4292

Sigmas (CONSTANT COEF. OF VAR. = .0993) and standard error

2.2777	3.3079	4.0155	4.8926	5.9917
.2297				

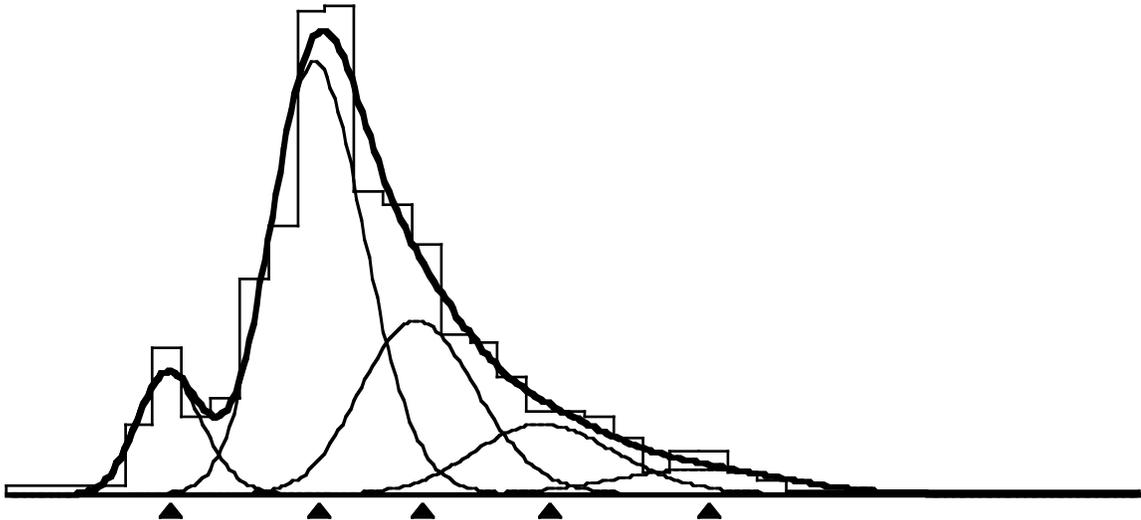
Degrees of freedom = 14

Chi-squared = 11.7257 (P = .6283)

Option number? [0 for list of options, -1 to STOP]

11

*Because the components in the Pike data have relatively small coefficients of variation, the Gamma fit is not very different from the Normal fit. The fitted components have slight positive skewness.*



Plot #006 Data: Heming Lake Pi ke 1965 Components: Gamma

Option number? [0 for list of options, -1 to STOP]

9

*Change from fitting mixtures of Gamma distributions to fitting mixtures of Lognormal distributions. Fitting Lognormal distributions is about as fast as fitting Normal distributions. Lognormal distributions are positively-skewed bell-shaped curves; Lognormals cannot take a wide range of shapes like Gamma distributions, but will work as well as Gammas in an application like this.*

SELECT A DISTRIBUTION

Enter 1 (Normal), 2 (Lognormal), or 3 (Gamma):

2

Distribution selected is Lognormal

Option number? [0 for list of options, -1 to STOP]

6

*Repeat the previous fit, assuming Lognormal distributions for the components.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

50

Display observed and expected counts as a table (Y/N) ?

Y

## MIX 2.3

---

Display observed and expected counts as a graph (Y/N) ?

**N**

Display variance-covariance matrix (Y/N) ?

**Y**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

**0**

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

Enter choice:

**0**

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),

2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),

4 (SIGMAS EQUAL).

Enter choice:

**3**

Is Kth sigma different (Y/N) ?

**N**

Do you want to abort (Y/N) ?

**N**

Number of iterations = 36

INTERVAL	EXPECTED COUNT	OBSERVED COUNT	RIGHT BOUNDARY
1	3.7216	4.0000	19.7500
2	11.9134	10.0000	21.7500
3	17.3461	21.0000	23.7500
4	13.7970	11.0000	25.7500
5	13.0945	14.0000	27.7500
6	26.7151	31.0000	29.7500
7	49.7509	39.0000	31.7500
8	65.1425	70.0000	33.7500
9	63.7451	71.0000	35.7500
10	52.3416	44.0000	37.7500
11	40.8507	42.0000	39.7500
12	32.4418	36.0000	41.7500
13	26.0357	23.0000	43.7500
14	20.8043	22.0000	45.7500
15	16.7411	17.0000	47.7500
16	13.7185	12.0000	49.7500
17	11.3487	12.0000	51.7500
18	9.3475	11.0000	53.7500
19	7.6530	8.0000	55.7500
20	6.2733	3.0000	57.7500
21	5.1579	6.0000	59.7500
22	4.2075	6.0000	61.7500
23	3.3436	3.0000	63.7500
24	2.5456	2.0000	65.7500

