MIXTURES OF DISTRIBUTIONS

by

RAY LEE MILLER, JR.

B. S., University of Missouri, 1957

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1964

Approved by:

Leslie F. Marcus

Major Professor

TABLE OF CONTENTS

INTRODUCTION

Mixtures of distributions occur when there is an overlapping of two or more distributions so that it is difficult to separate them into their respective components. The main statistical problem is to estimate as accurately as possible the true proportions of overlapping contributed by each distribution and their respective parameters. Mixtures of distributions may also be considered as a form of contagion.

There are many examples of mixtures of distributions or contagious distributions in our environment. Feller (1943) has distinguished two types of these distributions: 1) true contagion and 2) apparent contagion. Student's typing errors, payroll check errors, or bank statement and ledger errors are some of the situations in which the occurrence of one "favorable" event might affect the probability of another event happening. This is true contagion. Heterogeneity of decaying radioactive material or atmospheric data are examples which involve apparent contagion.

The purpose of this paper is to describe mixtures of distributions; to develop their fundamental distributions and interrelationship; and give examples of estimating the parameters of mixtures of exponential distributions by the method of moments and the method of maximum likelihood.

HISTORICAL REVIEW

As early as 1894 Karl Pearson had attacked the complicated problem of mixed frequency distributions (Rider, 1961). In particular, he considered dissecting[1] nonnormal populations into normal components. Pearson considered only two distributions in each of which a certain character is distributed normally. The statistical problem reduced to that of estimating the two mean values μ_1, μ_2 standard deviations σ_1, σ_2 and the proportions of mixture ρ and $(1 - \rho)$ from the observed frequency distribution. The five parameters were estimated by the method of moments and a solution to the estimates depended upon a suitably chosen root of a nonic (ninth degree equation) constructed from the first five moments of the observed frequency distribution.

In 1920 Greenwood and Yule developed a very general scheme for contagious events through their studies (diseases and accidents) concerned with the nature of frequency distributions representative of multiple happenings. Due to this generality their formulas became too complex for most practical applications (Feller, 1943).

About three years later Polya and Eggenberger were considering a similar problem which later led them to consider a special model of true contagion which is the simplest case of the general Greenwood-Yule scheme. Thus, Feller (1943) warns

[1]"Dissection" means point estimation of the parameters in a parametric mixture model.

that "in order to decide whether or not there is true contagion, it is not sufficient to consider the distribution of events, but a detailed study of the correlation between various time intervals is necessary."

It wasn't until 1939 when J. Neyman applied contagious distributions to entomology and bacteriology that there was a significant advancement in the study of contagious distributions. Neyman (1939) considered the distribution of larvae in a field which had been divided into plots of equal areas. In the experiments described by Neyman "the attempts to fit the Poisson Law... failed almost invariably with the characteristic feature that, as compared with the Poisson Law, there were too many empty plots and too few plots with only one larva." Thus, this is an excellent example of true contagion because the appearance of one larva in a plot seemed to increase the probability of finding at least one more larva in the same plot. From the related distributions that Neyman derived to fit the experimental data there evolved the generalized Neyman's distribution.

GENERAL PROBLEM OF MIXED DISTRIBUTIONS

The problem of mixtures of two distributions is to know exactly with what proportion each distribution contributes in the area where overlapping exists so that each distribution can be reconstructed in that area to learn what effect the second or first distribution had on the other. For example, it may sometimes happen that two normally distributed populations are mixed with nearly the same mean but different standard deviations. Cassie (1954) illustrated this case by graphing the length of 63 fish meshed in the cod-end of a trawl versus their cumulative frequency and obtained a curve indicating positive kurtosis. It was desirable to know exactly the proportion each distribution contributed in the area where overlapping occurred. Thus, reconstruction in that area would show how each distribution affected the other.

CRITERIA OF MIXTURES OF DISTRIBUTIONS

Credit for much of the following is due to Feller (1943) and Conover (1962).

Theorem:

If $F(x,a)$ is a distribution function depending on the parameter a, and $H(a)$ is a distribution function, then

$$G(x) = \int F(x,a)\, dH(a)$$

is a distribution function.

Proof:

It is sufficient to show that $G(x)$ satisfies the following three conditions:

(i) $G(x)$ is monotone non-decreasing, i.e., $G(x+b) \geq G(x)$ if $b > 0$.

(ii) $G(x)$ is right continuous, i.e., $G(x+0) = \lim_{b \to 0} G(x+b) = G(x)$.

(iii) $\lim_{x \to \infty} G(x) = 1 \quad . \quad \lim_{x \to -\infty} G(x) = 0$

Because $F(x,a)$ is a distribution function and therefore monotone non-decreasing, (i) is true. Thus,

$$G(x+b) = \int F(x+b,a)\, dH(a) \geq \int F(x,a)\, dH(a) = G(x) \; ; \; b > 0.$$

(ii) is right continuous since

$$G(x+0) = \int F(x+0,a)\, dH(a) = \int F(x,a)\, dH(a) = G(x)$$

where $F(x,a)$ must have been right continuous. The last condition is true because

$$\lim_{x \to \infty} F(x,a) = 1 \qquad \text{and} \qquad \lim_{x \to -\infty} F(x,a) = 0 \; ;$$

thus,

$$\lim_{x \to \infty} G(x) = \lim_{x \to \infty} \int F(x,a)\, dH(a) = \int \lim_{x \to \infty} F(x,a)\, dH(a)$$

$$= 1.dH(a) = 1$$

because $H(a)$ is also a distribution function. Also,

$$\lim_{x \to -\infty} G(x) = \lim_{x \to -\infty} \int F(x,a)\, dH(a) = \int \lim_{x \to -\infty} F(x,a)\, dH(a)$$

$$= 0.dH(a) = 0.$$

Therefore, both $F(x,a)$ and $H(a)$ determine that $G(x)$ is a distribution also.

