NRL Report 6226

Statistical Decomposition of a Mixture of Two Gaussian Populations

H. Bixhorn

Optimization Techniques Branch
Applied Mathematics Staff

March 29, 1965

APPROVED FOR PUBLIC
RELEASE - DISTRIBUTION
UNLIMITED

U.S. NAVAL RESEARCH LABORATORY
Washington, D.C.

CONTENTS

Abstract	ii
Problem Status	ii
Authorization	ii
INTRODUCTION	1
ESTIMATION BY THE METHOD OF MOMENTS	2
Case I: $f(x)$ is Asymmetric	2
Case II: $f(x)$ is Symmetric	3
ESTIMATION BY THE METHOD OF MAXIMUM LIKELIHOOD	5
APPLICATIONS	6
ACKNOWLEDGMENTS	7
Appendix A - SOME CONCEPTS IN THE THEORY OF ESTIMATION	8
REFERENCES	9
Appendix B - COMPUTATION OF ESTIMATES BY THE METHOD OF MOMENTS	10

ABSTRACT

The problem of decomposing a population that is a mixture of two Gaussian subpopulations is approached by two statistical methods: the method of moments and the method of maximum likelihood. The density function of such a population is of the form

$$f(x) = \omega_1 f_1(x) + \omega_2 f_2(x)$$

where

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{\frac{-(x-\alpha_1)^2}{2\sigma_1^2}} \quad \text{and} \quad f_2(x) = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{\frac{-(x-\alpha_2)^2}{2\sigma_2^2}},$$

with the mixing constants satisfying the equation $\omega_1 + \omega_2 = 1$. The problem of decomposing the original population is equivalent to the problem of determining the values of the parameters α_1, α_2, σ_1, σ_2, ω_1, and ω_2, for if these values are known, the Gaussian subpopulations are completely determined. The purpose of this report is to obtain estimates a_1, a_2, s_1, s_2, w_1, and w_2 of the parameters α_1, α_2, σ_1, σ_2, ω_1, and ω_2, respectively, and to investigate some statistical properties of these estimates.

Computer programs which obtain solutions by the method of moments have been written and successfully used on the Naval Research Electronic Computer (NAREC). On the other hand, certain difficulties arise in applying the method of maximum likelihood, but studies are in progress to obtain solutions by this method.

PROBLEM STATUS

This report completes one phase of the problem; work on other phases of the problem is continuing.

AUTHORIZATION

NRL Problem F04-05
Project RR 009-03-45-5802

Manuscript submitted November 24, 1964.

STATISTICAL DECOMPOSITION OF A MIXTURE OF TWO GAUSSIAN POPULATIONS

INTRODUCTION

In various fields of scientific research, there is a frequent occurrence of populations consisting of two Gaussian subpopulations. The density function of such a population is of the form

$$f(x) = \omega_1 f_1(x) + \omega_2 f_2(x) \tag{1}$$

where

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\alpha_1)^2}{2\sigma_2^2}} \quad \text{and} \quad f_2(x) = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\alpha_2)^2}{2\sigma_2^2}},$$

with the mixing constants satisfying the equation $\omega_1 + \omega_2 = 1$. It is often the case that the given total population is known to be of this form, but it is not known which two Gaussian populations compose the total population and in what proportion they are present. The problem of decomposing the original population into two Gaussian components is equivalent to the problem of determining the values of the parameters α_1, α_2, σ_1, σ_2, ω_1, and ω_2, for if these values are known, the Gaussian subpopulations are completely determined. The purpose of this report will therefore be to obtain estimates a_1, a_2, s_1, s_2, w_1, and w_2 of the parameters α_1, α_2, σ_1, σ_2, ω_1, and ω_2, respectively, and to investigate some statistical properties of these estimates.

Before proceeding with the problem of estimation of parameters, it should be mentioned that the decomposition of any population whose density is of the form of Eq. (1) into two Gaussian subpopulations is unique. That is to say, if

$$\begin{aligned} f(x) &= \frac{\omega_1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\alpha_1)^2}{2\sigma_1^2}} + \frac{\omega_2}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\alpha_2)^2}{2\sigma_2^2}} \quad (\text{where } \omega_1 + \omega_2 = 1) \\ &= \frac{\omega_3}{\sqrt{2\pi}\,\sigma_3} e^{-\frac{(x-\alpha_3)^2}{2\sigma_3^2}} + \frac{\omega_4}{\sqrt{2\pi}\,\sigma_4} e^{-\frac{(x-\alpha_4)^2}{2\sigma_4^2}} \quad (\text{where } \omega_3 + \omega_4 = 1), \end{aligned} \tag{2}$$

then

$$\begin{aligned} \omega_1 &= \omega_3, & \alpha_1 &= \alpha_3, & \sigma_1 &= \sigma_3 \\ \omega_2 &= \omega_4 & \alpha_2 &= \alpha_4 & \sigma_2 &= \sigma_4. \end{aligned} \tag{3}$$

This can be proved by performing certain algebraic operations on Eq. (2), letting x approach ∞ and then demonstrating that contradictions result if Eqs. (3) do not hold.