---

---

25                    4.9630                    5.0000

Variance-covariance matrix for parameter estimates;  
 pi( 5) and all fixed parameters are excluded

```

0.2305E-03
0.2229E-03 0.1451E-01
-0.1720E-03-0.4170E-02 0.5892E-02
-0.2099E-03-0.7950E-02-0.1273E-03 0.6533E-02
0.2305E-02 0.1767E-01-0.7453E-02-0.9688E-02 0.2160E+00
0.3509E-02 0.1086E+00-0.3795E-01-0.5707E-01 0.1810E+00 0.9370E+00
0.1182E-01 0.4658E+00-0.4134E-01-0.3117E+00 0.5967E+00 0.3447E+01 0.1765E+02
>>
0.1321E-01 0.5433E+00 0.1744E+00-0.4279E+00 0.6286E+00 0.3759E+01 0.2441E+02
>> 0.4888E+02
0.4617E-02 0.1718E+00 0.1524E+00-0.1129E+00 0.1721E+00 0.1043E+01 0.9481E+01
>> 0.2649E+02 0.2358E+02
0.5770E-03 0.2442E-01-0.1051E-01-0.1178E-01 0.5445E-01 0.2148E+00 0.7053E+00
>> 0.6700E+00 0.7747E-01 0.7617E-01

```

Fitting Lognormal components

Proportions and their standard errors

```

.09967    .51889    .22677    .10710    .04757
.01518    .12045    .07676    .08083    .04847

```

Means and their standard errors

```

23.0735    33.6069    41.1029    49.8826    60.4670
.4647        .9680        4.2013        6.9918        4.8562

```

Sigmas (CONSTANT COEF. OF VAR. = .1028) and standard error

```

2.3722    3.4551    4.2258    5.1284    6.2166
.2760

```

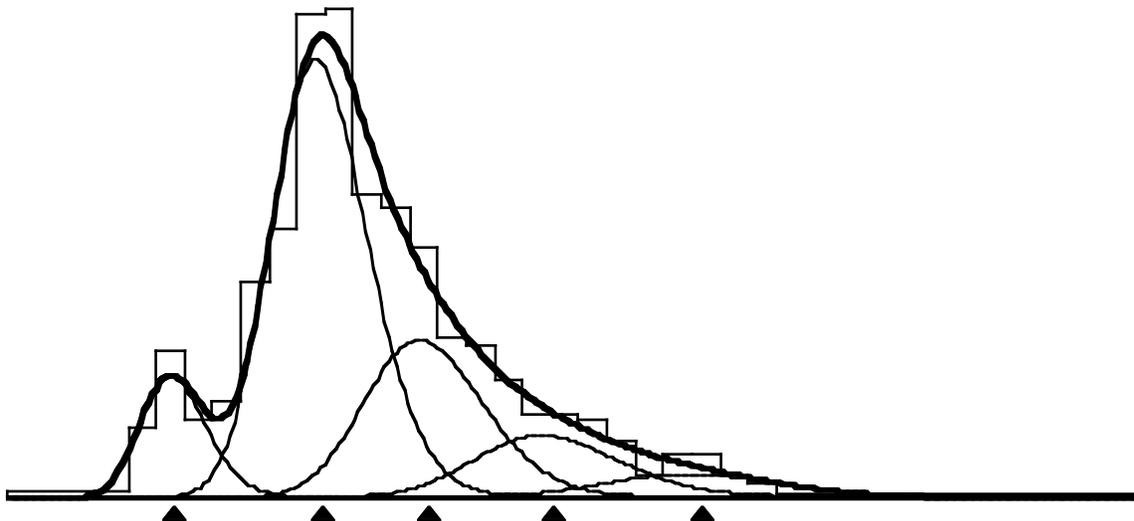
Degrees of freedom = 14

Chi-squared = 11.9477                    (P = .6105)

Option number? [0 for list of options, -1 to STOP]

11

*Because the components in the Pike data have relatively small coefficients of variation, the Lognormal fit is not very different from the Normal or Gamma fit. The fitted components have slight positive skewness.*



Plot #007 Data: Heming Lake Pi ke 1965 Components: Lognormal

Option number? [0 for list of options, -1 to STOP]

6

*Repeat the previous fit, but force the means to lie along a von Bertalanffy growth curve. Option 6 fails here. Option 6 will not usually succeed in fitting a growth curve unless the means are already very close to the best fit, especially when proportions and sigmas are also being estimated.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

50

Display observed and expected counts as a table (Y/N) ?

N

Display observed and expected counts as a graph (Y/N) ?

N

Display variance-covariance matrix (Y/N) ?

N

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

0

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

---

Enter choice:

4

Is Kth mean different (Y/N) ?