Definition:

A distribution function $G(x)$ is called a "mixture of distributions" if

$$(1) \qquad G(x) = \int F(x,a)\, dH(a)$$

where $F(x,a)$ is an arbitrary non-degnerate cumulative distribution function[1] (c.d.f.), depending on a parameter $\underline{a}$, and another c.d.f. $H(a)$. The domain of variation of $\underline{a}$ determines the range of integration. If $H(a)$ is a step function, then we must define a non-degenerate cumulative distribution $F(x,a_i)$ whose parameters are finite numbers $a_1, a_2, \ldots, a_i$; let $a_1 \in A_1, a_2 \in A_2, \ldots, a_i \in A_i$ where $A_1, A_2, \ldots, A_i$ are given sets of real numbers that make up the population. Let p_i be the weight attached to the population A_i which are mixed at random in proportions $p_1 : p_2 : \ldots : p_i$ ($p_i \geq 0$ and $\sum p_i = 1$) where p_i are real constants. Thus, a family of distribution functions may be obtained by letting the parameters vary over A_i independently of each other. The function

(2) $$G(x) = \sum p_i \, F(x,a_i)$$

or more simply

(3) $$G(x) = \sum p_i F_i(x)$$

is a cumulative non-degenerate distribution whose components are $F_i(x)$ and the real values p_i are weights (Medgyessy, 1961, p. 1).

Teicher (1960, 1961, and 1963) and Robbins (1948 and 1949) give a more rigorous approach to mixtures of distributions.

---

[1] $B(x)$ is said to be a degenerate distribution function if $B(x)=0$ when $x \leq s$ (constant) and if $B(x) = 1$ when $x > s$. Its graph consists of a single step of height one at $x = s$. A probability distribution is degenerate if its members are : one when $x = s$ (integer) and 0 (otherwise); its graph consists of a single point of height one at $x = s$.

DEVELOPMENT OF FUNDAMENTAL DISTRIBUTIONS OF MIXTURES

Feller (1943) gives the following definitions of contagion:

Definition:

True contagion. Each "favorable" event has a direct effect of increasing or decreasing the probability of some future event happening.

Definition:

Apparent contagion. Inhomogeneity in populations where the events are independent of each other.

Definition:

A simple Poisson distribution function with parameter *a* must be of the form

$$\pi(n;a) = e^{-a} \frac{a^n}{n!}, \quad a > 0, \quad n = 0,1,2, \ldots, \tag{4}$$

where *a* gives the expected number of "events."

Definition:

A distribution function P(x) is a compound Poisson distribution if it has the form

$$P(x) = \sum_{n \le x} \int_0^{\infty} e^{-a} \frac{a^n}{n!} \, dH(a), \quad n = 0,1,2, \ldots, \tag{5}$$

where H(a) is the distribution function of a non-negative random variable and *a* is distributed according to the cumulative probability law H(a).

From (5) one can go to the probability function, first considered by Greenwood and Yule (1920),

$$\Pi_n = \int_0^\infty e^{-a} \frac{a^n}{n!} \, dH(a) \quad . \tag{6}$$

Thus, (6) is called the compound Poisson distribution.

If H(a) is a step function, then the probability function of the compound Poisson distribution becomes

$$\Pi_n = \frac{1}{n!} \sum_{i=0}^{\infty} e^{-a_i} a_i^n p_i \quad . \tag{7}$$

By defining Pearson's type III distribution, two special cases of the Polya-Eggenberger distribution can be developed.

<u>Definition</u>:

The Pearson type III distribution is the distribution of the probability density function

$$h(a) = \frac{1}{\Gamma(\frac{g}{d})} \left(\frac{1}{d}\right)^{\frac{g}{d}} (a-c)^{\frac{g}{d}-1} e^{-\frac{1}{d}(a-c)} \quad ; a > c, \frac{g}{d} > 0, \frac{1}{d} > 0 \tag{8}$$

where $\underline{a}$, $\frac{g}{d}$, and $\frac{1}{d}$ are constants.

To obtain a more condensed form of Pearson's type III distribution, let $k = \frac{1}{d}$ and $t = \frac{g}{d}$, then (8) becomes

$$h(a) = \frac{k^t}{\Gamma(t)} (a-c)^{t-1} e^{-k(a-c)} \quad . \tag{9}$$

If C = 0, then h(a) in (9) has the origin of its distribution shifted by the amount (- C) and

$$(10) \qquad h(a) = H'(a) = \begin{cases} \dfrac{K^t}{\Gamma(t)} a^{t-1} e^{-Ka} & , a \geq 0 \\ 0 & , a < 0 \end{cases}$$

where <u>a</u> ranges over the desired values, k and t being constants determined by comparing the resulting compound Poisson distribution with the actual observed data.

The Polya-Eggenberger distribution may be obtained as a special case of (6). Substitute (10) into (6).

$$(11) \qquad \Pi_n = \int_0^\infty e^{-a} \frac{a^n}{n!} \left[\frac{K^t}{\Gamma(t)} a^{t-1} e^{-Ka} \right] da$$

$$(12) \qquad \Pi_n = \frac{1}{n!} \frac{K^t}{\Gamma(t)} \frac{1}{(1+K)(1+K)^{n+t-1}} \int_0^\infty [a(1+K)]^{n+t-1} e^{-a(1+K)} (1+K)\, da$$

which is of the form

$$\Gamma(t) \equiv \int_0^\infty x^{t-1} e^{-x}\, dx \quad .$$

Therefore, (12) becomes

$$(13) \qquad \Pi_n = \frac{K^t}{\Gamma(t)} \frac{1}{n!} \frac{\Gamma(n+t)}{(1+K)^{n+t}} \quad , K > 0 \; , t > 0 \; , n \leq 0, 1, 2, \cdots$$

which is the Polya-Eggenberger distribution.