ESTIMATION BY THE METHOD OF MOMENTS

A general method of estimation which is widely used because of its practical expediency is the method of moments, developed by Karl Pearson. This method involves equating population moments to sample moments and then solving for the unknown population parameters. This can be demonstrated for the problem in question.

The population moments $\mu_1', \mu_2, \mu_3, \ldots, \mu_\nu, \ldots$, are defined by

$$\mu_1' = \int_{-\infty}^{\infty} x f(x)\, dx ,$$

and

$$\mu_\nu = \int_{-\infty}^{\infty} (x - \mu_1')^\nu \; f(x)\, dx \qquad (\nu = 2, 3, \ldots) .$$

Thus,

$$\mu_1' = \int_{-\infty}^{\infty} x \left[\frac{\omega_1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-a_1)^2}{2\sigma_1^2}} + \frac{\omega_2}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-a_2)^2}{2\sigma_2^2}} \right] dx$$

and

$$\mu_\nu = \int_{-\infty}^{\infty} (x - \mu_1')^\nu \left[\frac{\omega_1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-a_1)^2}{2\sigma_1^2}} + \frac{\omega_2}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-a_2)^2}{2\sigma_2^2}} \right] dx \qquad (\nu = 2, 3, \ldots) .$$

These moments are clearly seen to be functions of the population parameters. Given a set of n sample values $(x_1, x_2, \ldots, x_n)$ drawn from the population, the sample moments $m_1', m_2, \ldots, m_\nu, \ldots$, are defined by

$$m_1' = \frac{1}{n} \sum_{i=1}^{n} x_i , \quad \text{and} \quad m_\nu = \frac{1}{n} \sum_{i=1}^{n} (x_i - m_1')^\nu \qquad (\nu = 2, 3, \ldots) .$$

By equating population moments to sample moments we get the following system of equations in the unknown parameters:

$$\mu_1'(\alpha_1, \alpha_2, \sigma_1, \sigma_2, \omega_1, \omega_2) = m_1'$$

and

$$\mu_\nu(\alpha_1, \alpha_2, \sigma_1, \sigma_2, \omega_1, \omega_2) = m_\nu \qquad (\nu = 2, 3, \ldots) .$$

The solutions of this system of equations are denoted by a_1, a_2, s_1, s_2, w_1, and w_2 and are called the method of moments estimates. Applying this method and using a great deal of algebraic manipulation, we obtain the following results.

Case I: $f(x)$ is Asymmetric

For this case it is sufficient to consider the first five moments plus the constraint $\omega_1 + \omega_2 = 1$. The resulting system of equations was reduced to the ninth-degree polynomial equation

$$24p_2^9 - 28\lambda_1 p_2^7 + 36m_3^2 p_2^6 - (24m_3\lambda_2 - 10\lambda_1^2)\, p_2^5 - (148m_3^2\lambda_1 + 2\lambda_2^2)\, p_2^4$$

$$+ (288m_3^4 - 12\lambda_1\lambda_2 m_3 - \lambda_1^3)\, p_2^3 + (24m_3^3\lambda_2 - 7m_3^2\lambda_1^2)\, p_2^2 + 32m_3^4\lambda_1 p_2 - 24m_3^6 = 0\,, \tag{4}$$

where

$$\lambda_1 = 9m_2^2 - 3m_4, \text{ and } \lambda_2 = 30m_2m_3 - 3m_5,$$

which must be solved for p_2.

Once p_2 is found, the roots v_1 and v_2 of the quadratic

$$v^2 - p_1 v + p_2 = 0\,, \tag{5}$$

where

$$p_1 = \frac{2m_3^3 - 2m_3\lambda_1 p_2 - \lambda_2 p_2^2 - 8m_3 p_2^3}{p_2(4m_3^2 - \lambda_1 p_2 + 2p_2^3)}\,,$$

must be determined.

Finally, the estimates are given by

$$a_1 = v_1 + m_1'\,, \qquad a_2 = v_2 + m_1' \tag{6}$$

$$w_1 = -\frac{v_2}{v_1 - v_2}\,. \qquad w_2 = \frac{v_1}{v_1 - v_2} \tag{7}$$

$$s_1^2 = m_2 - \frac{1}{3}\frac{m_3}{v_2} - \frac{1}{3}p_1 v_1 + p_2 \tag{8}$$

$$s_2^2 = m_2 - \frac{1}{3}\frac{m_3}{v_1} - \frac{1}{3}p_1 v_2 + p_2\,. \tag{9}$$

Although it was stated above that the decomposition is unique, this does not mean that the estimates will be unique. Equation (4) will often have more than one real root, and this in turn may lead to more than one real value for each of the estimates a_1, a_2, s_1, s_2, w_1, and w_2. If this should happen, the "best" estimate may have to be that which best agrees with other aspects of the problem under study.

Case II: $f(x)$ is Symmetric

The case where $f(x)$ is symmetric may be further divided into two cases: (a) $\alpha_1 = \alpha_2$, and (b) $\alpha_1 \neq \alpha_2$, but $\sigma_1 = \sigma_2$ and $\omega_1 = \omega_2$.

A. $\alpha_1 = \alpha_2$:

This case is a most deceptive one. As can be seen in Fig. 1, when two Gaussian density functions with equal means are added, the result itself may appear to be a Gaussian density function. Hence, if a histogram plotted from the values of a sample seems to approach a bell-shaped curve, it should not be assumed (unless other aspects

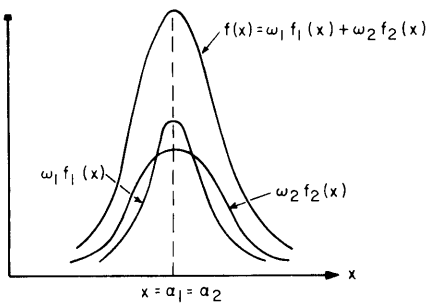

Fig. 1 - Superposition of two Gaussian density functions with equal means ($\alpha_1 = \alpha_2$)

of the problem warrant it) that the population is in itself, Gaussian. It may instead be a mixture of two Gaussian populations.

For a Gaussian population, the equations $\mu_4 = 3\mu_2^2$ and $\mu_6 = 5\mu_2\mu_4$ hold. A simple test, then, of whether a given symmetrical histogram approaches a Gaussian density curve is given by whether or not $m_4 = 3m_2^2$ and $m_6 = 5m_2 m_4$. Now, if these latter relations do not approximately hold, an attempt should be made to represent the population as the superposition of two Gaussian populations. We equate the two estimates α_1 and α_2 to m_1'

$$a_1 = a_2 = m_1'$$

and require that the estimated variances, s_1^2 and s_2^2, be roots of the equation

$$(m_4 - 3m_2^2)\, y^2 + \left(m_2 m_4 - \frac{1}{5} m_6\right) y + \left(\frac{1}{5} m_2 m_6 - \frac{1}{3} m_4^2\right) = 0 .$$

(Notice that here we also require m_6, the sixth sample moment.) Finally, we obtain the estimated weights

$$w_1 = \frac{m_2 - s_2^2}{s_1^2 - s_2^2} \quad \text{and} \quad w_2 = \frac{m_2 - s_1^2}{s_2^2 - s_1^2} .$$

B. $\alpha_1 \neq \alpha_2$, but $\sigma_1 = \sigma_2$ and $\omega_1 = \omega_2$:

For this case we have

$$\omega_1 = \omega_2 = \frac{1}{2}$$

so that

$$a_1 = \left(\frac{3m_2^2 - m_4}{2}\right)^{1/4} + m_1'$$

$$a_2 = -\left(\frac{3m_2^2 - m_4}{2}\right)^{1/4} + m_1'$$

and

$$s_1 = s_2 = \left(m_2 - \sqrt{\frac{3m_2^2 - m_4}{2}}\right)^{1/2} .$$

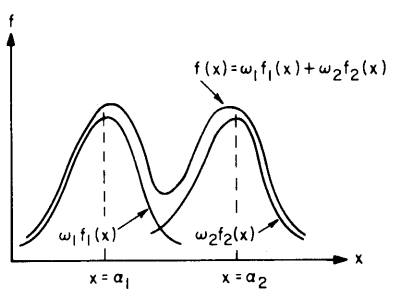

Fig. 2 - Superposition of two Gaussian density functions with different means ($\alpha_1 \neq \alpha_2$) but equal variances ($\sigma_1 = \sigma_2$) and weights ($\omega_1 = \omega_2$)

It can be shown that the estimates a_1, a_2, s_1, s_2, w_1, and w_2 are each asymptotically normally distributed with respective means α_1, α_2, σ_1, σ_2, ω_1, and ω_2, i.e., the estimates are asymptotically unbiased. The variance of each estimate tends to zero as the sample size $n \to \infty$. Thus, each estimate is a consistent estimate of the corresponding parameter.*

*For proofs of these statements, see H. Bixhorn, "Statistical Methods for Decomposing a Mixture of Two Normal Populations," thesis submitted in partial fulfillment of requirements for Master of Arts degree, The Catholic University of America, Washington, D.C. (1964).

ESTIMATION BY THE METHOD OF MAXIMUM LIKELIHOOD

Let $f(x;\underset{\sim}{\theta})$ be the density function of a population where $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ is a vector of unknown population parameters. The likelihood function of a sample $(x_1, x_2, \ldots, x_n)$ taken from this population is defined by

$$L(x_1, \ldots, x_n; \underset{\sim}{\theta}) = f(x_1; \underset{\sim}{\theta})\, f(x_2; \underset{\sim}{\theta}) \ldots f(x_n; \underset{\sim}{\theta}) .$$

The principle of the method of maximum likelihood is to take estimates of the unknown parameters that maximize the probability of obtaining the observed sample. The values $\hat{\theta}_1(x_1, \ldots, x_n)$, $\hat{\theta}_2(x_1, \ldots, x_n), \ldots, \hat{\theta}_p(x_1, \ldots, x_n)$ of the parameters $\theta_1, \ldots, \theta_p$, determined by the system of equations

$$\frac{\partial L}{\partial \theta_i} = 0 \qquad (i = 1, 2, \ldots, p) ,$$

are called maximum likelihood estimates of $\theta_1, \ldots, \theta_p$ if these values maximize L. Since L and $\log L$ attain their maximums for the same values of $\underset{\sim}{\theta}$, the preceding set of equations may be replaced by

$$\frac{\partial(\log L)}{\partial \theta_i} = 0 \qquad (i = 1, 2, \ldots, p) .$$

Now

$$\log L(x_1, \ldots, x_n; \underset{\sim}{\theta}) = \sum_{i=1}^{n} \log f(x_i; \underset{\sim}{\theta}) .$$

For the decomposition problem we have

$$f(x_i; \theta) = f(x_i; \alpha_1, \alpha_2, \sigma_1, \sigma_2, \omega_1, \omega_2) = \frac{1}{\sqrt{2\pi}} \left[\frac{\omega_1}{\sigma_1} e^{-\frac{(x_i - \alpha_1)^2}{2\sigma_1^2}} + \frac{\omega_2}{\sigma_2} e^{-\frac{(x_i - \alpha_2)^2}{2\sigma_2^2}} \right],$$

and thus the system of equations

$$\frac{\partial(\log L)}{\partial \alpha_j} = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{1}{f(x_i; \underset{\sim}{\theta})} \left[\frac{\omega_j}{\sigma_j^3} (x_i - \alpha_j)\, e^{-\frac{(x_i - \alpha_j)^2}{2\sigma_j^2}} \right] = 0 \qquad (j = 1, 2)$$

$$\frac{\partial(\log L)}{\partial \sigma_j} = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{1}{f(x_i; \underset{\sim}{\theta})} \frac{\omega_j}{\sigma_j^2} \left[\frac{(x_i - \alpha_j)^2}{\sigma_j^2} - 1 \right] e^{-\frac{(x_i - \alpha_j)^2}{2\sigma_j^2}} = 0 \quad (j = 1, 2)$$

$$\frac{\partial(\log L)}{\partial \omega_1} = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{1}{f(x_i; \underset{\sim}{\theta})} \left[\frac{1}{\sigma_1} e^{-\frac{(x_i - \alpha_1)^2}{2\sigma_1^2}} - \frac{1}{\sigma_2} e^{-\frac{(x_i - \alpha_2)^2}{2\sigma_2^2}} \right] = 0$$

must be solved for α_1, α_2, σ_1, σ_2, ω_1, and ω_2. ($\partial(\log L)/\partial\omega_2$ is not considered since $\omega_2 = 1 - \omega_1$.) No method has yet been found to reduce this system of five transcendental equations in five unknowns to a more workable form. However, programs are being written to find solutions of this system by the Newton-Raphson method. It is expected that this iterative process will quickly converge if method of moments estimates are taken as an initial approximation of the unknown parameter values.

It can be shown that the maximum likelihood estimates are asymptotically unbiased and consistent. In addition to these properties (which the method of moments estimates also have), the maximum likelihood estimates are asymptotically efficient. For proofs of these statements and an extension of the definition of efficient estimate to the case of more than one parameter, see the thesis by H. Bixhorn previously cited.

APPLICATIONS

The method of moments was applied to several samples taken from populations consisting of two Gaussian subpopulations. For the most general case (Case I previously mentioned) this involved the solution of a ninth-degree equation. The roots of this equation, and thus the estimates of the parameters, were obtained by means of an equation-solving routine programmed for use on the Naval Research Electronic Computer (NAREC).

The first sample to which the method was applied was artificially constructed as follows. One hundred values were taken from a table of random deviates of a Gaussian population with a mean value of 0 and a standard deviation of 1. Fifty additional values were taken from another table of random deviates of the same population, but each of these latter values y_i was transformed into $z_i = 2y_i + 2\,(i = 1, 2, \ldots, 50)$. Forming one sample from the first one hundred values and from the fifty "transformed" values gave a random sample from a population with a density function

$$f(x) = \frac{\omega_1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\alpha_1)^2}{2\sigma_1^2}} + \frac{\omega_2}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\alpha_2)^2}{2\sigma_2^2}},$$

where

$$\omega_1 = 2/3 \qquad \alpha_1 = 0 \qquad \sigma_1 = 1$$

$$\omega_2 = 1/3 \qquad \alpha_2 = 2 \qquad \sigma_2 = 2\,.$$

The method of moments yielded the following estimates for the sample:

$$w_1 = 0.672 \qquad a_1 = 0.006 \qquad s_1 = 1.001$$

$$w_2 = 0.328 \qquad a_2 = 2.021 \qquad s_2 = 1.999\,.$$

Figure 3 shows a histogram of this sample.

Other samples to which the method of moments was applied resulted from a problem being studied by the Cosmic Rays Branch of the Nucleonics Division. Helium nuclei are known to be present in primary cosmic radiation in two forms: the lighter isotope ^{3}He, and the heavier more abundant isotope ^{4}He. Each of these isotopes has certain observable physical properties, such as a characteristic "rate of energy loss" and "coulomb scattering." If many measurements are made, say, of the rate of energy loss of many ^{3}He (or ^{4}He) ions, these measurements would approximately obey a Gaussian distribution with a mean value equal to that isotope's characteristic value. Measurements of the rate

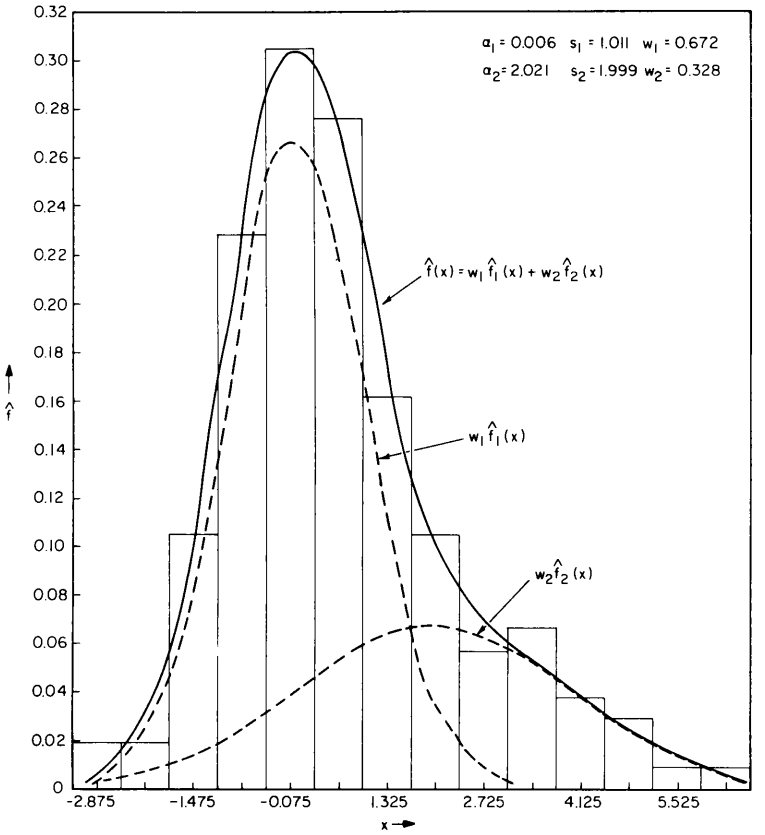

Fig. 3 - Histogram of artificially constructed sample with the density functions given by

$$\hat{f}_i(x) = \frac{1}{\sqrt{2\pi}\, s_i} e^{-\frac{(x-a_i)^2}{2 s_i^2}}$$

a_i, s_i, and w_i are method of moments estimates ($i = 1, 2$).

of energy loss of helium nuclei present in cosmic radiation are therefore considered to be drawn from a mixture of two Gaussian populations. The method of moments was used to decompose this population, and hence it was possible to estimate the relative abundance of each isotope of helium.

The measurements of coulomb scattering were also subjected to decomposition by the method of moments. Since the rate of energy loss is independent of coulomb scattering, an indication of the accuracy of the decomposition could be had by comparison of the w_1, w_2 values for both sets of measurements. In other words, the estimates of the relative abundance of each variety of helium should be approximately the same for the two sets of measurements.

Results of decomposition of the above samples, together with histograms of the samples and plots of population densities, are given in Appendices A and B.

ACKNOWLEDGMENTS

The author wishes to express his thanks to Miss J. Fisher for assistance in writing the method of moments programs and to Mr. R. McGill for useful suggestions on all phases of the work.

Appendix A

SOME CONCEPTS IN THE THEORY OF ESTIMATION

The following brief discussion will attempt to give the reader some understanding of the concepts of unbiased, efficient, and consistent estimates. It is assumed that the reader is familiar with the notions of populations, random variables, density functions, and means and variances as descriptions of density functions.

Suppose a random variable x has a density function $f(x;\theta)$, where θ is an unknown population parameter. For example, x may obey a Gaussian distribution of the form

$$f(x;m,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

where the value of the parameter m is unknown. A common problem in mathematical statistics is the estimation of such unknown parameters. This is done by drawing a sample of n items $x_1, x_2, \ldots, x_n$ from the population and inferring from this the value of the parameter. Usually there is a standard procedure by which a function of the sample values, say $g = g(x_1, x_2, \ldots, x_n)$, can be found to serve as an estimate. This function g is itself a random variable possessing a certain density function. In the above example, one could use the function

$$g(x_1, x_2, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

as an estimate of m. Clearly if the estimate g is to be of any use, it must at least approximate the true parameter value and, in some sense, converge to this value as the sample size $n \to \infty$. Further desirable properties of estimates are given below.

Let $f(x;\theta)$ be the density function of a random variable x and $g = g(x_1, x_2, \ldots, x_n)$ be an estimate of the population parameter θ. Then: (a) g is said to be an <u>unbiased</u> estimate if $E(g)$ (the mean of g) is equal to the value of θ. (b) It can be shown that, under certain general conditions, the variances of the class of all unbiased estimates of θ are bounded below by the number

$$M = \frac{1}{nE\left[\left(\frac{\partial[\log f(x;\theta)]}{\partial\theta}\right)^2\right]}.$$

An unbiased estimate g is said to be <u>efficent</u> if its variance $E[g - E(g)]^2 = E(g-\theta)^2$ is equal to M. Because the variance is a measure of dispersion of a random variable about its mean, the values assumed by an efficient estimate are clustered more closely about the value of θ than those of a nonefficient estimate. (c) If for any $\epsilon > 0$, $\delta > 0$, and sufficiently large n the inequality $P(|g-\theta| > \epsilon) < \delta$ holds, g is said to be a <u>consistent</u> estimate. Thus the probability that g differs from θ by an amount ϵ can be made arbitrarily small by taking a sufficiently large sample. Estimates are said to be asymptotically unbiased or efficient if the above-mentioned properties hold only as $n \to \infty$.

For a complete discussion of these concepts, see Refs. A1-A4.

REFERENCES

A1. Cramer, H., "The Elements of Probability Theory," John Wiley and Sons, Inc., New York (1955)

A2. Cramer, H., "Mathematical Methods of Statistics," Princeton University Press, Princeton, N.J. (1946)

A3. Fisz, M., "Probability Theory and Mathematical Statistics," 3rd ed., John Wiley and Sons, Inc., New York (1963)

A4. Hoel, P., "Introduction to Mathematical Statistics," 2nd ed., John Wiley and Sons, Inc., New York (1954)

Appendix B

COMPUTATION OF ESTIMATES BY THE METHOD OF MOMENTS

A program was written in NAREC floating-point language to compute estimates for Case I ($f(x)$ is asymmetric). This involved obtaining the roots of Eq. (4). When given a sample in floating-point decimal notation, tape 5330-A will compute the first five sample moments and the coefficients needed in Eq. (4), obtain roots of this equation, and then compute the estimates of the parameters.

Suppose Eq. (4) has n real roots ${}^1p_2, {}^2p_2, \ldots, {}^np_2 \, (n \leq 9)$. The following quantities will be printed out in the order given below:

0001	m_1'	0013	${}^1s_1^2$
0002	m_2	0014	${}^1s_2^2$
0003	m_3	0015	${}^2p_2$
0004	m_4	0016	${}^2p_1$
0005	m_5	0017	${}^2a_1$
0006	+)+)	0018	${}^2a_2$
0007	${}^1p_2$	0019	${}^2w_1$
0008	${}^1p_1$		----
0009	${}^1a_1$		----
0010	${}^1a_2$		----
0011	${}^1w_1$		----
0012	${}^1w_2$		etc.

Tape 5330-A must be used in conjunction with tape 2700, a floating-point subroutine, and tape 4197, an equation-solving routine. These tapes take the following locations:

	First Address	Last Address
Tape 2700	0000	03FF
Tape 4197	0800	08C2.

The following entries must be made in tape 5330-A:

location 103A, 103B } Δx

103C, 103D } Tolerance

location 103E, 103F } lb

1040, 1041 } ub.

These quantities are to be entered as floating-point decimal numbers. Their meanings are given in NAREC Bulletin #35, "Equation Solving Routine." It can be shown that any nonnegative root of Eq. (4) will lead to imaginary values for the estimates. The values of lb and ub should therefore always be negative. It is possible to find reasonable values of lb and ub by the following procedure: Draw a histogram of the sample and from this make rough estimates of a_1 and a_2. From Eq. (6), obtain $v_1 = a_1 - m_1'$ and $v_2 = a_2 - m'_1$. These values can now be inserted in Eq. (5) to solve for p_2. This value will usually be a very rough estimate of p_2, but it will give some idea of the interval in which the root p_2 can be found.

The sample size must be stored in location 700 as a fixed point hexadecimal number. Sample values should be placed in location 701 and consecutive locations in floating-point decimal notation.

The entire decomposition routine is to be entered at the left side of location 1110. The routine will run until all estimates are computed and printed out, and it will then come to an automatic stop.

It may happen that a value of p_2 leads to imaginary values for the estimates. For example, s_1^2 or s_2^2 may be a negative quantity, or Eq. (5) might have no real solutions. In the former case the printout will be the same as that shown above, but in the latter case, instead of a complete printout of $p_2, p_1, \ldots$, the symbol +)+) will appear. If either of these cases arises, it might still be possible to obtain real estimates by the method of moments. In many situations estimates of α_1 and α_2 (call them $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$) are known from other aspects of the problem. From Eqs. (7), (8), and (9) we can then solve for w_1, w_2, s_1, and s_2 by inserting

$$v_1 = \tilde{\alpha}_1 - m_1' \qquad v_2 = \tilde{\alpha}_2 - m_1'$$

$$p_1 = v_1 + v_2 \qquad p_2 = v_1 v_2 .$$

This was actually done for the decomposition shown in Fig. (B3) because the program obtained only imaginary values for the estimates. The decomposition shown in Fig. (B2) was also done by this method, not out of necessity, but to compare the results with those obtained by complete application of the method of moments [Fig. (B1)].*

*In Figs. (B1-B3), $\hat{f}(x) = w_1 \hat{f}_1(x) + w_2 \hat{f}_2(x)$, where $\hat{f}_1(x)$ and $\hat{f}_2(x)$ are the Gaussian density functions into which the sample was decomposed.

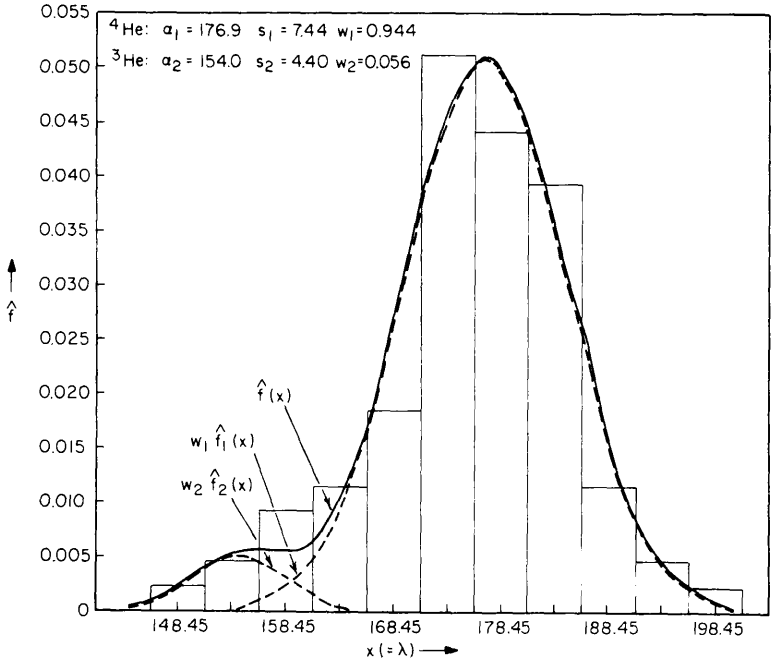

Fig. B1 - Histogram of sample consisting of 86 measurements of "rate of energy loss" λ of helium nuclei. The density functions of the sample are given by

$$\hat{f}_i(x) = \frac{1}{\sqrt{2\pi}\, s_i} e^{-\frac{(x-a_i)^2}{2 s_i^2}}$$

a_i, s_i, and w_i are method of moments estimates ($i = 1, 2$).

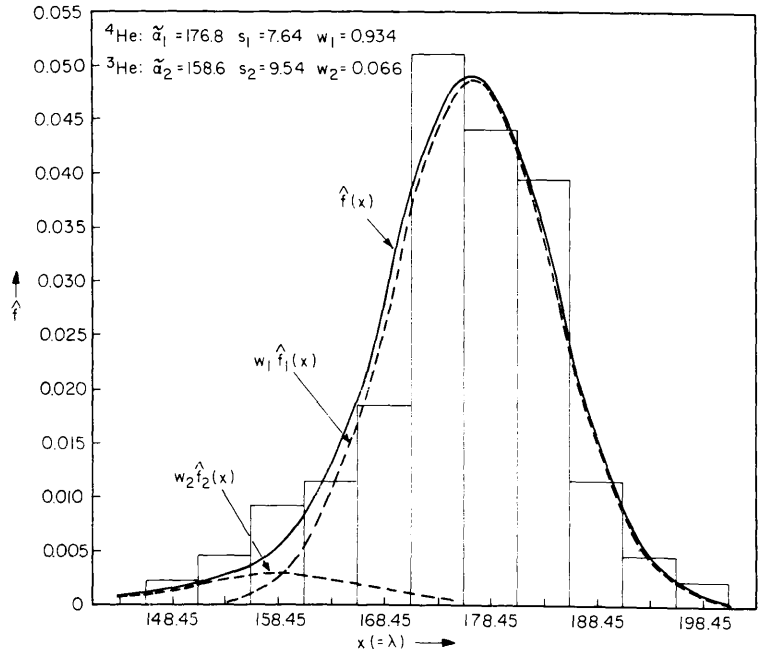

Fig. B2 - Histogram is same as in Fig. B1. The density functions of the sample are given by

$$\hat{f}_i = \frac{1}{\sqrt{2\pi}\, s_i} e^{-\frac{(x-\tilde{\alpha}_i)^2}{2 s_i^2}},$$

where $\tilde{\alpha}_i$ is an estimate of α_i obtained from other aspects of the problem, and s_i and w_i are method of moments estimates ($i = 1, 2$).

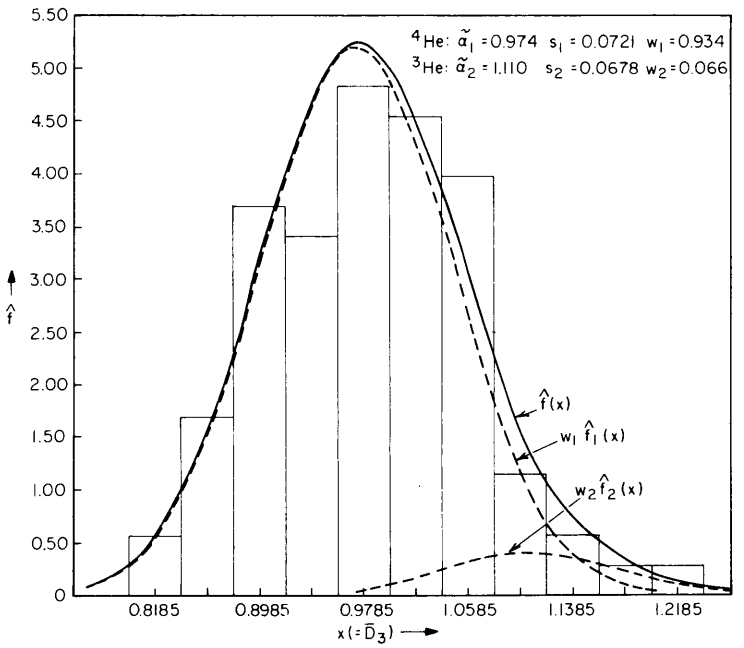

Fig. B3 - Histogram of sample consisting of 88 measurements of helium nuclei "coulomb scattering" $\bar{D}_3$. The density functions of the sample are given by

$$\hat{f}_i(x) = \frac{1}{\sqrt{2\pi}\, s_i} e^{-\frac{(x-\alpha_i)^2}{2 s_i^2}}$$

The estimates $\tilde{\alpha}_i$, s_i, and w_i are as described in Fig. B2 (i = 1, 2).

* * *

This page intentionally blank.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*

U.S. Naval Research Laboratory
Washington, D.C. 20390

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

STATISTICAL DECOMPOSITION OF A MIXTURE OF TWO GAUSSIAN POPULATIONS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

A final report on one phase of the problem.

5. AUTHOR(S) *(Last name, first name, initial)*

Bixhorn, H.

6. REPORT DATE

March 29, 1965

7a. TOTAL NO. OF PAGES

18

7b. NO. OF REFS

4

8a. CONTRACT OR GRANT NO.

NRL Problem F04-05

b. PROJECT NO.

RR 009-03-45-5802

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)

NRL Report 6226

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

None

10. AVAILABILITY/LIMITATION NOTICES

Unlimited availability. Copies available from Clearinghouse for Federal Scientific and Technical Information (CFSTI) Sills Building, 5285 Port Royal Road, Springfield, Virginia 22151.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Department of the Navy
(Office of Naval Research)

13. ABSTRACT

The problem of decomposing a population that is a mixture of two Gaussian subpopulations is approached by two statistical methods: the method of moments and the method of maximum likelihood. The density function of such a population is of the form

$$f(x) = \omega_1 f_1(x) + \omega_2 f_2(x)$$

where

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{(x-\alpha_1)^2}{2\sigma_1^2}} \quad \text{and} \quad f_2(x) = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{(x-\alpha_2)^2}{2\sigma_2^2}}$$

with the mixing constants satisfying the equation $\omega_1 + \omega_2 = 1$. The problem of decomposing the original population is equivalent to the problem of determining the values of the parameters α_1, α_2, σ_1, σ_2, ω_1, and ω_2, for if these values are known, the Gaussian subpopulations are completely determined. The purpose of this report is to obtain estimates a_1, a_2, s_1, s_2, w_1, and w_2 of the parameters α_1, α_2, σ_1, σ_2, ω_1, and ω_2, respectively, and to investigate some statistical properties of these estimates.

Computer programs which obtain solutions by the method of moments have been written and successfully used on the Naval Research Electronic Computer (NAREC). On the other hand, certain difficulties arise in applying the method of maximum likelihood, but studies are in progress to obtain solutions by this method.

DD FORM 1 JAN 64 1473

Security Classification

14. KEY WORDS	LINK A ROLE	LINK A WT	LINK B ROLE	LINK B WT	LINK C ROLE	LINK C WT
Gaussian populations - mixtures						
Gaussian populations - statistical analysis						
Helium isotope distribution - Gaussian approximations						
Gaussian approximations - computer programing						
Population mixtures - decomposition						
Parameter estimates - method of moments						
Parameter estimates - method of maximum likelihood						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.