N

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),  
 2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),  
 4 (SIGMAS EQUAL).

Enter choice:

3

Is Kth sigma different (Y/N) ?

N

Do you want to abort (Y/N) ?

N

PARAMETERS OUT OF RANGE AFTER 1 ITERATIONS

RESTORE PARAMETERS TO VALUES FROM PREVIOUS STEP

Proportions, Means, Sigmas:

.09967	.51889	.22677	.10710	.04757
23.0735	33.6069	41.1029	49.8826	60.4670
2.3722	3.4551	4.2258	5.1284	6.2166

Option number? [0 for list of options, -1 to STOP]

5

*Force the means to lie on a growth curve, using direct-search optimization to adjust the means while the proportions and sigmas are all held fixed. Option 6 will also work here if all the proportions and sigmas are held fixed.*

ESTIMATE SPECIFIED MEANS, SIGMAS FOR FIXED PROPORTIONS

Distribution selected is Lognormal

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),  
 2 (MEANS EQUAL), 3 (EQUALLY SPACED),  
 4 (GROWTH CURVE).

Enter choice:

4

Is Kth mean different (Y/N) ?

N

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),  
 2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),  
 4 (SIGMAS EQUAL).

Enter choice:

1

---

MIX 2.3

---

Which of the 5 sigmas are fixed?  
For each, in order, enter Y (FIXED) or N (FREE):  
**YYYYY**

Enter lower and upper limits for mean:  
**20 65**

Enter lower and upper limits for sigma:  
**1 7**

Enter initial steps for the first three means:  
**1 1 1**

Enter initial steps for the 5 sigmas:  
**.5 .5 .5 .5 .5**

Enter iteration limit:  
**100**

Enter convergence check frequency:  
**10**

Enter accuracy index:  
**1**

Do you want to abort (Y/N) ?  
**N**

Number of function evaluations = 120  
Number of restarts = 4  
Required standard deviation of vertex values = 0.100E+01

CONVERGENCE CRITERION NOT SATISFIED

Fitting Lognormal components

Proportions (ALL HELD FIXED)  
.09967 .51889 .22677 .10710 .04757

Means  
(Linf = 125.882; t1-t0 = 1.9536; k = .104495)  
23.2442 33.4280 42.6014 50.8645 58.3078  
GCRV GCRV

Sigmas  
2.3722 3.4551 4.2258 5.1284 6.2166  
FIXED FIXED FIXED FIXED FIXED

Degrees of freedom = 21

Chi-squared = 19.6888

Option number? [0 for list of options, -1 to STOP]  
**4**

*Revise the proportions, to adjust for the new means.*

## ESTIMATE PROPORTIONS FOR FIXED MEANS, SIGMAS

Distribution selected is Lognormal

Enter iteration limit:

**20**

Number of iterations = 6

Fitting Lognormal components

Proportions and their standard errors

.09880	.53898	.24467	.05141	.06613
.01417	.02688	.03046	.02742	.01785

Means (ALL HELD FIXED)

23.2442	33.4280	42.6014	50.8645	58.3078
---------	---------	---------	---------	---------

Sigmas (ALL HELD FIXED)

2.3722	3.4551	4.2258	5.1284	6.2166
--------	--------	--------	--------	--------

Degrees of freedom = 20

Chi-squared = 14.9367

Option number? [0 for list of options, -1 to STOP]

**6**

*Use Option 6 to find the best fit with the means on a growth curve  
and a constant coefficient of variation, while holding all  
proportions fixed.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

**30**

Display observed and expected counts as a table (Y/N) ?

**N**

Display observed and expected counts as a graph (Y/N) ?

**N**

Display variance-covariance matrix (Y/N) ?

**N**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

**1**

Which of the 5 proportions are fixed?

For each, in order, enter Y (FIXED) or N (FREE):

**YYYYY**

Constraints on means:

0 (NONE),                    1 (SPECIFIED MEANS FIXED),  
2 (MEANS EQUAL),         3 (EQUALLY SPACED),  
4 (GROWTH CURVE).

Enter choice:

4

Is Kth mean different (Y/N) ?

N

Constraints on sigmas:

0 (NONE),                    1 (SPECIFIED SIGMAS FIXED),  
2 (FIXED COEF. OF VARIATION),    3 (CONSTANT COEF. OF VARIATION),  
4 (SIGMAS EQUAL).

Enter choice:

3

Is Kth sigma different (Y/N) ?

N

Do you want to abort (Y/N) ?