Neyman's type A distribution can be obtained if a takes on the values bc only, where $c > 0$ is a constant and $b = 0, 1, \ldots$, and if a is distributed according to the Poisson law (Feller, 1943)

$$(14) \quad p_i = \text{Prob}\left\{a = bc\right\} = e^{-\lambda} \frac{\lambda^b}{b!}, \quad \lambda > 0,$$

then (7) becomes

$$\Pi_n = \sum_{b=0}^{\infty} \left[e^{-bc} \frac{(bc)^n}{n!} \right] e^{-\lambda} \frac{\lambda^b}{b!}$$

or

$$(15) \qquad \Pi_n = \frac{c^n}{n!} e^{-\lambda} \sum_{b=0}^{\infty} \frac{b^n}{b!} \left(e^{-c} \lambda \right)^b$$

which is Neyman's contagious distribution of type A.

RELATIONSHIPS BETWEEN DISTRIBUTIONS OF MIXTURES

Wilks (1961) stated that the gamma distribution is a Pearson type III distribution when the given distribution is a probability density function (p.d.f.) of the form

$$h(a) = \frac{a^{\mu-1} e^{-a}}{\Gamma(\mu)}$$

where μ is the mean and corresponds to t in equation (10).

According to Feller (1957), p. 131, the Polya-Eggenberger distribution can be linked to the negative binomial distribution by going through numerous limiting processes on the Polya-Eggenberger distribution. Also, Conover (1962) showed that the negative binomial distribution is a compound Poisson distribution. Gurland (1957) further related the negative binomial to other distributions not mentioned in this paper.

It was shown above that Neyman's contagious distribution of type A can be obtained by letting the parameter a in the compound Poisson step function equal bc only where a is distributed according to the Poisson law.

APPLICATION TO EXPONENTIAL DISTRIBUTIONS AND ESTIMATION OF PARAMETERS BY THE METHOD OF MOMENTS

Rider (1961) has stated that life characteristics of certain types of electronic components (resistors, capacitors, vacuum tubes, etc.) and complex systems of highspeed digital computers are very well described by exponential distributions. In experiments in life testing a probability density function of the following form may be assumed:

$$f(t) = \frac{1}{\theta} e^{-t/\theta} \quad ; \ \theta > 0, \ 0 \leq t < \infty \tag{16}$$

where θ is the mean lifetime between failures.

Suppose two populations of type (16) have been mixed in unknown proportions p and q =(1-p) with parameters θ_1 and θ_2 respectively. Then,

$$f(t) = p\theta_1^{-1} e^{-t/\theta_1} + (1-p)\theta_2^{-1} e^{-t/\theta_2} \tag{17}$$

At this point, the method of moments can be applied to estimate the parameters θ_1, θ_2, and p.

Let m_r' be the rth sample moment about zero. In particular, m_1', m_2', and m_3' are the moments of a random sample from (16). The estimators of p, θ_1, and θ_2 will be obtained by method of moments and denoted p^*, θ_1^*, and θ_2^*. Thus,

(18) $$p^* \theta_1^* + (1-p^*) \theta_2^* = m_1' \quad ,$$

(19) $$p^* \theta_1^{*2} + (1-p^*) \theta_2^* = \frac{1}{2} m_2' \, ,$$

(20) $$p^* \theta_2^{*3} + (1-p^*) \theta_2^{*3} = \frac{1}{6} m_3' \, .$$

From (18)

(21) $$p^* = (m_1' - \theta_2^*)/ (\theta_1^* - \theta_2^*)$$

Upon substituting (21) into (19) and (20), the following equations result:

(22) $$(m_1' - \theta_2^*) (\theta_1^* + \theta_2^*) = \frac{1}{2} m_2' - \theta_2^{*2} \quad ,$$

(23) $$(m_1' - \theta_2^*) (\theta_1^{*2} + \theta_1^* \theta_2^* + \theta_2^{*2}) = \frac{1}{6} m_3' - \theta_2^{*2} .$$

Equation (22) may be solved for θ_i^* (i = 1 or 2), the solution being

(24) $$\theta_i^* = (\frac{1}{2} m_2' - m_1' \theta_j^*) / (m_1' - \theta_j^*) \quad ,$$

where j = 2 or 1 according as i = 1 or 2. If equation (24) is substituted into (23), upon simplification we obtain θ_j^* .

$$(25) \quad 6(2m_1'^2 - m_2')\,\theta_j^{*2} + 2(m_3' - 3m_1'm_2')\,\theta_j^* + 3m_2' - 2m_1'm_3' = 0.$$

Solving this quadratic equation gives two roots θ_1^* and θ_2^*. Which root is designated θ_1^* and which θ_2^* is immaterial. p^* can be obtained by substituting θ_1^* and θ_2^* into (21) and will refer to the component θ_1 ; whereas, $1 - p^*$ will refer to θ_2.

The roots of (25) may not always be positive or even real. If every observation in a sample were equal to some constant $k > 0$, then it follows that $m_1' = k$, $m_2' = k^2$ and $m_3' = k^3$ and (25) is reduced to

$$(26) \quad k^2 \left[6(\theta - \tfrac{1}{3}k)^2 + \tfrac{1}{3}k^2 \right] = 0 ,$$

whose roots are imaginary. This may occur provided we have positive probability as seen from continunity considerations.

If $\theta_1 \neq \theta_2$, then our proposed estimators are consistent and the probability that $\theta_1^* > 0$, $\theta_2^* > 0$, $0 \le p^* \le 1$ approaches 1 as n tends to infinity. This happens because the estimators which are thought of as functions of (m_1', m_2', m_3'), are continuous at the point (μ_1', μ_2', μ_3'), where μ_i' are the population moments, and $\theta_1^* > 0$, $0 \le p^* \le 1$ if (m_1', m_2', m_3') is sufficiently close to (μ_1', μ_2', μ_3').