N

Number of iterations =     7

Fitting Lognormal components

Proportions and their standard errors

.09880	.53898	.24467	.05141	.06613
FIXED	FIXED	FIXED	FIXED	FIXED

Means (ON A GROWTH CURVE) and standard errors

(Linf =	109.042;	t1-t0 =	1.8226;	k =	.131107)
(s.e.:	22.213		.1690		.039767)
23.1774	33.7281	42.9824	51.0996	58.2194	
.4285	.3083	.4531			

Sigmas (CONSTANT COEF. OF VAR. = .1052) and standard error

2.4374	3.5469	4.5201	5.3737	6.1225
.1676				

Degrees of freedom = 20

Chi-squared = 13.6676                    (P = .8469)

Option number? [0 for list of options, -1 to STOP]

6

<p><i>Use Option 6 to go for the final fit with proportions all free, means on a growth curve, and constant coefficient of variation.</i></p>
---

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

---

30

Display observed and expected counts as a table (Y/N) ?

N

Display observed and expected counts as a graph (Y/N) ?

N

Display variance-covariance matrix (Y/N) ?

N

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

0

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),  
2 (MEANS EQUAL), 3 (EQUALLY SPACED),  
4 (GROWTH CURVE).

Enter choice:

4

Is Kth mean different (Y/N) ?

N

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),  
2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),  
4 (SIGMAS EQUAL).

Enter choice:

3

Is Kth sigma different (Y/N) ?

N

Do you want to abort (Y/N) ?

N

Number of iterations = 9

Fitting Lognormal components

Proportions and their standard errors

.10137	.58091	.21581	.04628	.05564
.01550	.04290	.04370	.04322	.04153

Means (ON A GROWTH CURVE) and standard errors

(Linf = 106.840; t1-t0 = 1.7469; k = .140071)  
(s.e.: 41.544 .2260 .076754)  
23.1899 34.1232 43.6276 51.8897 59.0719  
.4748 .4681 .9874

Sigmas (CONSTANT COEF. OF VAR. = .1091) and standard error

2.5300	3.7228	4.7597	5.6611	6.4446
.1944				

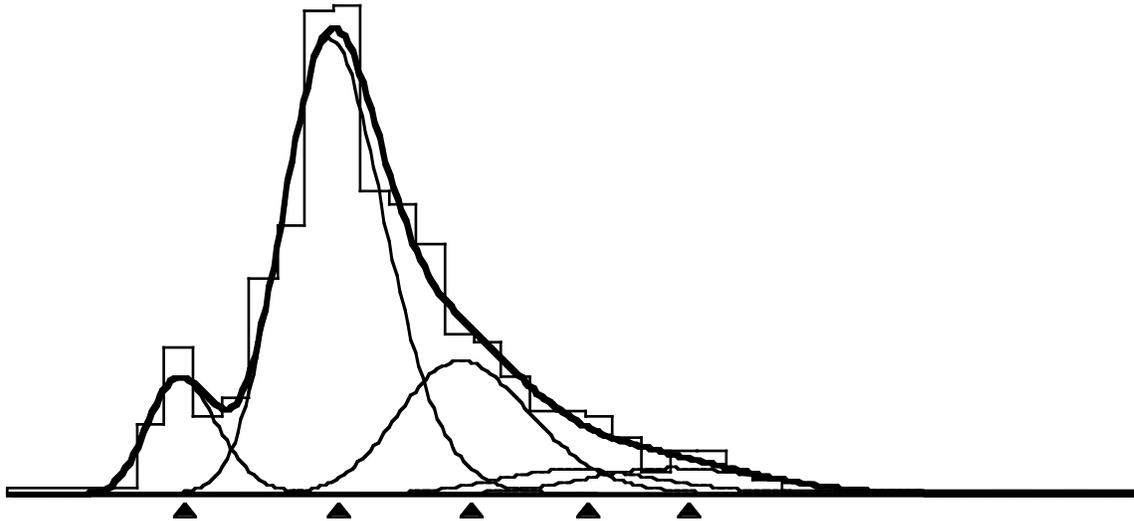
Degrees of freedom = 16

Chi-squared = 12.4566 (P = .7120)

Option number? [0 for list of options, -1 to STOP]

11

*Note that the means, from left to right, get progressively closer together, in accordance with the von Bertalanffy growth curve assumption.*



Plot #008 Data: Heming Lake Pi ke 1965 Components: Lognormal

Option number? [0 for list of options, -1 to STOP]

6

*The growth curve parameters  $L_{\infty}$  and  $K$  in the previous fit have very large standard errors; their coefficients of variation are about 50%. This suggests that there is not enough evidence in the data to support a von Bertalanffy growth-curve model, and that linear growth with equally-spaced means might fit almost as well. We will begin fitting this model with Option 6, holding the proportions and sigmas fixed while fitting the means.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

30

Display observed and expected counts as a table (Y/N) ?

N

Display observed and expected counts as a graph (Y/N) ?

N

---

Display variance-covariance matrix (Y/N) ?

**N**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

**1**

Which of the 5 proportions are fixed?

For each, in order, enter Y (FIXED) or N (FREE):

**YYYYY**

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

Enter choice:

**3**

Is Kth mean different (Y/N) ?

**N**

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),

2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),

4 (SIGMAS EQUAL).

Enter choice:

**1**

Which of the 5 sigmas are fixed?

For each, in order, enter Y (FIXED) or N (FREE):

**YYYYY**

Do you want to abort (Y/N) ?

**N**

Number of iterations = 10

Fitting Lognormal components

Proportions and their standard errors

.10137	.58091	.21581	.04628	.05564
FIXED	FIXED	FIXED	FIXED	FIXED

Means (EQUALLY SPACED) and standard errors

23.7755	33.8757	43.9760	54.0762	64.1765
.3706	.2729			

Sigmas and their standard errors

2.5300	3.7228	4.7597	5.6611	6.4446
FIXED	FIXED	FIXED	FIXED	FIXED

Degrees of freedom = 22

Chi-squared = 23.9032 (P = .3523)

Option number? [0 for list of options, -1 to STOP]

---

4

*Revise the proportions, to adjust for the new means.*

ESTIMATE PROPORTIONS FOR FIXED MEANS, SIGMAS

Distribution selected is Lognormal

Enter iteration limit:

**20**

Number of iterations = 6

Fitting Lognormal components

Proportions and their standard errors

.10385	.57470	.22737	.07187	.02220
.01491	.02694	.02813	.02061	.01034

Means (ALL HELD FIXED)

23.7755	33.8757	43.9760	54.0762	64.1765
---------	---------	---------	---------	---------

Sigmas (ALL HELD FIXED)

2.5300	3.7228	4.7597	5.6611	6.4446
--------	--------	--------	--------	--------

Degrees of freedom = 20

Chi-squared = 16.6046

Option number? [0 for list of options, -1 to STOP]

**6**

*Go for the final fit with all proportions free, means equally spaced,  
and constant coefficient of variation.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

**30**

Display observed and expected counts as a table (Y/N) ?

**N**

Display observed and expected counts as a graph (Y/N) ?

**N**

Display variance-covariance matrix (Y/N) ?

**N**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

**0**

Constraints on means:

---

```

0 (NONE),          1 (SPECIFIED MEANS FIXED),
2 (MEANS EQUAL),  3 (EQUALLY SPACED),
4 (GROWTH CURVE).
Enter choice:
3

Is Kth mean different (Y/N) ?
N

Constraints on sigmas:
0 (NONE),          1 (SPECIFIED SIGMAS FIXED),
2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),
4 (SIGMAS EQUAL).
Enter choice:
3

Is Kth sigma different (Y/N) ?
N

Do you want to abort (Y/N) ?
N

Number of iterations = 14

Fitting Lognormal components

Proportions and their standard errors
.10303  .62186  .20236  .06616  .00659
.01611  .03650  .03087  .02163  .00996

Means (EQUALLY SPACED) and standard errors
23.3395  34.4864  45.6333  56.7802  67.9271
.4942   .4344

Sigmas (CONSTANT COEF. OF VAR. = .1141) and standard error
2.6636  3.9357  5.2078  6.4799  7.7520
.1926

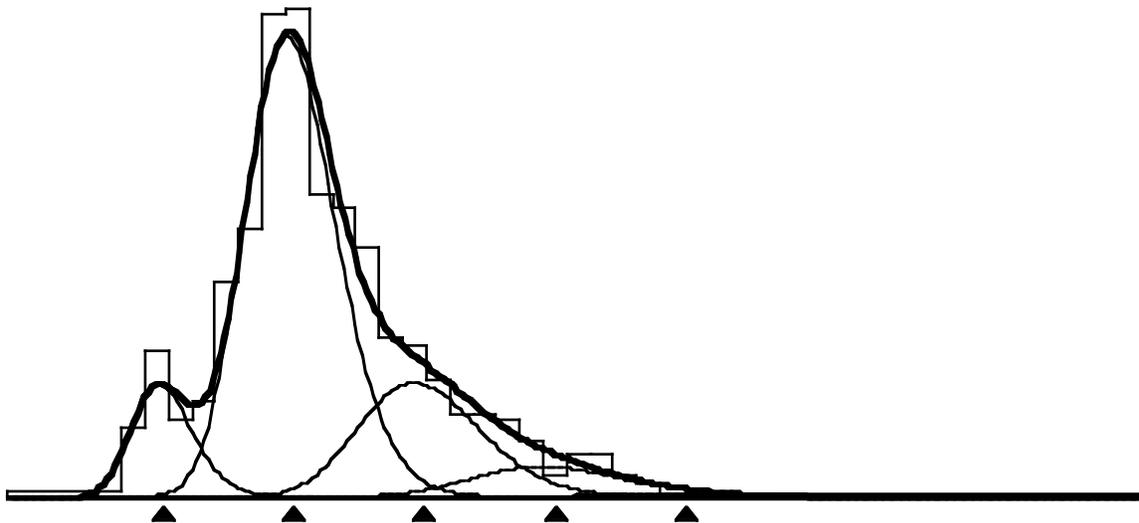
Degrees of freedom = 17

Chi-squared = 13.6015          (P = .6951)

Option number? [0 for list of options, -1 to STOP]
11