If $\theta_1 = \theta_2 = \theta$, then the behavior of the estimators change radically. Thus, $\mu_1' = \theta$, $\mu_2' = 2\theta^2$, $\mu_3' = 6\theta^3$; and therefore,

$$2\mu_1'^2 - \mu_2' = \mu_3' - 3\mu_1'\mu_2' = 3\mu_2'^2 - 2\mu_1'\mu_3' = 0$$

Thus, the coefficients in the quadratic equation (25), multiplied by $n^{\frac{1}{2}}$, are normally distributed in the limit as n approaches infinity, with zero means and finite and positive variances. This implies that θ_1^* and θ_2^* have no constant limits in probability and their imaginary parts do not become negligibly small as n increases. In particular, the estimators are not consistent in this case.

The reliability of the estimators θ_1^* and θ_2^* can be tested if it is temporarily assumed p is known. Otherwise, the calculation of the variances of the three estimators is difficult. With p known, only two sample moments are needed to estimate θ_1 and θ_2. Using equations (18) and (19),

$$\text{(27)} \qquad \theta_1^* = m_1' \pm \left(\frac{q}{2p}\right)^{1/2} \left(m_2' - 2m_1'^2\right)^{1/2}$$

where $q = (1 - p)$. If $\theta_1 \geq \theta_2$, use the upper sign; otherwise, use the lower sign if $\theta_1 \leq \theta_2$. One can see the shortcoming of the methods of moments because it is the knowledge about θ_1 and θ_2 that we desire. Knowing whether θ_1 is $>$ or $<$ θ_2 will tell us which pair of estimators is consistent.

Cramer (1958), p. 354, gives a theorem for the asymptotic variance of θ_1^* :

$$\text{(28)} \quad \text{Var } \theta_1^* = \mu_2(m_2')\left(\frac{\partial \theta_1^*}{\partial m_1'}\right)^2 + 2\,\mu_{11}(m_1', m_2')\frac{\partial \theta_1^*}{\partial m_1'}\,\frac{\partial \theta_1^*}{\partial m_2'} + \mu_2(m_2')\left(\frac{\partial \theta_1^*}{\partial m_2'}\right)^2 ;$$

Here $\mu_2(m_1')$ and $\mu_2(m_2')$ are the variances of m_1' and m_2' respectively with μ_{11} (m_1', m_2') the covariance of these two moments. The partial derivatives are to be evaluated at the point

$$\text{(29)} \qquad m_1' = p\,\theta_1 + q\,\theta_2 \qquad , \quad m_2' = 2(p\,\theta_1^2 + q\,\theta_2^2) \quad .$$

Rider (1961) gives formulas for finding the values of the coefficients of the partial derivatives in (29).

$$(30) \qquad \mu_2(m_1') = n^{-1}\left[(2p-p^2)\theta_1^4 - 2pq\theta_1^2\theta_2^2 + (2q-q^2)\theta_2^4\right]$$

$$(31) \qquad \mu_{11}(m_1', m_2') = 2n^{-1}\left[(3p-p^2)\theta_1^3 - pq\theta_1^2\theta_2 - pq\theta_1\theta_2^2 + (3q-q^2)\theta_2^3\right]$$

$$(32) \qquad \mu_2(m_2') = 4n^{-1}\left[(6p-p^2)\theta_1^4 - 2pq\theta_1^2\theta_2^2 + (6q-q^2)\theta_2^4\right]$$

If $\theta_1 > \theta_2$, the partial derivatives needed are

$$(33) \qquad \frac{\partial\theta_1^*}{\partial m_1'} = 1 - \frac{2^{1/2}\, q^{1/2}\, m_1'}{p^{1/2}(m_2' - 2m_1'^2)^{1/2}} \quad ,$$

$$(34) \qquad \frac{\partial\theta_1^*}{\partial m_2'} = \frac{q^{1/2}}{2^{3/2}\, p^{1/2}(m_2' - 2m_1'^2)^{1/2}} \quad .$$

At the point (29), these derivatives have values

$$(35) \qquad \frac{\partial\theta_1^*}{\partial m_1'} = \frac{-\theta_2}{p(\theta_1-\theta_2)} \quad , \quad \frac{\partial\theta_2^*}{\partial m_2'} = \frac{1}{4p(\theta_1-\theta_2)}$$

If $\theta_1 < \theta_2$, change the signs of the fractions on the righthand sides of (33) and (34), and again (35) is obtained.

If one substituted (30), (31), (32), (33) in (27), and simplified, the variance of the asymptotic distribution of $\theta_1^\#$ is

$$(36) \qquad \frac{1}{4np^2(\theta_1-\theta_2)}\left[p(6-p)\theta_1^4 - 4p(3-p)\theta_1^3\theta_2 + 2p(5-3p)\theta_1^2\theta_2^2 - 4p(1-p)\theta_1\theta_3^3 + (1-p^2)\theta_2^4\right] .$$

The variance of the asymptotic distribution of θ_2^* may be obtained by replacing p by q and interchanging θ_1 and θ_2 in (36).

Rider (1961) stated that data should not be assumed to have come from a mixed exponential distribution until it has been determined that they have not come from a single exponential distribution $\frac{1}{\theta} e^{-t/\theta}$. That is, θ of this distribution should be estimated and a chi-square test made to see whether the data conforms to this distribution. Should the hypothesis be rejected, then a mixed exponential population may be assumed.

It is possible that the chi-square test could give a wrong conclusion; whereupon, it would be impossible for the method of moments to estimate θ according to Rider (1961). Should the population be mixed and θ_1 nearly equal to θ_2, it still may be difficult to obtain valid estimates of them. Therefore, for practical purposes, estimators are not recommended until further research reveals some way to correct their shortcomings.