```

*Although the chi-square test above indicates an excellent fit, the equal-means constraint has pushed the 5<sup>th</sup> component almost off the histogram.*



Plot #009 Data: Heming Lake Pike 1965 Components: Lognormal

*We will finish this session by demonstrating how to lump all the oldest age-groups into a single component. This is useful if the right tail of the histogram is an ill-defined smear comprising several sparse, old age-groups. For this example, we will treat ages 4 and 5 as a single component and show that we can still get excellent estimates of the parameters of the first three age-groups.*

Option number? [0 for list of options, -1 to STOP]

8

*We begin by combining the rightmost intervals of the histogram. This is a good idea if the right tail ends in a sparse scatter of very large individuals, because putting them all into one grouping interval will make the fit less sensitive to them. This is not necessary for the pike data but we will demonstrate it anyway.*

REGROUP DATA OR RESTORE TO ORIGINAL GROUPING

Need to restore data to original grouping (Y/N) ?

N

Do you want to combine two adjacent classes (Y/N) ?

Y

Which boundary is to be deleted?

65.75

Do you want to combine two adjacent classes (Y/N) ?

Y

Which boundary is to be deleted?

63.75

---

Do you want to combine two adjacent classes (Y/N) ?

**N**

Regrouped data:

INTERVAL	OBSERVED COUNT	RIGHT BOUNDARY	Heming Lake Pike 1965
1	4.0000	19.7500	
2	10.0000	21.7500	
3	21.0000	23.7500	
4	11.0000	25.7500	
5	14.0000	27.7500	
6	31.0000	29.7500	
7	39.0000	31.7500	
8	70.0000	33.7500	
9	71.0000	35.7500	
10	44.0000	37.7500	
11	42.0000	39.7500	
12	36.0000	41.7500	
13	23.0000	43.7500	
14	22.0000	45.7500	
15	17.0000	47.7500	
16	12.0000	49.7500	
17	12.0000	51.7500	
18	11.0000	53.7500	
19	8.0000	55.7500	
20	3.0000	57.7500	
21	6.0000	59.7500	
22	6.0000	61.7500	
23	10.0000		

Do you want to store these data on a file (Y/N) ?

**N**

Option number? [0 for list of options, -1 to STOP]

**2**

*Initialize parameters for a 4-component fit. Components 1 to 3 correspond to components 1 to 3 in the previous fits, but component 4 now corresponds to the previous components 4 and 5 combined. We can begin with equal proportions, trusting Option 4 to improve them on the next step.*

READ A FULL SET OF PARAMETER VALUES

How many components? [Must be at least 1, at most 15]

**4**

Enter the 4 proportions:

**1 1 1 1**

Enter the 4 means:

**23 34 45 60**

Enter the 4 sigmas:

**2 3 4 6**

Proportions do not sum to 1. Do you want to re-scale (Y/N) ?

**Y**

---

Proportions  
.25000 .25000 .25000 .25000

Means  
23.0000 34.0000 45.0000 60.0000

Sigmas  
2.0000 3.0000 4.0000 6.0000

Option number? [0 for list of options, -1 to STOP]

4

*Revise the proportions, to adjust for the new means and sigmas.*

ESTIMATE PROPORTIONS FOR FIXED MEANS, SIGMAS

Distribution selected is Lognormal

Enter iteration limit:

20

Number of iterations = 6

Fitting Lognormal components

Proportions and their standard errors  
.10473 .57831 .24672 .07024  
.01380 .02356 .02167 .01239

Means (ALL HELD FIXED)  
23.0000 34.0000 45.0000 60.0000

Sigmas (ALL HELD FIXED)  
2.0000 3.0000 4.0000 6.0000

Degrees of freedom = 19

Chi-squared = 36.8714

Option number? [0 for list of options, -1 to STOP]

6

*Use Option 6 with proportions free, means equally spaced, and constant coefficient of variation, while holding fixed the mean and sigma of the 4<sup>th</sup> component. Because the 4<sup>th</sup> component does not represent a single age-group, it should not have to satisfy the same constraints as the first three components.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

30

---

Display observed and expected counts as a table (Y/N) ?

**N**

Display observed and expected counts as a graph (Y/N) ?

**N**

Display variance-covariance matrix (Y/N) ?

**N**

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

**0**

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

Enter choice:

**3**

Is Kth mean different (Y/N) ?

**Y**

Hold Kth mean fixed (Y/N) ?

**Y**

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),

2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),

4 (SIGMAS EQUAL).

Enter choice:

**3**

Is Kth sigma different (Y/N) ?

**Y**

Hold Kth sigma fixed (Y/N) ?

**Y**

Do you want to abort (Y/N) ?

**N**

Number of iterations = 10

Fitting Lognormal components

Proportions and their standard errors

.10239	.61921	.22138	.05703
.01594	.03682	.03255	.01353

Means (EQUALLY SPACED) and standard errors

23.2660	34.4931	45.7202	60.0000
.4670	.4222		FIXED

Sigmas (CONSTANT COEF. OF VAR. = .1144) and standard error

2.6619	3.9465	5.2310	6.0000
.1870			FIXED

---

Degrees of freedom = 16

Chi-squared = 13.1304 (P = .6632)

Option number? [0 for list of options, -1 to STOP]

6

*Repeat the previous fit, letting the mean and sigma of the 4<sup>th</sup> component go free.*

ESTIMATE PROPORTIONS, MEANS, SIGMAS WITH OR WITHOUT CONSTRAINTS  
AND/OR GIVE DIAGNOSTIC DISPLAYS

Distribution selected is Lognormal

Enter iteration limit [0 gives displays with current parameters]:

50

Display observed and expected counts as a table (Y/N) ?

N

Display observed and expected counts as a graph (Y/N) ?

N

Display variance-covariance matrix (Y/N) ?

N

Constraints on proportions:

0 (NONE), 1 (SPECIFIED PROPORTIONS FIXED).

Enter choice:

0

Constraints on means:

0 (NONE), 1 (SPECIFIED MEANS FIXED),

2 (MEANS EQUAL), 3 (EQUALLY SPACED),

4 (GROWTH CURVE).

Enter choice:

3

Is Kth mean different (Y/N) ?

Y

Hold Kth mean fixed (Y/N) ?

N

Constraints on sigmas:

0 (NONE), 1 (SPECIFIED SIGMAS FIXED),

2 (FIXED COEF. OF VARIATION), 3 (CONSTANT COEF. OF VARIATION),

4 (SIGMAS EQUAL).

Enter choice:

3

Is Kth sigma different (Y/N) ?

Y

Hold Kth sigma fixed (Y/N) ?

N

Do you want to abort (Y/N) ?

N

Number of iterations = 13

Fitting Lognormal components

Proportions and their standard errors

.10257	.61450	.22268	.06024
.01596	.03787	.03495	.02297

Means (EQUALLY SPACED) and standard errors

23.2958	34.4244	45.5531	59.3126
.4837	.4478		2.4084

Sigmas (CONSTANT COEF. OF VAR. = .1133) and standard error

2.6398	3.9009	5.1619	5.5745
.1908			2.2103

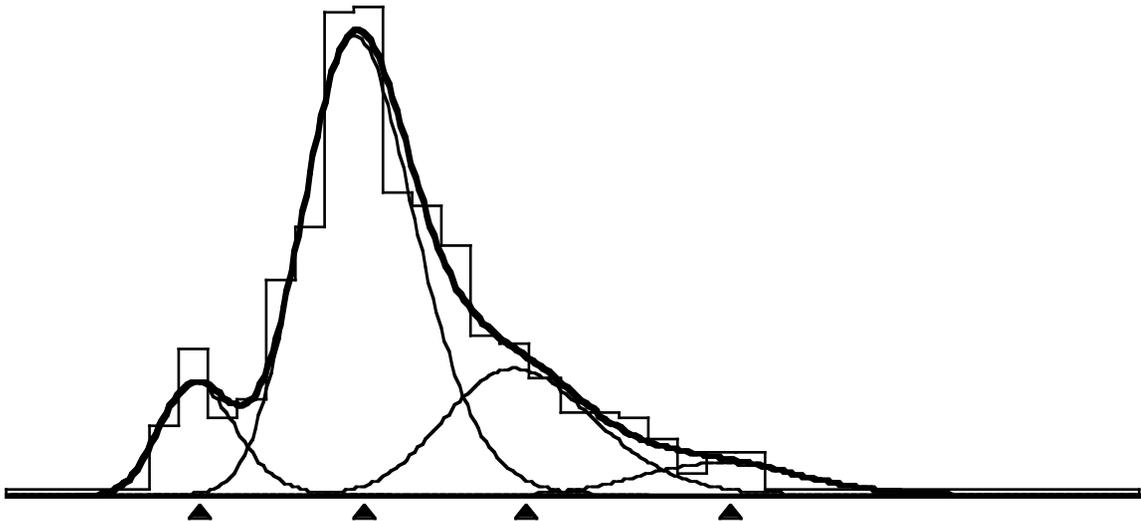
Degrees of freedom = 14

Chi-squared = 12.8913 (P = .5351)