APPLICATION TO EXPONENTIAL DISTRIBUTIONS AND ESTIMATION OF PARAMETERS BY THE METHOD OF MAXIMUM LIKELIHOOD ESTIMATORS

Mendenhall and Hader (1958) attacked the problem of estimating the parameters of a population that was obtained by mixing two exponential failure time distributions in unknown proportions, the population model being based upon censored sampling. After a predetermined length of time had elapsed or after a predetermined number of units had failed, the life test was concluded. It was assumed that each unit of the population conceptually contained a tag indicating the subpopulation to which the unit belonged. Of course, this unit or tag of information is only available after failure has occurred.

The estimation of population parameters will be by the method of maximum likelihood estimators. Consider a population composed of s = 2 subpopulations representing failure types, mixed in proportions p and $q = (1 - p)$, where $0 \le p \le 1$. The test termination time, T, is in units of size T at which time r units have failed, r_i from subpopulation (i) and $r = r_1 + r_2$. The time of failure of the jth unit from subpopulation (i), t_{ij} is observed. It is assumed that j always ranges from j = 1 to r_i when not specified. The (n-r) are the units which have not failed and yield no information as to the subpopulation they were drawn from. The cumulative failure probability distribution from equation (1) assuming G(x) = constant = is

$$F(t) = 1 - e^{-t/\theta} \quad , \quad 0 \le t < \infty$$

or to be more general

$$F_i(t) = 1 - e^{-t/\theta_i} \quad , \quad 0 \le t < \infty$$

where i = 1 or 2.

Now let $x = t/T$ and $\beta_i = \theta_i/T$, then

(37) $$F_i(x) = 1 - e^{-x/\beta_i} \quad , \quad 0 \le x < \infty \quad .$$

If p is the proportion of units belonging to subpopulation $i = 1$, then the cumulative distribution function for the population is

(38) $$F(t) = p F_1(t) + q F_2(t)$$

and the density function,

(39) $$f(t) = p f_1(t) + q f_2(t) \quad .$$

Also let

(40) $$H_i(t) = 1 - F_i(t)$$

and

(41) $$H(t) = 1 - F(t)$$

where $H(t)$ is the probability that a unit will survive to time t.

The probability of r_1 units failing due to cause (1) and r_2 units failing due to cause (2), and $(n-r)$ units surviving at $t=1$, given a random sample of n units, the multinomial

(42) $$P\left\{r_1, r_2, (n-r) \middle| n\right\} = \frac{n!}{r_1!\, r_2!\, (n-r)!} \left[pF_1(1)\right]^{r_1} \left[qF_2(1)\right]^{r_2} \left[H(1)\right]^{(n-r)}$$

The conditional density of obtaining ordered observations, $(x_{i1}, x_{i2}, \ldots , x_{ir_i} \mid r_i)$ and $x_{ij} \leq 1$ is

(43) $$P\left\{X_{i1}, X_{i2}, \ldots, X_{ir_i} \mid r_i \; ; \; X_{ij} \leq 1\right\} = \frac{r_i! \prod_{j=1}^{r_i} f_i(X_{ij})}{\left[F_i(1)\right]^{r_i}}$$

From (42) and (43) the likelihood for this sample becomes

(44) $$L = \frac{n!}{(n-r)!} H(1)^{(n-r)} p^{r_1} q^{r_2} \prod_{j=1}^{r_1} f_1(X_{1j}) \prod_{j=1}^{r_2} f_2(X_{2j})$$

(45) $$\frac{\partial \ln L}{\partial \beta_1} = \frac{K(n-r)}{\beta_1^2} - \frac{r_1}{\beta_1} + \frac{r_1 \bar{X}_1}{\beta_1^2}$$

(46) $$\frac{\partial \ln L}{\partial \beta_2} = \frac{(1-K)(n-r)}{\beta_2^2} - \frac{r_2}{\beta_2} + \frac{r_2 \bar{X}_2}{\beta_2^2}$$

(47) $$\frac{\partial \ln L}{\partial p} = \frac{K(n-r) + r_1}{p} - \frac{(1-K)(n-r) + r_2}{q}$$

where

$$K = \frac{p e^{-1/\beta_1}}{p e^{-1/\beta_1} + q e^{-1/\beta_2}} = \frac{1}{1 + \left(\frac{q}{p}\right) e^{(1/\beta_1 - 1/\beta_2)}} \quad .$$

At time t, the subpopulations would be mixed in the proportions p(t) and 1 - p(t).

(48) $$\therefore p(t) = \frac{p H_1(t)}{H(t)} \quad , \quad p(0) = p \quad .$$

Therefore, k = p(1), the conditional mixture proportion at the test termination time, x = 1.

When (43), (44), and (45) are equated to zero, the estimating equations are

$$(49) \qquad \hat{p} = \frac{r_1}{n} + \hat{K}\frac{(n-r)}{n} \quad ,$$

$$(50) \qquad \hat{\beta}_1 = \bar{X}_1 + \hat{K}\frac{(n-r)}{r_1} \quad ,$$

$$(51) \qquad \hat{\beta}_2 = \bar{X}_2 + \frac{(1-\hat{K})(n-r)}{r_2} \quad ,$$

where

$$(52) \qquad \bar{X}_i = \frac{\sum_{j=1}^{r_i} X_{ij}}{r_i} \quad , \quad \hat{K} = \frac{1}{1 + \left(\frac{\hat{q}}{\hat{p}}\right) e^{(1/\hat{\beta}_1 - 1/\hat{\beta}_2)}} \quad .$$

Now solve (49), (50), (51) and (52) simultaneously to obtain estimates of β_1, β_2, and p. If (49), (50) and (51) are substituted into (52) and (52) is solved for $\hat{k}$, a single equation is obtained

$$\hat{K} = h(\hat{K}) \qquad , \ (0 \leq \hat{K} \leq 1) \ ,$$

where $h(\hat{k})$ is a function of $\hat{k}$. $\hat{k}$ can be obtained by plotting $h(\hat{k}) - \hat{k}$ versus $\hat{k}$ and obtaining the solution where $h(\hat{k}) - \hat{k} = 0$. When $\hat{k} = 0$, $h(\hat{k}) - \hat{k}$ will be positive or zero.

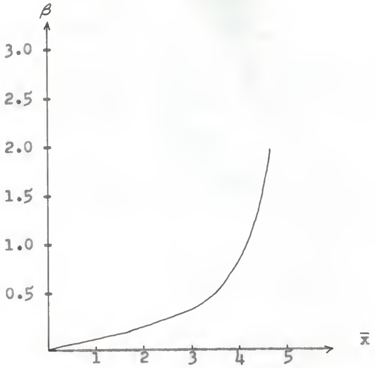

Figure 1. Maximum likelihood estimator of β as a function of $\bar{x}$ based on a sample from a truncated exponential distribution. Measurements expressed in units of truncation time T.

Mendenhall and Hader (1958) give a procedure to find a good approximation of $\hat{k}$ for samples drawn from a single truncated exponential distribution. Actually, it is obtained by making a modification of the maximum likelihood estimate. Where the exponential distribution is assumed to be truncated at time T, then the maximum likelihood estimate of β_1 is a solution of

$$(\beta_i - \bar{x}_i)(e^{-1/\beta_i} - 1) = 1 \qquad . \tag{53}$$

Thus, solutions of $\hat{\beta}_i$ are obtained graphically from figure 1 where $\hat{\beta}_i$ is given as a function of $\bar{x}_i$. Subpopulation (1) is chosen as the smaller $\bar{x}$ and $\hat{\beta}_{10}$ is the corresponding value $\bar{x}$ obtained from figure 1. Now substitute into (50) and solve for $\hat{k}_o$.

(54) $$\hat{\beta}_{10} = \bar{X}_1 + \hat{K}_o \frac{(n-r)}{r_1}$$

Now the quantity $A_o = g(\hat{k}_o) - \hat{k}_o$ can be solved. Thus, a solution to equations (49), (50), (51) and (52) is possible if $A = 0$. Since $g(0) \geq 0$, then the value of $\hat{k}$ which will satisfy $A = 0$, must be $\hat{k} < k_o$ or $\hat{k} > k_o$. It all depends upon whether A_o is negative or positive.

Let

(55) $$W = \frac{\hat{q}}{\hat{p}} e^{\left(1/\hat{\beta}_1 - 1/\hat{\beta}_2\right)} ,$$

(56) $$d\hat{K} = \frac{-dA}{1 + g(\hat{K})^2 \left(\frac{dW}{d\hat{K}}\right)} ,$$

where

(57) $$\frac{dW}{d\hat{K}} = -W(n-r)\left[1/\hat{q} + 1/\hat{p} + 1/r_1\hat{\beta}_1^2 + 1/r_2\hat{\beta}_2^2\right] .$$

Choosing

$$dA = -A_o$$

(58) $$\hat{k}_1 = \hat{k}_o + d\hat{k}_o$$

(59) $$\hat{k}_1 = \hat{k}_o + \frac{A_o}{\left[1 + g(\hat{K}_o)^2 \left(\frac{dW}{d\hat{K}_o}\right)\right]} .$$

Therefore, this iterative process may be repeated until one achieves the degree of accuracy desired. Then the solution for $\hat{k}$ obtained can be substituted into the estimating equations (49), (50) and (51) to find the estimates p, β_1, and β_2.

Should it happen that $r_1 = 0$, then there could be no estimates of β_1 obtained from the estimating equations. Really, this is not a problem in the practical sense because it can reasonably be concluded that β_1 must be very large or else $p_1 = 0$. So, let us adopt the convention that when $\hat{\beta}_1 = \infty$, we shall mean that β_1 is very large when $r_1 = 0$. In experimental work it is desired to choose T and n large enough that the probability that $r_1 = 0$ or $r_2 = 0$ is very small. Actually, one could not expect to get any information on failure parameters if he is not willing enough to test until some failures are observed.

The maximum likelihood procedure appears to give satisfactory results under the following conditions:

1) when the sample size is large. Tukey (1960), p. 463, stated that a statistical problem is usually called a

large sample problem when $1/n$ is so small everything else can be neglected.

2) when the test termination time, T, is large relative to θ_1 and θ_2 because $\beta_i = \theta_i/T$ must be very small to have an efficient estimation $E(r_i)$ proportional to β_i. When n and T are small, the estimates are badly biased and have large variances. Thus, it would seem desirable to investigate other estimators having better small sample properties.

Rao (1948) stated that for higher efficiency the estimates of the parameters must be found by the method of maximum likelihood because this gives rise to the "best" estimates in large samples. Thus, in practice one would not use the method of moments with large samples since it is not as efficient as the method of maximum likelihood.

The relative magnitudes of θ_1 and θ_2 will be known in most experimental situations; hence, the estimating procedure may be modified in a simple way when this is true. This modification will reduce the bias and variances of the estimates.

The disadvantage of maximum likelihood estimates is that the equations leading to them are usually non linear and thus difficult to solve.

ACKNOWLEDGMENT

The author wishes to express his thanks to his Major Professor, Dr. Leslie F. Marcus, Professor of Statistics, for his constructive criticisms and patience in the preparation of this paper. Also, appreciation is expressed to Dr. Gary F. Krause, Assistant Professor of Statistics, and to Dr. John W. Meux, Assistant Professor of Mathematics, for reviewing the manuscript.

REFERENCES

Cassie, R. M.
Some uses of probability paper in the analysis of size frequency distribution. Australian Journal of Marine and Freshwater Research, 1954, 5:513-522.

Conover, W. J.
Contagious distributions. M. A. Dissertation, Catholic University of America, Washington, D. C., 1962. 1-30 p.

Cramer, H.
Mathematical methods of Statistics. Princeton: Princeton University Press, 1958. 354 p.

Davis, D. J.
An analysis of some failure data. J. Amer. Statist. Assoc., 1952, 47:113-150.

Feller, W.
On a general class of "contagious" distributions. Ann. Math. Statist., 1943, 14:389-400.

Feller, W.
An introduction to probability theory and its applications. New York: John Wiley and Sons, 1962. 131 p.

Greenwood, M. and G. U. Yule
An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. J. Roy. Statist. Soc., 1920, 83:255-279.

Gurland, J.
Some interrelations among compound and generalized distributions. Biometrika, 1957, 44:265-267.

Medgyessy, P.
Decomposition of superpositions of distribution functions. Hungarian Acad. of Sciences, Budapest, 1961. 1 p.

Mendenhall, W. and Hader, R. J.
Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. Biometrika, 1958, 45:504-520.

Neyman, J.
On a new class of "contagious" distributions applicable in entomology and bacteriology. Ann. Math. Statist., 1939, 10:35-57.

Rao, C. R.
The utilization of multiple measurements in problems of biological classification. J. Roy. Statist. Soc. Ser. B, 1948, 10:163-170.

Rider, P. R.
The method of moments applied to a mixture of two exponential distributions. Ann. Math. Statist., 1961, 32:143-147.

Robbins, H.
Mixture of distributions. Ann. Math. Statist., 1948, 19:360-369.

Robbins, H. and E. J. G. Pitman
Application of the method of mixtures to quadratic forms in normal variates. Ann. Math. Statist., 1949, 20:552-553.

Teicher, H.
On the mixtures of distributions. Ann. Math. Statist., 1960, 31:55-73.

Teicher, H.
Identifiability of mixtures. Ann. Math. Statist., 1961, 32:244-248.

Teicher, H.
Identifiability of finite mixtures. Ann. Math. Statist., 1963, 34:1265-1269.

Tukey, J. W.
A survey of sampling from contaminated distributions. Palo Alto: Stanford University Press, 1960, 463 p.

Wilks, S. S.
Mathematical statistics. New York: John Wiley and Sons, 1962, 171 p.

APPENDIX

Numerical Example

Since the method of maximum likelihood estimation is more efficient than the method of moments where the sample size n is very large, the method of maximum likelihood has been chosen to illustrate a numerical problem (Mendenhall and Hader, 1958) to utilize the theory developed in this paper.

In this numerical example it was the policy to remove all ARC-1 VHF communication transmitter-receivers which had operated for 630 hours from the aircraft. Thus, T = 630 hours was the fixed time at which the sample was censored. Units which failed were removed from the aircraft for maintenance. The failure times for the ith population are assumed to have a cumulative failure probability distribution defined by

$$F_i(t) = 1 - \exp -t / \theta_i \quad , \quad 0 \le t < \infty .$$

In some cases the apparent failure of the units were not confirmed and upon arrival at the maintenance center were found to operate satisfactorily. This makes it desirable to estimate that fraction of unconfirmed failures in the population. Thus, the sample of failures may be subdivided into two subpopulations, 1) confirmed failures and 2) unconfirmed failures (shown in tables 1 and 2 respectively).

To identify the two subpopulations let the unconfirmed failures be called subpopulation (1) and the confirmed failures as subpopulation (2). Using tables 1 and 2, one obtains the

Table 1. Confirmed failures. Hours to failure for ARC-1 VHF radio transmitter receivers*

16	224	16	80	128	168	144	176	176	568
392	576	128	56	112	160	384	600	40	416
408	384	256	246	184	440	64	104	168	408
304	16	72	8	88	160	48	168	80	512
208	194	136	224	32	504	40	120	320	48
256	216	168	184	144	224	488	304	40	160
488	120	208	32	112	288	336	256	40	296
60	208	440	104	528	384	264	360	80	96
360	232	40	112	120	32	56	280	104	168
56	72	64	40	480	152	48	56	328	192
168	168	114	280	128	416	392	160	144	208
96	536	400	80	40	112	160	104	224	336
616	224	40	32	192	126	392	288	248	120
328	464	448	616	168	112	448	296	328	56
80	72	56	608	144	408	16	560	144	612
80	16	424	264	256	528	56	256	112	544
552	72	184	240	128	40	600	96	24	184
272	152	328	480	96	296	592	400	8	280
72	168	40	152	488	480	40	576	392	552
112	288	168	352	160	272	320	80	296	248
184	264	96	224	592	176	256	344	360	184
152	208	160	176	72	584	144	176	-	-

Table 2. Unconfirmed failures. Hours to failure for ARC-1 VHF radio transmitter receivers*

368	136	512	136	472	96	144	112	104	104
344	246	72	80	312	24	128	304	16	320
560	168	120	616	24	176	16	24	32	232
32	112	56	184	40	256	160	456	48	24
200	72	168	288	112	80	584	368	272	208
144	208	114	480	114	392	120	48	104	272
64	112	96	64	360	136	168	176	256	112
104	272	320	8	440	224	280	8	56	216
120	256	104	104	8	304	240	88	248	472
304	88	200	392	168	72	40	88	176	216
152	184	400	424	88	152	184	-	-	-

*Tables reproduced from Biometrika 1958, 45(3 and 4): 509.

following data:

$$n = 369 \quad r_1 = 107 \quad r_2 = 218 \quad r = r_1 + r_2 = 325$$

where r_i is the number of units failed in each subpopulation and n is the random sample size.

$$n-r = 369 - 325 = 44 \qquad \bar{x}_1 = \frac{\bar{t}_1}{T} = 0.3034862$$

$$\bar{x}_2 = \frac{\bar{t}_2}{T} = 0.3644677$$

t is the time the unit survived which is a proportional part of the fixed time T at which the life of a unit is censored.

Now one can form the estimating equations by making use of equations (49), (50) and (51).

(60) $\hat{p} = 0.2900 + 0.1192\ \hat{k}$

(61) $\hat{\beta}_1 = 0.3035 + 0.4112\ \hat{k}$

(62) $\hat{\beta}_2 = 0.5663 - 0.2018\ \hat{k}$

One can simplify the process of obtaining an iterative solution by making use of table 3.

The first step of the procedure is to use $\bar{x} = 0.303$ and enter figure 1. Thus, the corresponding estimate of β_1 is $\hat{\beta}_{10} = 0.380$; and the corresponding value of $\hat{k}$, $\hat{k}_o = 0.186$ is obtained by using the estimating equation (61). Knowing k_o, $\hat{\beta}_{20}$ and $\hat{p}_o$, can be easily obtained. Table 3 shows these values in row v = 0.

Table 3. Record of iterations

v	$\hat{k}_v$	$\hat{\beta}_{1v}$	$\hat{\beta}_{2v}$	$\hat{p}_v$	w_v	$g(\hat{k}_v)$	A_v
0	0.186	0.380	0.529	0.312	4.622	0.1779	-0.0081
1	.166	.3718	.5328	.3098	5.024	.1660	.0000
2	.167	.3721	.5326	.3099	5.002	.1666	- .0004
3	.165	.3713	.5330	.3097	5.046	.1654	.0004

Next we compute

(63) $$g(\hat{k}_o) = \frac{1}{1 + \left(\frac{\hat{q}_o}{\hat{p}_o}\right) \exp\left[1/\hat{\beta}_{10} - 1/\hat{\beta}_{20}\right]} \quad ,$$

and $A_o = g(\hat{k}_o) - \hat{k}_o = -0.0081$ follows (Mendenhall and Hader, 1958). The value of $\hat{k}$ will be obtained whenever there exists a solution to the maximum likelihood equations, i.e., $A = 0$. Since A can be either positive or zero when $\hat{k} = 0$ and negative when $\hat{k} = 0.186$, then there exists a solution for $\hat{k}$ when $0 < \hat{k} < 0.186$. Therefore the value $\hat{k}$ has when $i = 1$ must be less than 0.186. By use of equation (56) we can calculate the change in $\hat{k}$, $d\hat{k}_o$.

(64) $$d\hat{k}_o = \frac{A_o}{1 + g(\hat{k}_o)^2 \left(\frac{dw_o}{d\hat{k}_o}\right)} = \frac{(-0.0081)}{1 + (0.1779)^2\ (-19.04)}$$

$$= -0.02.$$

Thus, $$\hat{k}_1 = \hat{k}_o + d\hat{k}_o = 0.186 - 0.02 = 0.166 \quad ,$$

$$\hat{\beta}_{11} = 0.3718, \qquad \hat{\beta}_{21} = 0.5328, \qquad \hat{p}_1 = 0.3098.$$

Since the values of these estimating parameters make $A_1 = 0.0000$, then essentially these estimates are the maximum likelihood estimates of the parameters. If a bound is desired on the iteration error, then it can be obtained by calculating A for $\hat{k}_2 = 0.167$ and $\hat{k}_3 = 0.165$. If the solution for $\hat{k}$ is taken to be 0.166 and since $A_2 = -0.0004$ is negative and $A_3 = 0.0004$ is positive, then the absolute value of the iterative error for $\hat{k}$ is less than 0.001.

The estimate of the unconfirmed failures is $\hat{p} = 0.3098$ and one can find their estimated average life to be

$$\hat{\theta}_1 = \hat{\beta}_1 T = (0.3718)(630) = 234.2 \text{ hours.}$$

The estimate of the average life for the confirmed failures is

$$\hat{\theta}_2 = \hat{\beta}_2 T = (0.5328)(630) = 335.7 \text{ hours.}$$

Although the accuracy above was good for the estimation equations when only one iteration was used, it may be that one or two iterations on the maximum likelihood equations may not be sufficient. Therefore, one should put bounds on the iterative error as was done in this numerical example.

It can be seen that the estimates of the average life of units from two subpopulations, θ_1 and θ_2, can be quite useful in maintaining excellent service in the communications field as one can anticipate maintenance requirements before any actual break down in service results. Thus, one can see the importance and use of estimation theory.

MIXTURES OF DISTRIBUTIONS

by

RAY LEE MILLER, JR.

B. S., University of Missouri, 1957

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1964

Mixtures of distributions occur when there is overlapping of two or more distributions so that it is difficult to separate them into their respective components. Feller has distinguished two types of these distributions, true contagion and apparent contagion.

Karl Pearson attacked the problem of mixed frequency distributions in the early 20th century. About 25 years later Greenwood and Yule developed a very general scheme for contagious events but this proved to be too complex for most practical applications. Several years later Polya and Eggenberger considered a special model of true contagion which was the simplest case of the general Greenwood-Yule scheme. The next significant advancement in the study of contagious distributions was in 1939 when Neyman applied them to entomology and bacteriology. Since that time Neyman's distributions have been generalized by several workers among whom were Feller and Gurland.

Several definitions of distributions have been cited and developed. Some of these were the compound Poisson, Neyman's contagious distribution of type A, Polya-Eggenberger, and Pearson's type III distributions. Also, a theorem of mixtures of distributions was stated and proved.

Applications to exponential distributions and estimation of parameters were done by methods of moments and methods of maximum likelihood. Since the method of maximum likelihood estimation is more efficient than the method of moments where the sample size n is very large, a numerical example illustrating the use of the method of maximum likelihood was applied to an exponential distribution.