Option number? [0 for list of options, -1 to STOP]

11

*This fit is more satisfactory than the one previously obtained with 5 equally-spaced means (Plot #009). The present fit finds almost exactly the same values for the parameters of components 1 to 3, but treats age-groups 4 and 5 as a single component.*



Plot #010 Data: Heming Lake Pike 1965 Components: Lognormal

Option number? [0 for list of options, -1 to STOP]

-1

Execution of MIX terminated.

### Standard licence agreement for MIX users

This Licence Agreement is made and entered into this \_\_\_\_\_ day of \_\_\_\_\_, 19\_\_ BETWEEN

ICHTHUS DATA SYSTEMS, a duly registered partnership under the laws of the Province of Ontario, Canada

OF THE FIRST PART

and

hereinafter called the Licensee OF THE SECOND PART.

WHEREAS the Licensor has developed a program relating to the statistical analysis of mixtures of distributions;

AND WHEREAS the Licensee is desirous of obtaining from the Licensor a licence to use the said program and the Licensor is desirous of granting a licence to the Licensee to allow its use of the program;

NOW THEREFORE THIS AGREEMENT WITNESS THAT in consideration of the mutual covenants, conditions and terms hereinafter set forth, and for other good and valuable consideration,

The Licensor hereby grants to the Licensee the nonexclusive right to use the program which is known as MIX subject to the terms and conditions of this Agreement and the Licensee hereby accepts such licence solely upon such terms and conditions:

1. The distribution fee for this licence of one physical copy of the program is \$225.00 (Canadian). Distribution fees are due and payable in advance. Distribution fees do not include local, provincial, state or federal taxes or any governmental taxes or duties whatsoever and the Licensee hereby agrees to pay all such taxes and/or charge which may be imposed upon the the Licensee or Licensor with respect to the distribution, possession and use of the program pursuant to this agreement.

2. The Licensee agrees that the program is and at all times the property of the Licensor; the Licensee shall have no right or interest therein except as expressly set forth in this Agreement.

3. The Licensee may:

(i) use the program on a single machine

(ii) copy the program into any machine-readable or printed form for backup or modification purposes only;

(iii) modify the program and/or merge it into another program for use on a single machine; the terms of this Agreement shall continue to apply to the portion of the program used.

4. The Licensee covenants and agrees not to transfer the program and licence to any other party; should possession of the program or a copy thereof be transferred to another party, this licence is automatically terminated and the Licensee shall be liable to pay to the Licensor any damages suffered by the Licensor as a result of the Licensee having breached this Agreement.

5. This licence is effective until terminated. It shall be terminated:

(i) by the licensee destroying its copy of the program together with any copies, modifications and/or merged portions of the program; or

(ii) by the Licensee breaching any of the terms and conditions of this Agreement.

6. THE LICENSOR DOES NOT MAKE ANY WARRANTIES, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER WHATSOEVER, INCLUDING WITHOUT LIMITATION, THE MERCHANTABILITY AND FITNESS OF THE PROGRAM FOR ANY PARTICULAR PURPOSE. SHOULD THE PROGRAM PROVE DEFECTIVE, THE LICENSEE (AND NOT THE LICENSOR) SHALL ASSUME THE ENTIRE COST OF ANY REPAIR, SERVICING AND/OR CORRECTION.

7. Should the Licensor be required to take any legal proceedings to enforce this Agreement, its full cost of doing so shall be paid by the Licensee.

8. If any part, term, or provision of the Agreement shall be held illegal, unenforceable or in conflict with any law of a federal, provincial, state or Government whatsoever having jurisdiction over this Agreement, the remaining portions or provisions shall not be affected thereby.

9. This Agreement shall be construed and enforced according to the laws of the Province of Ontario. If the Licensee is located outside Canada, the parties hereto agree that any dispute arising in connection with this agreement shall nonetheless be determined by the Ontario Court System.

10. This Agreement contains all the agreements, representations, and understandings of the parties hereto and supersedes any previous agreements, and/or commitments oral or written.

11. Each of the undersigned warrants that he/she has the authority to bind to this Agreement the party which he/she represents.

IN WITNESS WHEREOF, the parties hereto have executed this Agreement as of the day and year first above written.

ICHTHUS DATA SYSTEMS

(Licensor)

per: \_\_\_\_\_

Date:

(Licensee)

per: \_\_\_\_\_

Date: