LEVEL IV

PURDUE UNIVERSITY

DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

(12)

(6)

ON MIXTURES OF DISTRIBUTIONS: A SURVEY AND SOME NEW RESULTS ON RANKING AND SELECTION.

by

(10) Shanti S. Gupta, Purdue University

and

Wen-Tao Huang, Academia Sinica, Taiwan

(15) N00014-75-C-0455

(14) MMS-79-22

Department of Statistics
Division of Mathematical Sciences
(9) Mimeograph Series #79-22

(11) Aug[redacted]79

(12) 69

291 730

ON MIXTURES OF DISTRIBUTIONS: A SURVEY AND SOME NEW RESULTS ON RANKING AND SELECTION*

by

Shanti S. Gupta, Purdue University

and

Wen-Tao Huang, Academia Sinica, Taiwan

0. Introduction and Summary

There is a large body of literature on the mixture of distributions going over about the last eighty years. Since Pearson [84] considered the estimation of the parameters of the mixture of two normal densities in 1894, many more papers have appeared related to the problem of statistical inference about the parameters of mixture and probabilistic properties of mixture densities. In 1960, Teicher [120] started the study of general considerations of identifiability of mixtures of distributions. Since then the interest in the mathematical aspects of mixtures has received an increasing amount of attention, and the approach to the statistical inference of mixtures has also seen more development. Recently, the studies of mixtures and related topics in statistics and probability have developed even more so, that these can be classified as a new area. For this reason, the present authors decided to review (survey) some of the literature dealing with some aspects of this area which seemed important to them. The topics covered relate to probabilistic properties, estimation, hypotheses testing, and multiple decision (selection and ranking) procedures.

*This research was supported by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University.

The applications of mixtures of distributions can be found in many fields such as ecology, taxonomy, fishery, biology, plant and animal breeding, psychology and engineering, etc. In biology it is often desired to measure certain characteristics in natural populations of some particular species. The distribution of such characteristics may vary markedly with age of the individuals. Age is difficult to ascertain in samples from populations. In such cases the biologist observing the population as a whole is dealing with a mixture of distributions, the mixing in this case is done over a parameter depending on the unobservable variate "age". In agriculture remotely sensed unlabelled observations from several crops are available and sometimes along with some labelled observations informatior is also available about the distribution of individual crop population. On the basis of such information one wishes to estimate the acreage of a particular crop or all crops as proportion of the total acreage.

In statistical applications of mixtures, the mixture of densities can be used to approximate some parameter(s) associated with a density. For example, the coefficient of skewness of Fisher's transformation $z = \frac{1}{2} \log (\frac{1+r}{1-r})$ of the correlation coefficient decreases more rapidly than the excess of its kurtosis when the sample size increases. The usual normal approximation for its distribution can be improved by mixing it with a logistic distribution. The resulting mixture approximation which can be used to estimate the probabilities and the percentiles, compares favorably in both accuracy and simplicity (see [78]).

In this paper we restrict ourselves to probabilistic properties, estimation, hypotheses testing and multiple decisions. In Section 1 we review these main results concerning probabilistic properties of mixing distributions including the identifiability, scale mixture, infinite divisibility, atomicness and perfectness. In Section 2 we survey results on estimation theory which include the method of moments, method of maximum likelihood estimation, method of least squares, Bayesian estimation method, and method of curve fitting. For the hypotheses testing problem, we give those results which provide tests for hypothesis whether an observed sample is mixed from two samples with certain unknown proportion; we also give those results which test if the mean of the mixture population is equal to some known value. All these are treated in Section 3. And finally in the last section (Section 4) we study some selection problems of mixture populations. We use the subset selection formulation when the sample size is small and also study the case of large sample using the indifference zone approach.

At the end of the paper we have given a reasonably comprehensive and useful bibliography concerning the topics discussed in this paper and also the topic of experimental designs. This last topic is not discussed in this paper and hence the papers dealing with it are marked with a * in the bibliography.

1. Probabilistic Properties

Let $G(\theta)$ be a cumulative distribution function. Let $F(x,\theta)$ be a cumulative distriubtion function in x for each θ on the support of G. Assume $F(x,\theta)$ is Borel measurable in θ for every x. Then, $H_G(x)$ defined by $H_G(x) = \int_{-\infty}^{\infty} F(x,\theta)dG(\theta)$ is a distribution function, which is called a G-mixture of F and G is referred to as a mixing distribution. When G is a discrete distribution, $H_G(x)$ is called a finite mixture and G is referred to as a finite mixing distribution. Let the domain of θ be denoted by Θ and $\sigma(\Theta)$ denote a σ-algebra of Θ such that each point of Θ is contained in $\sigma(\Theta)$. Let $\mathcal{G}$ denote a class of mixing distributions on $(\Theta, \sigma(\Theta))$. Let $\mathcal{H}$ denote the class of all G-mixture of F for all

$G \in \mathcal{G}$. Let M denote a mapping from $\mathcal{G}$ to $\mathcal{H}$ such that for each $G \in \mathcal{G}$, $M(G) = \int_{-\infty}^{\infty} F(x,\theta)dG(\theta)$. Class $\mathcal{H}$ is called identifiable if M is one-to-one so that one can identify some unique mixing distribution G_0 when a certain $H_0 \in \mathcal{H}$ is given.

1A. Identifiability of Mixtures

Some basic properties of mixture was studied by Robbins in 1948 [95]. Teicher [120] extended and generalized this work. Teicher [121] initiated the study of identifiability problem. In [121], location and scale parameter mixtures are considered, i.e. when θ is, respectively, the location and the scale parameter of $F(x,\theta)$. Sufficient conditions for the identifiabilities of $\mathcal{H}$ when θ is, respectively, the location and scale parameter, are given. It is also shown by Teicher [121] that $\mathcal{H}$ is identifiable if $\{F(x,\theta), \theta \in \Theta\}$ is an additively closed family, i.e. $F(x,\theta_1)*F(x,\theta_2) = F(x,\theta_1+\theta_2)$, the operation is the usual convolution. In [122] necessary and sufficient conditions for identifiability of finite mixtures are given. Important distributions such as normal and gamma are shown to be identifiable under finite mixing. Some sufficient conditions are also given for the class of binomial distributions to be identifiable.

These results are largely extended by Yakowitz and Spragins [127]. They consider the general case of p-dimensional distributions. Using the methods of linear algebra, the authors obtain a necessary and sufficient condition for identifiability of finite mixtures. This condition is very useful since it is easy to check. They conclude that the family of p-dimensional Gaussian distributions, the family of Cauchy distributions, the family of non-degenerate negative binomial distributions, the family of products of n exponential distributions (for fixed integer n), and the union of the family of p-variate

Gaussian and the family of products of n exponential distributions are all identifiable. Using a result given by [122], Mohanty [76] showed that the finite mixture of Laguerre distributions is also identifiable. Chandra [14] has proved some results given by Teicher [122] and Yakowitz and Spragins [127] by some other methods. Recently, Blum and Susarla [9] gave a short and clear set of equivalent conditions for identifiability. Let $A = \{F(x,\cdot): x \in R\}$. Denote $C_0(\Theta)$ the Banach space of continuous functions on Θ which vanish at ∞ and the norm is given by the sup norm. Blum and Susala [9] showed that if $A \subset C_0(\Theta)$, then $\mathcal{H}$ is identifiable if, and only if A generates $C_0(\Theta)$ in the sup norm.

1B. Scale Mixtures

When the mixture is defined in the form $H_G(x) = \int_0^\infty F(\frac{x}{\theta})dG(\theta)$, the mixture is called the scale mixture. This kind of mixture has special interest both in probability theory and statistics. It is easy to see that the density and the associated characteristic function of $H_G(x)$ can be written, respectively, as

$$h_G(x) = \int_0^\infty \frac{1}{\theta}f(\frac{x}{\theta})dG(\theta), \quad \varphi_H(t) = \int_0^\infty \varphi_f(t\theta)dG(\theta).$$

In terms of random variables, we denote them by $Z =_d XY$ ($=_d$ means equality in distribution) where X, Y and Z are, respectively, associated with $F_X(x)$, $G_Y(\theta)$ and $H_Z(x)$. It is interesting to note that the class of scale mixtures is closed under the operation of scale mixing, i.e. if $F \in \mathcal{S}$, the class of scale mixtures, then $H \in \mathcal{S}$ where $H(x) = \int_0^\infty F(\frac{x}{\theta})dG(\theta)$ where $G(\theta)$ is some distribution function on $(0,\infty)$. Define $\mathcal{H}_X = \{H_Z: Z =_d xy\}$. Then, we have for $a \geq 0$, $0 \leq p \leq 1$, $F_1, F_2 \in \mathcal{H}_X \Rightarrow p\,F_1(ax) + (1-p)F_2(ax) \in \mathcal{H}_X$. The conditions for the identifiability in the case of scale mixture can be put in another form in terms of moment conditions which is given by Keilson and Steutel [64] as follows. If $X \neq 0$ a.s. and $E|X|^{\epsilon} < \infty$ for some $\epsilon > 0$ and if $E|Z|^{\epsilon} < \infty$, then $E\,Y^{\epsilon} < \infty$ and there exists one-to-one correspondence between Z and Y ($\mathscr{M}$ and $\mathscr{G}$). Now, if we assume X to be a normal with mean 0, or, the kernel of the mixture, i.e. $F(x,\theta)$ is the normal distribution function with mean 0; we can characterize the class of mixtures. Let $\mathscr{M}(\Phi)$ denote the class of scale mixtures (variance mixtures) of normal distribution with mean 0. From the Bernstein's representation theorem for completely monotone functions (see [37 p. 415]) we can conclude that $f \in \mathscr{M}(\Phi)$ if, and only if, $\varphi_f(t)$, the characteristic function of f, is an even function and $\varphi_f(\sqrt{t})$ is completely monotone on $(0,\infty)$. We recall that $h(x)$ is completely monotone on $(0,\infty)$ if $(-1)^n h^{(n)}(x) \geq 0$ for $x > 0$ and $n = 0,1,2,\ldots$. Accordingly, by checking the conditions, it can be seen that the Cauchy distributions, the Laplace distributions, student t-distributions and the symmetric stable distributions are all in the class $\mathscr{M}(\Phi)$. This was obtained by Kelker [66]. Also, logistic and double exponential distributions belong to $\mathscr{M}(\Phi)$ ([1]). To characterize $\mathscr{M}(\Phi)$ in another type, we restate a result of Schoenberger [104] as follows: $f \in \mathscr{M}(\Phi)$ if, and only if there exists a function Ψ such that $\varphi_f(t) = \Psi(t^2)$ and for $\underset{\sim}{t}' = (t_1, t_2, \ldots, t_p)$, $\varphi_f(t) = \Psi(|\underset{\sim}{t}|^2)$, a p-dimensional characteristic function for each p ($p = 1,2,\ldots$). It was shown in [64] that $\mathscr{M}(\Phi)$ is closed under the multiplication of densities with suitable renorming if the product is integrable. If Z has density $f_Z(z)$ which is symmetric about 0, Mohanty [76] showed that a necessary and sufficient condition for $Z =_d N\,Y$, where N denotes the zero mean normal random variable, is that for some k $(\frac{-d}{dz})^k f_z(\sqrt{z}) \geq 0$, for $z > 0$. He also found some special correspondence between Z and Y. If $f_Z(z) = e^{-z}[\frac{1}{1+e^{-z}}]^2$ is logistic, then $G_Y(y) = 2\sum_{1}^{\infty}(-1)^{k-1}k^2 y^{-3}\exp(-k/2y^2)$ i.e. $\frac{1}{2}Y$ is the asymptotic distribution of the well-known Kolmogorov's goodness of fit statistic. This result is useful for Monte Carlo studies. It was also found that if Z is double exponent, then $\frac{1}{2} Y^2$ is exponential. Finally, it is interesting to ask how broad is the classs $\mathcal{M}(\Phi)$. For given $\alpha_1, \alpha_2 (\frac{1}{2} < \alpha_1 < \alpha_2)$, let $F \in \mathcal{M}(\Phi)$ with $F(x_1) = \alpha_1$ and $F(x_2) = \alpha_2$. Let $\mathcal{M}(\Phi;\ x_1,\ x_2,\ \alpha_1,\ \alpha_2) = \{H(x):\ H(x_1) = \alpha_1,\ H(x_2) = \alpha_2,\ H(x) \in \mathcal{M}(\Phi)\}$. Then, Efron and Olshen [35] showed that, there exists an $F^* \in \mathcal{M}(\Phi;\ x_1,\ x_2,\ \alpha_1,\ \alpha_2)$ such that $F^*(x') = \max H(x')$ for $x' \in (x_1,\ x_2)$ and $F^*(x'') = \min H(x'')$ for $x'' \notin (x_1,\ x_2)$ where the maximum and minimum are over the set $H(\Phi;\ x_1,\ x_2,\ \alpha_1,\ \alpha_2)$.

If X is considered to be a gamma distribution of order α $(0 < \alpha < \infty)$, we denote the class of mixture by $\mathcal{M}_G(\alpha)$. Then, we note that $\mathcal{M}_G(\alpha) \subset \mathcal{M}_G(\beta)$ for $\alpha < \beta$. When $\alpha = 1$, $\mathcal{M}_G(1)$ is the mixture of exponential density and for $f_Z \in \mathcal{M}_G(1)$, $f_Z(x)$ is completely monotone. $\mathcal{M}_G(1)$ plays a key role in stochastic processes reversible in time. Kingman [67] showed that any density $\mathcal{F}(x)$ on $(0,\infty)$ can be approximated arbitrarily closely by a finite mixture of exponential densities and this mixture is in $\mathcal{M}_G(1)$. Let $\mathcal{M}_G^S(\alpha)$ denote the class of mixtures of $\Gamma_\alpha + \Gamma_\alpha^D$, where Γ_α denotes the gamma of order α and Γ_α^D denotes the dual of Γ_α. Then, for $\alpha < \beta$, $\mathcal{M}_G^S(\alpha) \subset \mathcal{M}_G^S(\beta)$ and $\lim_{\alpha\to\infty} \mathcal{M}_G^S(\alpha) = \mathcal{M}(\Phi)$ (see [64]).

Another important property concerning mixture is the infinite divisibility. We recall that a random variable X is infinitely divisible (i.d.) if, for any positive integer n, there exist independently identically distributed random variables $X_1, X_2, \ldots, X_n$ such that $X =_d X_1 + X_2 + \ldots + X_n$. It has been shown [81] that in many families of i.d. distribution functions, the property of i.d. is preserved under the operation of mixing. Furthermore, for certain families, this property still holds even when mixing and convolution are applied repeatedly. To find such a class, define $\mathcal{L}_0$ to be a set of all real positive characteristic functions that are log-convex on $(0,\infty)$. Then it is shown in [64] that $\mathcal{L}_0$ is closed under (a) mixing (b) raising to a positive power (c) scaling (d) multiplication and thus (e) any combination of (a), (b), (c) and (d).

For the scale mixing of normal distributions, Kelker [66] showed that if the mixing distribution is non-degenerate and finite, i.e. $G(b) = 1$ for some finite b, then $H_G(x)$ is not i.d.. On the other hand, we note that $H_G(x)$ is i.d. if $G(x)$ is i.d. (see [64]).

Following the notation $\mathcal{M}_G(1)$ introduced earlier, that is, the class of mixtures with mixand a standard exponential density, it is shown [64] that each element in $\mathcal{M}_G(1)$ is i.d. Also, each element in $\mathcal{M}_G^S(1)$ is i.d.

Now we consider the power mixture. Let $\varphi_X(t)$ be an infinitely divisible characteristic function. We define $\phi_Z(t) = \int_0^\infty \{\varphi_X(t)\}^\theta dG_Y(\theta)$ as the power mixture of $\varphi_X(t)$ (or equivalently X). Then, it is easily seen that $\phi_Z(t)$ is i.d. and the class $\mathcal{M}(X)$, the set of all power mixtures of $\varphi_X(t)$, is closed under mixing and convolution (see [64]).

It is interesting to note that all scale mixture of Cauchy distributions are i.d. (see [66]). For the scale mixing Steutel [116] characterizes a big class which are i.d.. We state it as follows. If $\varphi(t)$ is i.d., then $\psi(\theta) = \frac{\theta}{\theta - \log \varphi(\theta)}$ is an i.d. characteristic function for $\theta > 0$ and $\phi_Z(t) = \int_0^\infty \psi(\theta) dG(\theta)$ is also i.d..

1C. Some Other Properties

Let $(X, \mathcal{A}, P)$ be a probability space. A set $A \in \mathcal{A}$ is called atom of P if for each $B \subset A$ such that $B \in \mathcal{A}$ either $P(B) = 0$ or $P(B) = P(A)$. P is called atomic if each positive measurable set contains an atom. P is non-atomic otherwise. The atomic or non-atomic properties of the mixture measures are not always preserved. In [90] an example was given where the mixture of a non-atomic measure is atomic. However, on a real line or a subset of a real line Ω, if the probability measure $P(x,\cdot)$ is non-atomic for each x, then the mixture is always non-atomic for any probability measure $G(\theta)$. For other more general cases, three sufficient conditions were given in [90] for the non-atomicness of the mixture when the mixand measure is non-atomic.

The perfectness of a probability measure was first discussed in the book [42]. This concerns the approximation of a measurable set by a closed or open set. For a probability space $(X, \mathcal{A}, P)$, P is called perfect if for every $\mathcal{A}$-measurable real-valued function f on X and every subset S of the real line for which $f^{-1}(S) \in \mathcal{A}$, there is a linear Borel set $T \subset S$ such that $P(f^{-1}(S)) = P(f^{-1}(T))$. To check the perfectness of P, Sozonov [101] showed that P is perfect if, and only if, for each $\mathcal{A}$-measurable real-valued function f on X there is a linear Borel set $A_f \subset f(X)$ such that $P(f^{-1}(A_f)) = 1$. Accordingly, it is easy to see that a discrete measure is perfect. We call a mixture measure perfect or non-perfect according to whether the mixing measure is perfect or not. Rodine [97] conjectured that perfect mixtures of perfect measures are perfect. It was shown to be false by Ramachandran [89]. However, it is true that the perfect mixtures of discrete measures (thus perfect) are perfect (see [89]). In general, perfect mixture of non-perfect measures can be perfect. The perfectness of the mixture and mixand measures does not guarantee the perfectness of the mixing measure (see [89]).

2. Estimation

Let $H(x) = \int_\Omega F(x,\alpha)dG(\alpha)$ be the mixture distribution. If $G(\alpha)$ is discrete, then $H(x)$ is given by $H(x) = \sum_{i=0}^{\infty} \theta_i F(x,\alpha_i)$. When the summation is finite, $H(x)$ is called finite mixture. In this section, we study the problem of estimating $G(\alpha)$ based on independent observations from $H(x)$. However, for the most part we will discuss the case of finite mixtures. For the case of finite mixtures, the study is then for the estimation of

θ_i and α_i. The methods for estimation can be classified as the method of moments, method of maximum-likelihood estimation, the minimum square method, Bayesian estimation method and the method of curve fitting. In this problem, all mixture distributions are assumed to be identifiable so that the estimations of parameters make sense. Some important classes of continuous and discrete distributions which are identifiable have been mentioned in Section 1A.

2A. Method of Moments

In 1894 K. Pearson [84] studied the dissection of asymptotic and symmetric frequency curves into two components of normal frequency distributions. This may be the earliest paper that investigated the estimation of parameters in the finite mixture case by the use of the method of moments. Let $\Phi(x, \mu, \sigma^2)$ denote the normal cdf with mean μ and variance σ^2. The mixture is given by $H(x) = \alpha \Phi (x, \mu_1, \sigma_1^2) + (1-\alpha) \Phi (x, \mu_2, \sigma_2^2)$. K. Pearson [84] computed the first five moments and by equating the population moments to the sample moments he obtained a nonic (9th degree) equation. Solving for these equations he finally obtained the estimates for α, μ_1, σ_1, μ_2 and σ_2. However, the estimates are not unique. He used the data of 1000 crabs from Naples. For study of the frequency distributions of the breadth of forehead of crabs, assuming the crabs were from two different species, he considered the ratio of the forehead to the body-length as the abscissae of the curve. Applying the method he developed, he arrived at two sets of solutions. This lack of uniqueness of solutions bothered Pearson and he suggested choosing the set of estimates which resulted in the closest agreement between the sixth central moment of the sample and the corresponding moment of the mixture which are supposed to be fitted. Charlier in 1906 [16] suggested a somewhat simpler but still laborious, solution of the moment equations involving a cubic and the ratio of two other polynomials. Burrau in 1934 [13] computed certain functions of the moments which are expressed in terms of the five parameters to be estimated. In the same year Strömgren [117] computed some tables and charts to aid calculation of solutions of some equations which are derived using some given function of moments. Again in the same year of 1934, Pollard [86] considered the dissection of a symmetric density into three components of normal density. Under some assumption Pollard was able to reduce eight parameters to five. Since the density is assumed symmetric so that odd moments are zero. Since five equations are needed for the five unknown parameters, the first eight moments are computed. Pearson's solution [84] are not applicable in this case. However, the development is analogous.

Instead of moment equations, one might expect the application of techniques involving iteration for maximum likelihood equations. This has been done, in fact, by Rao [91] for special case $\sigma_1 = \sigma_2$. This assumption simplifies the problem considerably. However, the calculations involved are still quite cumbersome.

In 1967, Cohen [21] again derived the nonic equation which was first obtained by Pearson. Cohen considerably reduced the total computational effort otherwise required. Some special cases considered by Cohen are $\sigma_1 = \sigma_2 = \sigma$ or $\theta_1 = \theta_2 = \theta$. Some conditional maximum likelihood and conditional chi-square estimates were also discussed. An example was provided to illustrate the procedure proposed for the estimates. However, the problem of lack of uniqueness of solutions still remained. Another solution to the example given by Cohen [21] was provided by Hawkins [50].

In general, multiple solutions for the estimate of parameters are possible. When multiple solutions occur, either solution would be the one of interest and should be examined with an eventual choice of a preferred solution in mind. And when a clear decision can not be made on the basis of any tests, a larger sample should be taken if conditions so permit. Even if some tests are possible, the confidence of conclusion of the estimates are far weakened. Having multiple solutions for estimate is one of the shortcomings for the application of the method of moments.

Later Rao [92] considered the same problem for the special case of equal variances and his results led to a simple set of equations having a unique solution. Rao's method was later programmed for computer's use by Hasselbled [48] and was found to work very well.

Gregor [43] based on the idea of Doetsch [30] as provided by Medgyessi [73] constructed an algorithm which can be used to find the mean of each component with the aid of a Fourier transformation of the given density function. The method of reduction of variances was utilized to determine the unknown variance and frequencies of the components (using the continued fraction approximation for the error function). To test the goodness of fit Kolmogorov-Smirnov test statistics were used.

Day [27] considered the estimate of the proportion of mixture α by the method of moments when each component is a multivariate normal with common variance matrix. For the univariate case, some simulation results showed the estimate behave reasonably nearly as well as maximum likelihood estimate. However, when the dimensionality of the component is larger, the estimates appear poor.

John [56] considered a related but different model of problem. It was assumed that the sample of size N was the result of mixing a random sample of size N_1 from a p-variate normal population with mean $\underset{\sim}{\mu}_1$ and the covariance matrix $\underset{\sim}{\Sigma}$ with an independent random sample of size N_2 from another p-variate normal population with mean $\underset{\sim}{\mu}_2$ and covariance matrix $\underset{\sim}{\Sigma}$. It was desired to estimate N_1, N_2, μ_1, μ_2 and $\underset{\sim}{\Sigma}$. The method of moments was considered for the case $p = 1$. It has been shown that in this case there is an unique of the solution for the estimates. The same method proposed can be applied to the general case of $p > 1$. Asymptotic normality of the moment estimates was also studies by John [56]. For $p = 2$, an example was worked out using the proposed method.

When the components are other than normal, Mendenhall and Hader [75] considered the exponential populations. Rider [93] also considered the same case with less restrictions. He derived the estimates by the method of moments. It was shown that the estimate obtained were consistent. However, it is not clear that the estimates always exist. Cohen [20] considered the cases of mixture of the Poisson distributions and a mixture of one Poisson and one binomial. In the former case, he considered the estimates based on the first two sample moments and the zero sample frequency. Again, he considered the mixture of truncated Poisson distributions with missing zero classes. For the latter case, he used the technique of factorial moments. As the author pointed out, in practice, the more difficult and most important problem is to determine which components are appropriate to fit the data. Rider [94] also considered the case of Poisson mixture, and computed asymptotic variances. When the components are binomial, Blischke [5] used the technique of factorial moments to obtain some relations among moments and parameters. First three moments were computed to obtain three equations so that a unique solution is possible for three unknown parameters. However, the estimates obtained by Blischke [5] have the unpleasant property of assuming complex as well as indeterminate values with positive probability, though this probability tends to zero as sample size increases to infinity. He also showed that the moment estimates $\hat{\underset{\sim}{\theta}} \equiv (\hat{p}_1, \hat{p}_2, \hat{\alpha})$ are asymptotically normal and consistent. Blischke also considered asymptotic relative efficiency (ARE) of the moment estimates $\hat{\underset{\sim}{\theta}} = (\hat{p}_1, \hat{p}_2, \hat{\alpha})$. The ARE of $\hat{\underset{\sim}{\theta}}$ is defined as the ratio of $\sigma^2_{\theta*}/\sigma^2_{\hat{\theta}}$ where $\sigma^2_{\theta*}$ is the Cramér-Rao lower bound of $\theta*$ which is the maximum likelihood estimate. When the components of the estimates $\hat{\underset{\sim}{\theta}}$ are considered jointly, a joint asymptotic relative efficiency (JARE) of $\hat{\underset{\sim}{\theta}}$ relative to the maximum likelihood estimate $\theta*$ was also considered defined by the square of the ratio of the areas of the ellipse of concentration of the respective asymptotic normal distributions. It was proved that the joint asymptotic efficiency is given by $\det(\Sigma_{\theta*})/\det(\Sigma_{\hat{\theta}})$ where $\Sigma_{\hat{\theta}}$ is the convariance matrix of $\hat{\underset{\sim}{\theta}}$. For some special values of p_1, p_2 and α, Blischke [5] computed both ARE and JARE and it was found that neither ARE nor JARE are monotone with respect to n. However, for the limiting case, they always attain the value 1.

When the number of binomial components is larger than 2, Blischke [7] considered a general case of r binomial components with 2r-1 parameters to be estimated. He also applied the method of moments to obtain the first estimate. Then, he considered another efficient estimate based on the moment estimates. This construction of alternative estimate was made at the suggestion of Le Cam [69]. By Neyman's linearization technique BAN estimates were also constructed. Asymptotic relative efficiency and joint asymptotic relative efficiency of the moment estimates were discussed by Blischke [7]. A numerical example for the comparisons of the method of moment and other two alternative estimates was given.

The results for the mixture of r binomials can also be obtained for a number of other distributions. For example, they are applicable to mixtures on p (with known k) of negative binomialy and hence to its special cases, the Pascal and geometric distributions. As regards other cases, Bliss and Fisher [8] , Shenton and Wallington [107] and Katti and Gurland [62] have discussed the negative binomial which is a compound Poisson distribution. Sprott [115] and Katti and Gurland [61] discussed the case of the Poisson-binomial distributions which is the Poisson mixture of parameter n of binomial. The case of the Poisson-negative binomial was studied by Katti and Gurland [60]. For the Neyman contagious distributions (see [80]) Shenton [105] discussed efficiency of the moment estimates. And for a two parameter beta-distribution mixture on parameter p of binomial which is the so-called negative hypergeometric by Shenton [106] the moment estimates were studied by Skellan [108]. Mosimann [77] studied the mixture of multinomials. Falls [36] considered a mixture distribution of two Weibull distribution each with different scale and different shape parameter. Moment estimates were proposed and some graphical illustration and a numerical example were given by Falls [36]. For some other details reference should be made to Blischke [5] and Isaenko and Urbakh [55].

Moment estimates are usually not considered very efficient except for some cases such as the normal, binomial and Poisson distributions. Methods more efficient such as the method of maximum likelihood are more desirable. However, in many cases, such as for example when more unknown parameters need to be estimated, the maximum likelihood equations are found complicated and almost intractable. Under this situation, one may still consider the moment estimates.

For some further studies on the efficiency of moment estimates reference should be made to [105], [106], [115], [5], [7], [39], [48], [118] and [51].

2B. Methods of Maximum Likelihood

In many cases, maximum likelihood estimates are considered to be more efficient than the moment estimates. For the problem of estimating of parameters in the distribution of mixtures most authors treated it by the method of moments in the early years. In 1966, Hasselblad first considered the estimation problem by the method of maximum likelihood. The population from which we sample obeys a density function which is a mixture of k normal densities. Taking logarithms of the likelihood functions and differentiating with respect to each parameters μ_i (mean), σ_i^2 (normal variance) and α_i (mixture proportion) $i = 1, 2, \ldots, k$, and equating them to zero Hasselblad [48] obtained 3k-1 independent equations with 3k-1 unknown parameters. By substitution of some equal quantity in some equation into another equation, he obtained the first iteration scheme. A rough estimate from the truncation method is used as an initial guess for this scheme. The idea of the generalized steepest descent method proposed by Goldstein was applied. It can be shown that the direction traveled by the procedure at each iteration possesses a positive inner product with respect to the gradient. For an alternative treatment of the 3k-1 equations, Hasselblad [48] applied the Newton iteration method, and finally he obtained a matrix equation of an iteration scheme. The investigation of the variances of the estimates are important. Hasselblad [48] gave the explicit formula for the second partials of logarithms of the likelihood-function and from these, the information matrix and thus the variance-covariance matrix of the estimates was approximated. Some details of the asymptotic variances of the estimates of the means, proportions and standard deviations were also given. However, it should be pointed out that the solutions are limited to grouped data in which all class intervals are of equal width. And, in practice, these results obtained would not be likely to show satisfactory unless some conditions should be met, for example, grouping intervals should be narrow, a large sample must be available, and when k = 2, it is desired to have sample size 1000 or more and when k is large, even larger sample sizes are needed. When the separation between component means are insufficient and unable to obtain k distinct sample modes, the estimates obtained are very likely to be unreliable.

For the same problem, Behboodian [4] showed that the maximum likelihood estimates for the component mean μ_i and component variance σ_i^2 are, in fact, respectively, a weighted sample mean $\hat{\mu}_i = \sum_{j=1}^{n} \hat{w}_{ij} X_j$ and the weighted sample variance $\hat{\sigma}_i^2 = \Sigma\, \hat{w}_{ij} X_j^2 - \hat{\mu}_i^2$ for $i = 1, 2, \ldots, k$ where $\hat{w}_{ij}$ are the values of w_{ij} obtained by replacing μ_i, σ_i^2 and α_i by $\hat{\mu}_i$, $\hat{\sigma}_i^2$ and $\hat{\alpha}_i$ and w_{ij} satisfies $w_{ij} = f_i(x_j)/nf(x_j)$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n$. Furthermore, w_{ij}'s satisfy

$$\sum_{j=1}^{k} \hat{\alpha}_j \hat{w}_{ji} = \frac{1}{n} \qquad (i = 1, 2, \ldots, n) \qquad \text{and}$$

$$\sum_{j=1}^{n} \hat{w}_{ij} = 1 \qquad (i = 1, 2, \ldots, k),$$

where $f_i(x)$ and $f(x)$ are, respectively, the densities of ith component and the mixture distribution. In fact, these also have been obtained by Wolfe [126]. He considered the case of multivariate normal density $f_i(\underset{\sim}{x}, \theta_i)$ for each component and he introduced the so-called "probability of membership" of a vector $\underset{\sim}{x}$ in type i which is defined as $P(i|\underset{\sim}{x}) = \dfrac{\alpha_i f_i(\underset{\sim}{x}, \underset{\sim}{\theta}_i)}{f(\underset{\sim}{x})}$ where $f(\underset{\sim}{x})$ is the mixture density. He, furthermore, obtained that the ML of $\hat{\alpha}_i$, $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are given by

$$\hat{\alpha}_i = \frac{1}{n} \sum_{j=1}^{n} \hat{P}(i|\underset{\sim}{x}_j), \quad \hat{\mu}_{ij} = \frac{1}{n\,\hat{\alpha}_i} \sum_{r=1}^{n} \hat{P}(i|\underset{\sim}{x}_r)\, X_{rj} \quad \text{and}$$

$$\hat{\sigma}_{ij}(s) = \frac{1}{n\,\hat{\alpha}_i} \sum_{r=1}^{n} \hat{P}(s|x_r)\,(X_{ir}-\hat{\mu}_{si})\,(X_{jr}-\hat{\mu}_{sj})$$

where μ_{ij} and x_{ij} are the j-th component of $\underset{\sim}{\mu}$ and $\underset{\sim}{x}_i$ and $\sigma_{ij}(s)$ is the (i,j) element in Σ_s which is the covariance of the s-th component. These results are more general than that of Behboodian [4]. It is obvious that $\hat{w}_{ij}$ are the functions of observations $x_1, x_2, \ldots, x_n$. To solve for $\hat{w}_{ij}$, one has to solve the simultaneous functional equations which are rather complicated. However, the relations among $\hat{\mu}_i$, $\hat{\sigma}_i^2$ and $\hat{\alpha}_i$ are given which are useful for the computations of some quantities when some other quantities are obtained.

In 1969, Day [27] considered the mixture of two p-multivariate normal populations with equal covariance matrix Σ. There are $\frac{1}{2}p^2 + \frac{5}{2}p + 1$ unknown parameters which are to be estimated. As usual, taking logarithms and differentiating in turn with respect to each unknown parameter and equating to zero, a set of equations are obtained. By introducing the quantity $P(i|\underset{\sim}{x}_j)$, the probability that observation $\underset{\sim}{x}_j$ comes from the component i, Day was able to express the maximum likelihood estimates of unknown parameters in terms of the estimates of $P(i|\underset{\sim}{x}_j)$, denoted by $\hat{P}(i|\underset{\sim}{x}_j)$ which can be simply expressed in terms of some quantities which are functions of $\hat{\alpha}$ and the estimated Mahalanobis distance in terms of the maximum likelihood estimates. Finally, an iterative scheme was set up. If the initial guesses are close to the real values satisfying the scheme, it can be shown that the sequences generated by the iterative process converge to the solutions. However, solutions may not be unique. For example, when $p \geq 3$, and the Mahalanobis distance between the two components $\Delta^2 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\ \Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'$ is small, say less than 2 and the sample size is small, the solutions are nearly multiple. In this situation, one has to check up at each local maximum to determine where the over-all maximum lies. And this is some shortcoming. By repeating the iterative process from enough different starting points, all the local maxima can be found. However, the maximum likelihood estimates are invariant under linear transformation. This property is helpful for the simulation study. These estimates are, of course, asymptotically mimimum variance unbiased for $\Delta > 0$. Instead of estimating the mean and variance, it seems more interesting to estimate the generalized distance Δ. The asymptotic variance of $\hat{\Delta}$ is given by $r(\Delta)/n$ where $\{r(\Delta)^{-1}\} = E\{(\frac{\partial \log f(x)}{\partial \Delta})\}^2$. When Δ is small, Day showed that $(r(\Delta)^{-1}) = \frac{3}{2}\alpha^2(1-\alpha^2)(1-2\alpha)^2\Delta^4 + O(\Delta^6)$ ignoring the correlation of $\hat{\alpha}$ and $\hat{\Delta}$. When Δ is large, $(r(\Delta)^{-1})$ is approximated by $\alpha(1-\alpha)(1+2\alpha(1-\alpha)\Delta^2)(1+\alpha(1-\alpha)\alpha^2)^{-2}$. For more than 2 components, it is proposed that the analogous iterative process can be developed.

When the component multivariate densities $f_i(\underset{\sim}{x})$ are all specified, there are k-1 proportion parameters which remain unknown and need to be estimated. Peters and Coberly [85] gave a necessary condition that if $\hat{\underset{\sim}{\alpha}}$ is a maximum likelihood estimate (MLE) then $\hat{\underset{\sim}{\alpha}}$ satisfies a fixed point equation $\hat{\underset{\sim}{\alpha}} = G(\hat{\underset{\sim}{\alpha}})$ where for componentwise $kG(\alpha_i)$ is the sum of the ratios of each component of density to the density of the mixture. In order to to find this fixed point, some properties of $\hat{\underset{\sim}{\alpha}}$ and G were found. It was shown that G is a local contraction at $\hat{\underset{\sim}{\alpha}}$ if the rank of $M = (f_i(\underset{\sim}{x}_j))_{n\times k}$ is k and $\hat{\underset{\sim}{\alpha}}$ is a MLE and is an interior point. In fact, if $\hat{\underset{\sim}{\beta}}$ is an interior point such that $\hat{\underset{\sim}{\beta}} = \lim G^n(\underset{\sim}{\beta})$, then $\hat{\underset{\sim}{\beta}}$ is a MLE. When $k = 2$, and β is an interior point in its domain, $G^n(\beta)$ converges to the MLE. It should be pointed out that the fixed point $\hat{\beta}$ satisfying $\hat{\beta} = G(\hat{\beta})$ is not unique. A method is suggested to choose a starting point which is based on the maximum-likelihood classifier. An example was used to show the iterations needed for the accuracy of 0.5×10^{-i} $(i = 2, 3, 4)$ starting from 7 different points. For the accuracy 0.5×10^{-4}, the iterations for the worst case never exceed 70.

For a finite mixture of k exponential families with r unknown parameters in each component density, there are rk + k - 1 parameters including the k-1 unknown proportions to be estimated. Hasselblad [49] derived a set of equations for the successive substitutions iteration scheme. For a practical computation, an initial estimates are necessary and three methods for these estimates are proposed. However, one of them is the initial guess. This can often be made by the mode of the sample or other information obtained directly from data. It was found that the initial estimates is relatively unimportant as long as it is in the admissible range. For some special distributions such as Poisson, binomial, and exponential, exact iterative procedures were given and a numerical example for each case was provided. Asymptotic variances for the Poisson example were derived. For the binomial case, with k = 2, the moment estimates proposed by Blischke [7] was applied to the same data given in the example, and some comparisons between the MLE and the moment estimates were made. It was found that the MLE estimate are superior than the moment estimates in some sense for the small sample study of size 100. The MLE always lies in the admissible range whenever the initial guess is in the same range which is not the case for the moment estimates. Also the variance of the MLE is smaller than that of the moment estimates. However, the asymptotic variance may be very large if the sub-populations are not well separated. Therefore sample sizes of 1000 or more are always desirable for the MLE. It can be expected that the moment estimates may be very bad when sample sizes are small. Day [27] has shown that when the components are multivariate normal, the moment estimates are essentially useless.

The joint asymptotic relative efficiency comparisons in [15] and [118] show that the MLE are much more efficient than the moment estimates, especially, when $\Delta \equiv |\mu_2 - \mu_1|/\min(\sigma_1, \sigma_2) \leq 2$. Hosmer [53] used Monte Carlo simulation to study the MLE for $\Delta \leq 3$ with $\sigma_1 \neq \sigma_2$ and with relatively small sample size $n \leq 300$. This is interesting because both [49] and [27] suggested large sample size as strongly desirable, especially, when the two components are not well-separated. Using the iterative procedure proposed in [48], Hosmer used a stopping time $N = i$ whenever $|L(\underset{\sim}{\phi}^{(i+1)}) - L(\underset{\sim}{\phi}^{(i)})| \leq 10^{-4}$ and took $\hat{\underset{\sim}{\phi}} = \underset{\sim}{\phi}^{(i+1)}$. Otherwise, he suggested $N = 999$ with $N \geq 10$. In the preceding L is the likelihood function, $\underset{\sim}{\phi} = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$, and $\underset{\sim}{\phi}^{(0)}$ is the initial estimate. There is a strong indication that the initial guess $\underset{\sim}{\phi}^{(0)}$ does not seem to have much effect on the MLE $\hat{\underset{\sim}{\phi}}$. With sample size $n = 100$, and $\underset{\sim}{\phi}^{(0)} = (0.3, 0, 1, 1, 1.5)$, for each of 10 different samples, $\hat{\underset{\sim}{\phi}}$ was computed using three quite different initial guesses. In 7 of the 10 samples the values of $\hat{\underset{\sim}{\phi}}$ obtained by starting with the three different guesses were the same and in two other samples 2 of the 3 initial gueses concluded the same $\hat{\underset{\sim}{\phi}}$. The three values of $\hat{\underset{\sim}{\phi}}$ were significantly different in only one sample. For the true parameters $\underset{\sim}{\phi} = (0.3, 0, 1, \mu_2, 1.5)$, $\mu_2 = 1, 2, 3$, simulations for the MLE obtained from 10 samples of size 100 and for true parameters $\underset{\sim}{\phi} = (0.3, 0, 1, 3, 1.5)$, simulations for MLE obtained from 10 samples of size 300 indicate that the MLE may not be accurate enough to provide useful estimates. Hence, the poor behavior of the estimates of the parameters for these examples considered shows that the MLE, though much more efficient than the moment estimates and perhaps the best method available, may be still highly unsatisfactory for even the moderate sample sizes.

The main difficulty in the problems of estimation of mixture is that the data are mixed. When two components are not well separated, some of the data can be from either component with high probability. If the data can be identified the component of origin or when the data contain information about the mixing proportion, the problem may be easier, and, the sample size may be reduced and the estimates still give the same information for the unknown parameter. For this interesting conjecture, Hosmer [53] did the study by using the Monte Carlo method. First, he classified the data into three types. The first type data is mixed and it is called model 0 (M0) sample. A sample where the component of origin of each observation is known with certainty will be called known data. Two types of known data are possible according to whether or not the known data contains information about the mixing proportion. A sample which contains both mixed and known data and where the known data contains no information about the mixing proportion will be referred to as a model 1 (M1) sample, as for example, in the case when 20 male fish and 20 female fish are arbitrarily taken. A model 2 (M2) sample will be referred to the case when the sample contains both mixed and known data, and information about the mixing proportion is contained in the relative number of observations from the two components in the known data. An example of M2 sample would be the case where 100 fish are taken and then the fish are classified as male and female. Let n denote the sample size of M0 sample and let m denote the sample size of M1 or M2 sample. Let the proportion of m to n be denoted by $r = \frac{m}{n}$. The intent in considering M1 or M2 samples is that one needs only a small amount of known data to improve on the M0 sample. The Monte Carlo study followed the same assumptions given in [53] which have been mentioned above except that $\sigma_2 = 2.25$ instead of $\sigma_2 = 1.5$. In this study r was restricted to be 0.1, 0.2 and 0.3 with each value of n and $\underset{\sim}{\phi}$. For given n, M0 sample was generated as a mixed sample. The known samples for the M1 sample were generated by starting with the first observation generated for the mixed sample and noting the population of origin of each observation successively until exactly rn/2 were obtained from each component. These observations became the known sample and the remaining n(1-r) observations the mixed sample. The known samples for M2 sample were constructed by noting the population of origin of the first nr observations for the mixed sample. The observations from the first component formed one known sample and the observations from the second component formed the other known sample. For $n = 100$ and $\underset{\sim}{\phi} = (0.5, 0, 1, 1, 2.25)$, 10 samples were generated and the M0, M1 and M2 estimates were computed from each sample. The mean, variance and mean squared error of these estimates were tabulated. The cases for $n = 100$, and for $\mu_2 = 2$, $\mu_2 = 3$ and for $n = 300$, $\mu_2 = 3$, respectively, were also tabulated. From these Monte Carlo results, it is noted that for most parameters, and for various sample sizes considered and the different values of the ratio r, the M1 and M2 estimates tend to have smaller variance and mean squared error than those of M0. The variances and the mean square errors of M1 and M2 estimates tend to decrease as r increases. When $n = 100$ and $r = 0.1$, the M1 estimates seem to have smaller variances and mean square errors than those of M2. It is found that the estimates obtained using both the mixed and known data were more accurate than those computed from the saml1 samples. The conjecture that, if the components are not well separated and if part of the mixed sample can be correctly classified or if the mixed sample can be supplemented by a small sample of known data, the estimates would be more accurate, was supported by the Monte Carlo results.

As another direction for the study of statistical properties of the estimates for the parameters in the mixture density, Tubbs and Coberly [125] did the study of the so-called sensitivity of the estimates for the mixing proportions. They considered the three bivariate normal mixture and applied the Monte Carlo method. When the original data from each components were shifted (in location and direction), the variations of the estimates for the mixing proportion suggested that the estimates were sensitive. Four kinds of estimate were considered. They were MLE, moment estimate (ME) minimum chi-square estimate, (MCE), and the classification estimate (CE), the last being simply the proportion of the sample which is classified into the ith class by the maximum-likelihood classifier. Mean square errors for each kind of estimates were plotted in [125]. It is interesting to note (based on the Monte Carlo result) that the ordering of the four estimates, according to the degree of sensivity, would be (CE, MLE) $\geq$ MCE $\geq$ ME. However, it is also apparent that the particular type of shift deviation from the model would result in a different ordering. Hence, if the suspected deviation is known to be of one particular type or direction, a specialized experiment should be done to investigate the sensitivity under that alternative.

2C. Method of Least Squares, Bayesian Approach and Some Other Methods

It is known from previous sections that samples of small size do not, in fact, provide good solutions either for method of moments or for method of maximum likelihood. Besides, the computations of estimates using either of these methods are cumbersome and some difficulties such as lack of uniqueness may occur. Therefore, it is desired to study some other methods for estimation. Most of results described in this section are restricted to the estimation of mixing distribution. In 1968 Choi and Bulgren [19] considered the case of estimating the mixing distribution when the component densities are completely specified. Let $H_n(x)$ denote the empirical distribution associated with the observed sample $\underset{\sim}{x}$ of size n. If the mixing proportion $\underset{\sim}{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ is used they considered the integral squared errors given by $S_n(\underset{\sim}{\alpha}) \equiv \int (H_\alpha(x) - H_n(x))^2 dH_n(x)$ where $H_\alpha(x)$ is the cdf of the mixture associated with $\underset{\sim}{\alpha}$. In fact, they considered the case of finite mixture and showed that there exists the solution $\hat{\underset{\sim}{\alpha}}(= \hat{\underset{\sim}{\alpha}}_n(\underset{\sim}{x}))$ which minimizes $S_n(\underset{\sim}{\alpha})$ for all $\underset{\sim}{\alpha}$ in the admissible domain. This $\hat{\underset{\sim}{\alpha}}$ is then used as the estimate of the mixing proportion. It has been shown that $\hat{\underset{\sim}{\alpha}}$ converges to the true unknown value of $\underset{\sim}{\alpha}$ with probability one if continuity conditions are assumed for H in θ_i (parameter in mixand density) and α_i $(i = 1, 2, \ldots, k)$. Furthermore, asymptotic normality is also shown for the estimate $\hat{\underset{\sim}{\alpha}}$ if non-singularity condition holds for the matrix $(E(H(x,\theta_i)H(x, \theta_j)))$, $i,j = 1, 2, \ldots, k$. Rate of convergence of $\hat{\underset{\sim}{\alpha}}$ is shown to be $0(\ln n/\sqrt{n})$ for all $n \geq n_0$ with probability one. These asymptotic properties are very helpful for the study of the estimates. In 1969, Choi [18] considered the case of estimating the mixing proportion and unknown parameters in the component densities when the functional form of the component distribution is specified. He used the same criterion of the integral squared errors. The same optimal asymptotic properties are shown to hold if some other extra conditions on the first and second derivatives of $H(x,\alpha)$ with respect to α_i $(i = 1,2,\ldots,k)$ are satisfied. It should be noted that the parameters to be estimated in this situation are given by $G = (\alpha_1, \alpha_2, \ldots, \alpha_k, \theta_1, \theta_2, \ldots, \theta_k)$ when θ_i's are the parameters in the ith component distribution. Some Monte Carlo studies are made in [19]. Each component is assumed to be a univariate normal density with common variance 1. The number of components ranges from 2 to 5. Sample size ranged between 10 and 400. Simulations were repeated 500 times and mean square errors were computed. It was found that mean square errors are small when sample sizes are at least as large as 200 and the mean square errors were not largely effected by the number of components. The result of Choi [18] can be extended to the case of continuous mixing distribution by taking a sequence of distributions as its approximation. The criterion of errors considered by [19] and [18] in fact can be extended to become $\int (H_\alpha(x) - H_n(x)) dW(x)$ where $W(x)$ is some weight function. As Bartlett and MacDonald [2] have studied, a good choice of $W(x)$ is not easy. The special case $k = 2$ has been studied in [2] and for $k \geq 3$, the situation is quite complicated. The criterion of errors considered in [19] is, in fact, the Cramér-Von Mises type or Wolfowitz distance between two sample functions. If this distance is defined to be the Kolmogorov type $\sup_x |H_\alpha(x) - H_n(x)|$, then the solutions $\hat{\alpha}(= \hat{\alpha}_n(x))$ to minimize this distance have been considered by Deely and Druse [28]. This paper is related to the empirical Bayes approach of Robbins [95]. They considered the problem of estimating the general mixing distribution $G(\alpha)$ by choosing a sequence of discrete didtributions $\{G_n(\alpha)\}$, where for each n, $G_n(\alpha)$ depends on the sample X_n of size n, such that $G_n(\alpha)$ converges weakly to $G(\alpha)$ with probability one. For each n, an admissible $\hat{G}_n(\alpha)$ is chosen so that the minimum of the uniform distance between $H_{G_n(\alpha)}(x)$ and $H_n(x)$ is attained. For each sample size n, a sequence $\{\hat{G}_n(\alpha)\}$ is obtained to approximate the real $G(\alpha)$. Under some mild conditions, it has been shown that $\hat{G}_n(\alpha) \to G(\alpha)$ at any continuity point of G with probbility one. The existance of such $\hat{G}_n(\alpha)$ for each n is guaranteed and its computation involves a linear programming problem. To be more general, suppose d is any metric for the topology of weak convergence of probabilities on the sample space (see Parthsarthy [83]). Let $\mathcal{G}$ denote the set of all mixing distribution function $G(\alpha)$ defined on Ω, the parameter space. For the topology of weak convergence, suppose $\mathscr{G}$ is compact and for a sequence $\{\mathscr{G}_i\}_1^\infty$ of subclasses of $\mathscr{G}$ satisfies $\overline{\bigcup_1^\infty \mathscr{G}_i} = \mathscr{G}$. If $\hat{G}_n(\alpha)$ is so chosen such that for each n, $\hat{G}_n(\alpha) \in \mathscr{G}_n$ and $d(H_{\hat{G}_n(\alpha)}, H_n)$ attains its infimum for all $G_n(\alpha) \in \mathscr{G}_n$, then it is shown [14] that $\hat{G}_n(\alpha)$ converges weakly to $G(\alpha)$ with probability one if $F(x,\alpha)$ is continuous with respect to α. The results in [28] can thus be obtained by taking some special metric satisfying some conditions. Some other conditions for the weak convergence of $\hat{G}_n(\alpha)$ have also been studied in [14]. Using another approach, Blum and Susarla [9] considered a partition of parameter space Ω. A step function G_n is constructed such that on each division of the partition, the constant value is given according to some weight which are controlled by the local maximum and minimum values of the mixture density on this division. When the mixture density $h_G(\cdot)$ is unknown, an estimate $\hat{h}_n(\cdot) \pm \epsilon_n$ $(= \hat{h}_n(\cdot, x_1, x_2, \ldots, x_n))$ satisfying $\sup_x |\hat{h}_n(x) - h_G(x)| \to 0$ a.s. is used to replace $h_G(\cdot)$. If some conditions similar to continuity in both x and θ are satisfied by the component density $f(x,\theta)$, then the weak convergence of $\hat{G}_n$ to the real mixing distribution $G(x)$ holds almost surely. Furthermore, when θ is a location or scale parameter, it has been shown that $|h_{\hat{G}_n}(x) - h_G(x)| \to 0$ a.s. and $E(h_{\hat{G}_n}(x) - h_G(x))^2 = 0(n^{-c_1})$ where $c_1 = \min(2c, 1-2c)$ for some constant c satisfying $\epsilon_n = n^{-c}$. The construction of $\hat{G}_n$ is possible by linear programming though not simple. One question may be raised how the partition of Ω is taken so that for practical application, the convergence of $\hat{G}_n$ would be more reasonable. Comparison with methods given by [28] and [18] the fundamental property of the weak convergence of the $\hat{G}_n$ are all satisfied. However, the computational feasibility of the Choi's method [18] is not clearly established.

If the observations from the mixture population are restricted to the positive integer value, Rolph [98] first considered Bayes estimation of $G(\alpha)$. Some assumptions were made by Rolph. Ω is a finite interval and considering $f(x,\alpha)$ as a function of α, say, $q_x(\alpha)$ for a fixed x, $q_x(\alpha) = \sum_{i=0}^{\infty} a_i(x)\alpha^i$ (In fact, continuity of q in α is sufficient). Then, the unconditional mass function (mixture mass function) can be expressed as a summation of sequence of ith moment of $G(\alpha)$. Properly putting some prior distribution of the set $\mathcal{G}$ of distributions defined on Ω, consider the Bayes estimate $\hat{G}$ which minimizes the risk associated with some loss function $L(\hat{G}, G)$. Under some conditions, the Bayes estimate of G is just the expectation of the posterior distribution. The Bayes estimate $\hat{G}$ is thus determined by taking the distribution with $(\hat{m}_1, \hat{m}_2, \ldots)$ as its moments where each $\hat{m}_i$ is the expected ith moment under the posterior distribution. Consider the loss function of the form $\sum_1^{\infty} \gamma_i(m_i(\hat{G}) - m_i(G)^2)$ where $m_i(G)$ is the ith moment of G. Suppose $\bar{G}_t$ and $\underline{G}_t$ are the two boundaries where distributions having $(\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_{t-1})$ as their moments then, the estimate $\hat{G}_n$ is defined as the convex combination of $\bar{G}_t$ and $\underline{G}_t$. It has been shown that the sequence $\{\hat{G}_n\}$ $(\hat{G}_n = \hat{G}_n(x_1, x_2, \ldots, x_n))$ is consistent. Relaxing the restriction to Ω being a half-line, Meeden [74] chose the prior distribution on the set of distribution on Ω in another way. Using some results of [98] Meeden [74] was able to show that the Bayes estimate under his set-up was consistent. These mathematical constructions and proof are complete; however, the practical computation for the estimate is not so easy and clearcut and still needs more investigation.

Properties of consistency or weak convergence are important and desirable and fundamental for our study of estimation of mixing distribution.

The above properties may not hold when sample size is small. Paying attention to the small sample property, Boes [11] considered the possibility of some estimates to attain the Cramér-Rao bound. Restricting to the case of finite mixture, he obtained the necessary and sufficient conditions for the attainment of the Cramér-Rao lower bound for the parameter α when $k = 2$. A uniformly minimum variance estimator of α was obtained which was also shown to be consistent [11]. When $k \geq 3$, some jointly efficient estimates were obtained by Boes [11]. By an estimate $\hat{\underset{\sim}{\theta}}(\underset{\sim}{x}) = (\hat{\theta}_1(\underset{\sim}{x}), \hat{\theta}_2(\underset{\sim}{x}), \ldots, \hat{\theta}_k(\underset{\sim}{x}))$ jointly efficient for $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ in some set U, we mean the ellipsoid of concentration of $\hat{\underset{\sim}{\theta}}(\underset{\sim}{x})$ centered at $\underset{\sim}{\theta}$, coincides with the minimum ellipsoid of concentration. Again, by considering the risk defined by $R(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}) = \sum_1^{k-1} a_i \operatorname{Var} \hat{\theta}_i$, where $\hat{\underset{\sim}{\theta}} \in U$ = set of all unbiased estimate of $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots \theta_k)$ and for some constants $a_1, a_2, \ldots a_{k-1}$. Then, it is obvious that $R(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}) \geq \sum_1^{k-1} a_i I^{ii}(\underset{\sim}{\theta})$ where $(I^{ij}(\underset{\sim}{\theta})) = (I_{ij}(\underset{\sim}{\theta}))^{-1}$ and $I_{ij}(\underset{\sim}{\theta}) = E[(\frac{\partial}{\partial\theta_i} \ln h)(\frac{\partial}{\partial\theta_j} \ln h)]$ and where h denotes the likelihood function. Denoting $L(\underset{\sim}{\theta}) = \sum_1^k a_i I^{ii}(\underset{\sim}{\theta})$, by $\underset{\sim}{\theta}^0$- efficient estimate of $\hat{\theta}$, we mean $R(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta}^0) = L(\underset{\sim}{\theta})$. Let $\Omega^0 = \{\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$: $\theta_i \geq 0, \sum_1^{k-1} \theta_i \leq 1\}$. Boes [12] has shown that if $\underset{\sim}{\theta}^*$ is a point in Ω^0 for which $L(\underset{\sim}{\theta})$ attains its maximum, then the $\underset{\sim}{\theta}^*$-efficient estimate $\hat{\underset{\sim}{\theta}}(\underset{\sim}{\theta}^*) = (\hat{\theta}_1(\underset{\sim}{\theta}^*), \hat{\theta}_2(\underset{\sim}{\theta}^*), \ldots, \hat{\theta}_k(\underset{\sim}{\theta}^*))$ is a minimax unbiased estimate for $\underset{\sim}{\theta}$ in the sense that $\sup_{\theta} R(\hat{\underset{\sim}{\theta}}(\underset{\sim}{\theta}^*), \underset{\sim}{\theta}) \leq \sup_{\underset{\sim}{\theta}} R(\hat{\underset{\sim}{\theta}}, \underset{\sim}{\theta})\ \forall\ \hat{\underset{\sim}{\theta}} \in U$. This is a very desirable result if such a minimax unbiased estimate can be found. Some examples were given by Boes. It is interesting to mention an example to see the simplicity of the estimate. If $k = 3$ and each component is uniform such that $f_1 = \frac{1}{2}$ in $(0,2)$, $f_2 = \frac{1}{2}$ in $(2,4)$ and $f_3 = \frac{1}{2}$ in $(1,3)$. For some (any) constants a_i it is seen that the minimax unbiased estimate is given by $(\hat{\theta}_1, \hat{\theta}_2) = (\frac{1}{2} + \frac{N_1-N_3}{n}, \frac{1}{2} + \frac{N_4-N_2}{n})$ where N_i = number of observations falling in $(i-1, i]$, $i = 1, 2, 3, 4$.

Finally, by the approach of curve fitting, Preston [88] proposed to fit the mixing distribution by piece-wise polynomial arcs. Here it is assumed that each component density is discrete. The estimations given in [28], [98] and [74] are all step function approximations to the mixing distribution. Hence, polynomial approximation would be more proferable and accurate if the approximations are appropriate. Let $\hat{G}(\alpha)$ denote the polynomial approximation of $G(\alpha)$. Preston [88] considered the estimate of form $\hat{G}(\alpha) = \sum_{i=1}^{m} \sum_{j=0}^{r} a_{ij} \ \ell_{ij}(\alpha)$, where

$$\ell_{ij}(\alpha) = \begin{cases} 0 & \alpha < \beta_i \\ [(\alpha-\beta_i)/(\beta_{i+1}-\beta_i)]^j & \beta_i \leq \alpha < \beta_{i+1} \\ 1 & \beta_{i+1} \leq \lambda \end{cases}$$

$(\ell_{i0}(\beta_i) = 1)$. $\{\alpha_{ij}\}$ are sequence of parameters and β_i are constants. Hence $\hat{G}(\alpha)$ is a polynomial of degree r. Denoting $L_{ij}(x) = \sum_{s\leq x} \int F(s,\alpha)d\ell_{ij}(\alpha)$, we have $\hat{H}(x) = \sum_{i=1}^{m} \sum_{j=0}^{r} \alpha_{ij} \ L_{ij}(x)$. Hence, if $\hat{G}(\alpha)$ is an estimate of $G(\alpha)$, $\hat{H}(x)$ should be an estimate of $H(x)$. Using the observed sample to form an empirical distribution function $H_n(x)$ as another estimate of $H(x)$, the parameters $\{\alpha_{ij}\}$ to be determined are thus so chosen that $\hat{H}(x)$ is as close to $H_n(x)$ as possible. Take the Kolmogorov type of criterion, $D(H_n,\hat{H}) \equiv \max_x |H_n(x)-\hat{H}(x)|$. $\{\alpha_{ij}\}$ are chosen that $D(H_n, \hat{H})$ is minimized subject to the constraint that $\hat{G}(\alpha)$ is a distribution function. Some special case that $\hat{G}(\alpha)$ is a step function, piece-wise linear, piece-wise quadratic, have been discussed. To study the goodness of the estimate $\hat{G}(\alpha)$ for $G(\alpha)$, a criterion $K(\varphi) = \sum_x \{\varphi(x) - \varphi_G(x)\} h(x)$ is defined. It is shown that $\hat{G}(\alpha)$ is good from an empirical Bayes point of view if $E(K(\varphi_{\hat{G}}))$ (the expectation is taken with respec to random sample) is small. Some numerical examples are studied and D and K are computed. However, for the practical and general purposes, a good choice of location of β_i is not clearly established. It is also obvious that if $H_n(x)$ is not close to $H(x)$, the estimate $\hat{G}(\alpha)$ would also be unreliable. Asymptotic properties of $\hat{G}(\alpha)$ are not given though it may be consistent or weakly convergent.

3. Testing Statistical Hypothesis

Most papers concerning the inferences about mixture densities are related to the estimation of parameters. In practical situation, it is desirable to know whether an observed sample is from a population which is a mixture of two known populations. Generally, we may be interested in knowing whether the distribution function of one population is a mixture of the distribution functions of the other two populations. This kind of inference is quite different to that of estimation. On the other hand, we may, sometimes, wish to know whether the mean of a mixture population is equal to some known values. This is the standard hypothesis testing problem.

Thomas [124] in 1969 considered the problem whether one population is a mixture of two other populations. Let the three populations be denoted, respectively, by π_1, π_2, and π_3 and the associated cdf be denoted by $F_1(x)$, $F_2(x)$ and $F_3(x)$. Let the nth random observation from π_i be denoted by X_{in} $(i = 1, 2, 3)$. Let R_i denote the rank of X_{i1} in the sample $\{X_{11}, X_{21}, X_{31}\}$. We will denote X_{i1} by X_i when there is no confusion. Thomas

[124] proposed a 0-1 valued statistic t which is defined by

$$t(R_1,R_2,R_3) = \begin{cases} 0 & \text{if } (R_1,R_2,R_3) \text{ is an even permutation of } (1,2,3), \\ 1 & \text{otherwise.} \end{cases}$$

It has been shown that if π_3 is really a mixture of π_1 and π_2, then $E(t) = \frac{1}{2}$. It was pointed out that, in fact, the mixture can be extended to $k (k \geq 3)$ components and with the same definition, the result holds. Suppose n_1 samples, n_2 samples and n_3 samples are drawn respectively from π_1, π_2, and π_3. Define a symmetrized U statistic by

$$t_n^* = \frac{1}{n_1 n_2 n_3} \sum_{i,j,k} t(R_{1i}, R_{2j}, R_{3k}) \tag{3.1}$$

where the summation is over all possible values of i, j and k and $n = \min(n_1, n_2, n_3)$. Then, t_n^* is asymptotically normal. In fact, it has been shown [124] that $(t_n^* - \frac{1}{2})\sqrt{n} \xrightarrow{\mathcal{L}} \Phi(0,1)$, the standard normal, if $F_1 \neq F_2$. Hence, t_n^* can be used for the test of the null hypothesis that F_3 is a mixture of F_1 and F_2. However, it is to be noted that the mixture of F_3 is not a necessary condition for $E(t) = \frac{1}{2}$.

Now consider the following situation of null and alternative hypotheses; H_0: $F_3(x) = \alpha F_1(x) + (1-\alpha)F_2(x)$ for all x for some $0 < \alpha < 1$. H_1: $F_3(x) = \alpha F_1(x) + (1-\alpha)F_2(x)$ has a nondegenerate solution at $x = a$ and no other finite solutions. Then, under H_1, it can be shown that

$$E(t) > \tfrac{1}{2} \text{ if, and only if, } f_3(a) - \alpha f_1(a) - (1-\alpha)f_2(a) > 0$$

while $E(t) = \frac{1}{2}$ under H_0. It can also be shown that $\mathrm{var}(t_n^*) \to 0$ under H_0 and H_1. Hence, the two-sided test

$$\text{Reject } H_0 \text{ if } \left| t_n^* - \tfrac{1}{2} \right| > \epsilon(b)$$

is consistent for testing H_0 against H_1 for some significance level b.

Let $R_j(1)$ denote the number of X_{1j}'s less than or equal to X_{2j} and let $R_j(3)$ be the number of X_{3r}'s less than or equal to X_{2j} and let $S_r(i)$ $(i = 1, 2)$ denote the number of X_{ij}'s less than or equal to X_{3r}. Then the statistic t_n^* defined by (3.1) can be expressed as

$$t_n^* = \frac{1}{n_1 n_3} \sum_{r=1}^{n_3} S_r(1) + \frac{1}{n_2 n_3} \sum_{j=1}^{n_2} R_j(3) - \frac{1}{n_1 n_2} \sum_{j=1}^{n_2} R_j(1).$$

From this and some other relations the proportion α can then be estimated by

$$(3.2) \qquad \hat{\alpha} = (n_1 \sum_{j=1}^{n_2} R_j(3) - \frac{1}{2} n_1 n_2 n_3)/(n_2 \sum_{r=1}^{n_3} S_r(1) + n_1 \sum_{j=1}^{n_2} R_j(3) - n_1 n_2 n_3).$$

Also, let $\delta = P_y\{X_1 < X_2\}$, then $\delta = \int_{-\infty}^{\infty} F_1(x) dF_2(x)$ and δ can be estimated by

$$(3.3) \qquad \hat{\delta} = \sum_{j=1}^{n_2} R_j(1)/n_1 n_2.$$

$$(3.4) \qquad \text{Let } \beta_i = \int_{-\infty}^{\infty} F_1(x)\, F_2(x)\, dF_i(x) \qquad (i=1,2).$$

Then, the probabilities $2\beta_1$ and $2\beta_2$ can, similarly, be estimated by considering these triples (X_{1i}, X_{1r}, X_{2j}) and (X_{1i}, X_{2s}, X_{2j}), respectively, where $i \neq r$, $j \neq s$.

Let $F_{in}(x)$ denote the empirical distribution functions associated with π_i $(i = 1, 2, 3)$. Suppose $\hat{\alpha}$ is calculated such that

$$(3.5) \qquad E(\hat{\alpha} - \alpha) = O(n^{-1})$$

$$(3.6) \qquad E(\hat{\alpha} - \alpha)^2 = \frac{v}{n} + O(n^{-2}).$$

Define

$$(3.7) \qquad \tau_n^2 = n \int_{-\infty}^{\infty} (\hat{\alpha} F_{1n}(x) + (1-\hat{\alpha})\, F_{2n}(x) - F_{3n}(x))^2\, dF_{3n}(x)$$

$$(3.8) \qquad \tau_n'^2 = n \int_{-\infty}^{\infty} (\hat{\alpha} F_{1n}(x) + (1-\hat{\alpha})\, F_{2n}(x) - F_{3n}(x))^2\, dF_3(x)$$

Then, it is shown by Thomas [124] that under the hypothesis that $F_3(x)$ is a mixture of $F_1(x)$ and $F_2(x)$, for any $\epsilon > 0$,

$$(3.9)\quad \lim_{n\to\infty} P_r \{|\tau_n^2 - \tau_n'^2| < \epsilon\} = 1.$$

By (3.6) - (3.8), we have, ignoring the terms $O(n^{-1})$

$$(3.10)\quad E(\tau_n'^2) = \frac{1}{3} + \frac{4}{3}\ v^2 - \frac{1}{2}\ \alpha(1-\alpha)(1+2\alpha) - \alpha(1-\alpha)(1-2\alpha)\ \delta$$

$$-2(v^2 - \alpha^2(1-\alpha))\beta_1 - 2(v^2 - \alpha(1-\alpha)^2)\ \beta_2.$$

Now suppose $F_3(x) = \alpha(x)\ F_1(x) + (1-\alpha(x))\ F_2(x)$. Thomas [124] considered the following hypotheses

$$H_0\text{: } \alpha(x) = \alpha, \text{ for all } x,\ 0 < \alpha < 1$$

$$H_1\text{: } \alpha(x) \neq \text{constant.}$$

Using the estimate of α given by (3.2), Thomas [124] was able to show that

$$\mathrm{Var}(\tau_n'^2) = O(1)$$

where $\tau_n'^2$ is defined by (3.8) and thus under H_1, for any $c > 0$

$$\lim_{n\to\infty} P_r \{|D_n| > c\} = 1$$

where D_n is the difference between the estimates of the two sides of (3.10). The critical region: Reject H_0 if $|D_n| >$ c so proposed by Thomas [124] is thus consistent and asymptotically unbiased. Note the treatment of tests is non-parametric.

For a parametric consideration, Johnson [58] studied the same problem that an observed sample was consistent with it being from a mixture of two symmetrical populations. Hence, for his case, he assumed F_1 and F_2 are specified and both have symmetrical densities with means μ_1 and μ_2 and common variance σ^2. Let X_j denote the jth observation from π_3. Johnson [58] considered the statistic

(3.11) $$\hat{\alpha}_x = (\bar{X}_n - \mu_2)/(\mu_1 - \mu_2)$$

which can be easily shown to be unbiased for α. For some given a define

(3.12) $$Y_j = \begin{cases} 1 & \text{if } X_j < a \\ 0 & \text{otherwise.} \end{cases}$$

Let $p_i = P_r\{X_1 < a | \mu_i\}$ $(i = 1, 2)$.

Consider another statsitic

(3.13) $$\hat{\alpha}_y = (\bar{Y} - p_2)/(p_1 - p_2)$$

which can also be seen to be unbiased for α. If $\hat{\alpha}_x$ and $\hat{\alpha}_y$ differ greatly, this may be regarded as evidence that X_i are not distributed as a mixture of the two given components. Along this approach, Johnson [58] was able to show that $n \operatorname{Var}(\hat{\alpha}_x - \hat{\alpha}_y)$ was independent of unknown α, and, therefore, the statistic $(\hat{\alpha}_x - \hat{\alpha}_y)[\operatorname{Var}(\hat{\alpha}_x - \hat{\alpha}_y)]^{-1/2}$ should have approximately a standard normal distribution. However, this approximation is too rough and inaccurate. For some special normal components, he used $\sqrt{n}(\hat{\alpha}_x - \hat{\alpha}_y)V^{-1/2}$ as a test statistic which is approximately standard normal for large n, where $V = n \operatorname{Var}(\hat{\alpha}_x - \hat{\alpha}_y)$ can be, in fact, calculated. Some computations of the test were also made for some special cases. Another test based on the statistic $U_i = |X_i - \frac{1}{2}(\mu_1 + \mu_2)|$ was proposed. It was noted that U_i always has the same distribution whether X_i comes from π_1 or π_2. The number of X_i's between μ_1 and μ_2 have a binomial distribution with parameters n and $\Phi(|\mu_1 - \mu_2|/\sigma) - \frac{1}{2}$ if π_3 is really a mixture of two normal components.

Comparisons of powers based on the two proposed tests has been made by Johnson [58] and it is shown that the latter test is more powerful. These tests are all based on simple statistics of observations. The choice of a defined by (3.12) and the distribution of the test statistics may be needed for some further studies.

For the problem of testing whether the mean of the mixture density is equal to some prefixed value , Blumenthal and Govindarajulu [10] considered that $F_3(x)$ with mean θ is a mixture with proportion α of two normal components $F_1(x)$ and $F_2(x)$ which have different means but common variance. They considered the hypothesis $H_0: \theta = 0$ vs $H_1: \theta > 0$. A Stein's two-stage procedure was proposed. First one computes the sample variance S_n^2 of sample of size $m(\geq 3)$ from π_3 which is defined by

$$S_m^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X}^2)$$

then, one takes a second sample of size N-m, where

$$N = \max\{m, [\frac{S_m^2}{z}]\},$$

$[x]$ denoting the greatest integer value of x not exceeding x and z denotes some specified constant. Then one computes $T = \sqrt{N}\,\bar{X}_N/S_m$ where $\bar{X}_N$ is the total sample observed. The critical region proposed for rejecting H_0 is: $T > t_{m-1,1-\alpha}$ where $t_{a,\alpha}$ denotes the 100α percentile of the t-distribution with d.f.. Let R denote the random unobservable number of observations among $X_1, X_2, \ldots, X_m$ which come from π_1. Let μ_1, μ_2 and σ^2 denote, respectively, the mean of $F_1(x)$ and $F_2(x)$ and their common variance. Then, it was shown that

$$P_r(T < t|R, S_m^2, N) = \Phi(a/\sigma\delta) + (\Delta/\delta\sqrt{n})\,\Phi(a/\sigma\delta) \cdot [c(\delta^2 - (a/\sigma)^2) - (m\,\alpha - \delta)] + O(1/(n - m))$$

where $\Delta = ((\mu_2 - \mu_1)/\sigma)$, $\delta = (1 + \Delta^2\,\alpha(1-\alpha))^{1/2}$, $a = s\,t - \theta\sqrt{n}$

and $c = \Delta^2\,\alpha(1-\alpha)(2\alpha-1)/6\delta^4$ given that $R = r$, $S_m^2 = s$ and $N = n$. If $(mz)^{1/2}/\sigma < 1$, then the cdf of T was to be $\Phi(\xi) - [\xi\varphi(\xi)(1+\xi^2)/8\delta^4(m-1)] \cdot [2+4\,\Delta^2\alpha(1-\alpha) + \Delta^2\alpha(1-\alpha)(2\alpha-1)^2] + [\Delta^3\sqrt{z}(2+\xi^2)\;\varphi(\xi)\;\alpha(1-\alpha)(2\alpha-1)/3\sigma\delta^4]$ with error term $0[\max\,(m^{-1.5}, m^{-0.5}/\sigma)]$ where $\xi = (t - (\theta/\sqrt{z}))$ and $\Phi(x)$ and $\varphi(\alpha)$ denote, respectively, the standard normal cdf and its density. Based on this distribution, the sizes of the Stein two-stage test were computed for some special given values of m, Δ, α and the first kind of error. The test is good in the sense that the size is small comparing to the one expected. However, in many situations, the values of α or even the values of Δ are unknown, and when this is the case, the two-stage test can not be carried out.

As it has been pointed out in part A of Section 2 that on many occasions, a difficulty that the statistician is confronted with for the estimation of the parameters in the mixture density is that it is unknown if the observed sample is mixed consisting of some other samples with specified or unspecified densities. This is a question that has been studied in this section.

4. Multiple Decision (Selection and Ranking) Problems for Mixture of Distributions

Suppose a population π consists of k subpopulations, say, $\pi_1, \pi_2, \ldots, \pi_k$ such that in a sample an individual observation comes from π_i with probability α_i ($i = 1, 2, \ldots, k$). Let $f_i(x)$ denote the density function of a random observation from π_i. Then the density of a random observation from π is given by a finite mixture $f(x) \equiv \sum_1^k \alpha_i\, f_i(x)$. In some situations, based on sampling from π, we are interested in selecting some π_j so that the associated α_j is the largest among all probabilities α_i $(i = 1, 2,\ldots,k)$. We call this kind of selection problem the first kind of selection in finite mixture. When the density $f_i(x)$ is degenerate at a certain point with probability mass one, this special situation becomes the problem for the selection of the most probable event in k categories i.e. the multinomial cell selections problem. On the other hand, suppose there are k populations, say, π_1', $\pi_2',\ldots,\pi_k'$ such that the density of a random observation from π_i' is given by a finite mixture $g_i(x) \equiv \sum_1^m \alpha_{ir} f_r(x)$ $(i = 1, 2,\ldots k)$, where each component density $f_r(x)$ is fixed, may be specified or unspecified. By sampling from each population, we are interested in selecting some π_j' so that the associated parameter α_{jr} is the largest (or smallest) among all $\alpha_{1r}, \alpha_{2r},\ldots\alpha_{kr}$ for some prefixed r. For convenience without loss of generality, we may take $r = 1$, that is in the mixture, we put the component $f_r(x)$ under main consideration in the first place so that we may consider the selection of the largest (smallest) α_{j1}. We call this kind of selection the second kind of selection in finite mixtures. When $m = 2$ and $f_1(x)$ and $f_2(x)$ are both degenerate with different values, the second selection problem becomes the usual selection of the best coin (see Gupta and Sobel [47]). It is to be noted that both kinds of selection occur in the compound decision problems as proposed by Robbins [96] in which mixing distributions correspond to some prior distributions. In this section we restrict ourselves to the second kind of selection. First of all, we consider the case when the sample size is small and then consider the large sample size situation. In this section, all component densities will be assumed identifiable.

4A. Samll Sample Size Case

In this part we impose no restriction on the parameter space. Based on the given samples of size n from each population we wish to select a subset of populations which includes the one we desire most with high probability which is pre-assigned before the experiment is carried out. This approach is called the subset selection formulation. One can refer to Gupta [46] for more details.

a) Procedures based on discriminant points

Suppose $\pi_1, \pi_2, \ldots, \pi_k$ are k populations such that the cumulative distribution function of π_i is a mixture of two components given by

$$G_i(x) = \alpha_i F(x-\theta_1) + (1-\alpha_i) F(x-\theta_2) \qquad i = 1, 2, \ldots, k$$

for some unknown $\alpha_i \in (0,1)$ with $\theta_1 < \theta_2$.

Let $\Omega = \{\underset{\sim}{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k): 0 < \alpha_i < 1\}$.

Let $X_{i1}, X_{i2}, \ldots, X_{in}$ denote n independent observations from π_i. To select a subset of populations containing the one associated with the largest α_i, we consider the following rule $R(x_0)$, which is based on some fixed point x_0, which selects a non-empty subset of populations when samples are taken. For a given point x_0, let N_i denote the number of observations from π_i that are less than or equal to x_0. We define $R(x_0)$: Select π_i, if and only if

$$N_i \geq \max_{1 \leq j \leq k} N_j - c$$

for some positive constant c.

Suppose θ_1 and θ_2 are known; without loss of generality we may assume $\theta_1 = 0$ and $\theta_2 = \Delta$. If F is specified, set $F_1(x_0) = F(x_0-\Delta)$. Then, since the random variable N_i is a binomial random variable with parameter n and $p_i \equiv \alpha_i F(x_0) + (1-\alpha_i) F_1(x_0)$ it follows that $p_i \leq p_j$, if, and only if, $\alpha_i \leq \alpha_j$. Since $G_i(x)$ is stochastically increasing with respect to α_i, the probability of a correct selection (CS: correct selection means selection of any subset which includes the population with the larger α_i) is thus minimized in the set $\{(\alpha,\alpha,\ldots,\alpha): 0 \le \alpha \le 1\}$ (see Desu [29]). We thus conclude,

<u>Theorem 1.</u> $$\inf_{\underset{\sim}{\alpha}\in\Omega} P_{\underset{\sim}{\alpha}}\{CS|R(x_0)\} = \inf_{0\le\alpha\le 1} \sum_{r=0}^{m} H^{k-1}(c + r; \alpha, x_0)\, h(r; \alpha, x_0)$$

where $H(i; \alpha, x_0) = \sum_{r=0}^{i} h(r; \alpha, x_0)$ and

$$h(r; \alpha, x_0) = \binom{n}{r} [\alpha F(x_0) + (1-\alpha) F_1(x_0)]^r [\alpha(1-F(x_0)) + (1-\alpha)(1-F_1(x_0))]^{n-r}.$$

To choose x_0, we see that when F is symmetric about 0, the best choice of x_0 is given by $x_0 = \frac{\Delta}{2}$. If F is not symmetric, by a geometrical argument, it is clear that it suffices to choose x_0 in $(0,\Delta)$ so that the right hand side of Theorem 1 attains its maximum. When Δ is unknown, we need to consider the infimum of the right hand side of Theorem 1 for all $\Delta > 0$ and then choose some $x_0 > 0$ so that a supremum is attained.

<u>Corollary 1</u>: Suppose $G_i(x) = \sum_{r=1}^{m} \alpha_{ir} F_r(x)$ is a finite mixture of m identifiable cumulative distributions function, $i = 1, 2, \ldots, k$. If for any $\beta_i > 0$, $\sum_{j=1}^{m} \beta_j = 1$, there exists x_0 such that $F_1(x_0) > \sum_{i=2}^{m} \beta_i F_i(x_0)$ and for this x_0, $\sum_{r=2}^{m} \alpha_{jr} F_r(x_0) \ge \sum_{r=2}^{m} \alpha_{ir} F_r(x_0)$ if and only if $\alpha_{j1} > \alpha_{i1}$. Then, for the selection of some populations associated with the largest α_{i1}, we have

$$\inf_{\underset{\sim}{\alpha}\in\Omega} P_{\underset{\sim}{\alpha}}(CS|R(x_0)) = \inf_{0\le p\le 1} \left[\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} \left[\sum_{j=0}^{i+c} \binom{n}{j} p^j (1-p)^{n-j}\right]^{k-1}\right]$$

Proof: Define $\delta_{ij} = \alpha_{ij}/(1-\alpha_{i1})$ $j = 2, 3, \ldots, n$, $i = 1, 2, \ldots, k$. Then,

$$G_i(x) = \alpha_{i1} F_1(x) + (1-\alpha_{i1}) F_{1i}(x) \text{ where}$$

$$F_{1i}(x) = \sum_{j=2}^{m} \delta_{ij} F_i(x) \text{ with } \delta_{ij} > 0, \sum_{j=2}^{m} \delta_{ij} = 1.$$

By given conditions, we have $\alpha_{j1} > \alpha_{i1}$ if, and only if, $p_j > p_i$ where

$p_i = \alpha_{i1} F_1(x_0) + (1-\alpha_{i1}) F_{1i}(x_0)$ which is the associated parameter of the binomial random variable N_i. The problem thus becomes the selection of the largest p_i which is discussed in Gupta and Sobel [47] and Gupta, Huang and Huang [44]. For k = 2 the infimum takes place at $p = \frac{1}{2}$ and for $k \geq 3$ asymptotic results and lower bounds are obtained.

We note that when $F_i(x) = F(x-\theta_i)$ with $\theta_1 > \theta_2 > \ldots > \theta_k$ the conditions in the corollary are satisfied if $\alpha_{jr}/(1-\alpha_{j1}) > \alpha_{ir}/(1-\alpha_{i1})$ for $r = 2, 3, \ldots, m$. The optimal choice of x_0 is impossible unless F and θ_i's are specified. For a detailed discussion of the computation c reference should be made to Gupta and Sobel [47] or Gupta, Huang and Huang [44].

<u>Corollary 2</u>: If $G_i(x) = \alpha_i \Phi(x;\ \theta_1, \sigma_1^2) + (1-\alpha_i) \Phi(x;\ \theta_2, \sigma_2^2)$ where $\Phi(x;\ \theta, \sigma^2)$ denotes the normal cdf with mean θ and variance σ^2, then

i) if $\theta_1 < \theta_2$ and $\sigma_1 = \sigma_2$, the best choice of x_0 is given by $(\theta_1 + \theta_2)/2$,

ii) if $\theta_1 = 0$ and $\theta_2 = \Delta > 0$, the best choice of x_0 is the real root in the interval $(0,\Delta)$ of the equation

$$(\sigma_2^2 - \sigma_1^2)x_0^2 + 2\sigma_1^2\Delta x_0 - \sigma_1^2\Delta^2 - 2\sigma_1^2\,\sigma_2^2\,(\ln\ \sigma_2 - \ln\ \sigma_1) = 0,$$

iii) if θ_1 and θ_2 are unknown and $\sigma_1 \leq \sigma_2$, then for any x_0,

$\inf p_{\underline{\alpha}}\ (CS|R(x_0)) = B(k,n,c)$ which is the same expression as on right hand side of Corollary 1.

The proof of Corollary 2 is straightforward and hence omitted.

Next, we consider the case of a mixture of three identifiable cdf's. Suppose

$$G_i(x) = \alpha_i F_1(x) + \beta_i F_2(x) + \gamma_i F_3(x) \quad \text{where}$$

$$0 < \gamma_i = 1-\alpha_i - \beta_i,\ 0 < \alpha_i,\ \beta_i < 1,\quad i = 1, 2, \ldots k.$$

We consider a rule which is based on two discriminant points, say, x_0 and x_1 ($x_0 < x_1$). Let N_i denote the number of samples from π_i which lie in (x_0, x_1). For the selection of the largest β_i, we propose the following rule:

$R(x_0, x_1)$: Select π_i iff

$$N_i \geq \max_j N_j - d$$

Then, we have the following theorem:

<u>Theorem 2</u>: If $F_i(x) = F(x-\theta_i)$ with $\theta_1 < \theta_2 < \theta_3$ and F is symmetric about 0, then, for $x_0 \in (\theta_1, \theta_2)$ and $x_1 \in (\theta_2, \theta_3)$ with $x_0 - \theta_1 = \theta_3 - x_1$,

$$\inf_{\underset{\sim}{\alpha}} p_{\underset{\sim}{\alpha}} \{CS|R(x_0, x_1)\} = B(k,n,d).$$

Proof: N_i is a binomial random variible with parameter $[F_1(x_1) - F_3(x_1) + F_3(x_0)] - F_1(x_0)]\alpha_i + [F_2(x_1) - F_3(x_1) + F_3(x_0) - F_2(x_0)]\beta_i + [F_3(x_1) - F_3(x_0)]$. The conditions of the choices of x_0 and x_1 and the symmetry of F imply the coefficient of α_i vanishes and the coefficient of β_i is strictly positive. Hence, $p_i < p_j$ if, and only if, $\beta_i < \beta_j$. This completes the proof.

There are (uncountably) many choices of x_0 and x_1, the discriminant points. However, the ones that maximize $F(x_1 - \theta_2) - F(x_1 - \theta_3) + F(x_0 - \theta_3) - F(x_0 - \theta_2)$ with $x_0 - \theta_1 = \theta_3 - x_1$ would be optimal in the sense that the infimum of the probability of a correct selection (with respect to the parameter space) is maximized.

Corollary 3: If $G_i(x) = \alpha_i\ \Phi(x;\ \theta_1,\ \sigma^2) + \beta_i\ \Phi(x;\ \theta_2,\ \sigma^2) + \gamma_i\ \Phi(x;\ \theta_3,\ \sigma^2)$ with $\theta_1 < \theta_2 < \theta_3$. Then, the optimal choices of x_0 and x_1 are those which maximize $\int_{-(\theta_3-x_1)}^{(x_1-\theta_2)} \varphi(t;0,\sigma^2)dt$ and minimize $\int_{-(\theta_2-x_0)-(\theta_3-\theta_2)}^{-(\theta_2-x_0)} \varphi(t;0,\sigma^2)\,dt$ with the restriction $x_0 - \theta_1 = \theta_3 - x_1$.

Proof: Proof follows from Theorem 2 and by noting that $\int_{x_0}^{x_1} \varphi(t;\theta_2,1)dt$ - $\int_{x_0}^{x_1} \varphi(t;\theta_3,1)dt = \left(\int_{-(\theta_3-x_1)}^{x_1-\theta_2} - \int_{-(\theta_2-x_0)-(\theta_3-x_2)}^{-(\theta_2-x_0)}\right) \varphi(t;0,1)dt$

b) Selection Procedures Based on Sample Means

We assume $G_i(x) = \alpha_i\ F_1(x) + (1-\alpha_i)\ F_2(x)$ such that $F_1(x) < F_2(x)$ for all x. For the subset selection of populations associated with the largest α_i, we propose

$$R_1:\ \text{Select } \pi_i \text{ if, and only if } \bar{X}_i \geq \max_j \bar{X}_j - c$$

Then, we have the following

Theorem 3: $\inf\limits_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}}\ (CS|R_1) = \inf\limits_{0\leq\alpha\leq 1} \int_{-\infty}^{\infty} H^{k-1}(x + c,\ \alpha)\ dH(x,\alpha)$

where

$$H(x,\alpha) = \Sigma\binom{n}{j}\ \alpha^j(1-\alpha)^{n-j}\ F_1^{*j} * F_2^{*(n-j)}(nx) \text{ with}$$

$F_i^{*r}(x)$ being the r convolutions of $F_i(x)$.

Proof: Since $G_i(x)$ is a stochastically increasing family of distributions with respect to α_i, hence $P_{\underset{\sim}{\alpha}}(CS|R)$ attains its infimum in the set $\{(\alpha,\alpha,\ldots,\alpha):\ 0 \leq \alpha \leq 1\}$. We also note that

$$P_r\{\bar{X}_i \le x\} = \sum_{j=0}^{n} P_r\{\sum_1^s Y_i + \sum_1^{n-s} Z_j \le n\,x \mid s=j\}\ P\{s=j\}$$

$$= \sum_{j=0}^{n} \binom{n}{j} \alpha_i^{j} (1-\alpha_i)^{n-j}\ F_1^{*j} * F_2^{*(n-j)}\ (nx)$$

where Y_i and Z_j are independent random observations corresponding to F_1 and F_2 respectively.

Corollary 3: If $F_i(x) = \Phi(x;\theta_i,\sigma_i^2)$ $(i = 1,2)$ with $\theta_1 > \theta_2$ and $\sigma_1 \le \sigma_2$, then,

$$\inf_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}} \{CS|R_1\} = \inf_{0\le\alpha\le1} \left[\sum_{j=0}^{n} \sum_{i=0}^{n} \binom{n}{j}\binom{n}{i} \alpha^{i+j}(1-\alpha)^{2n-i-j}\ \Phi(t(\theta_1,\theta_2,\sigma_1,\sigma_2,c))\right]$$

where $t(\theta_1,\theta_2,\sigma_1,\sigma_2,c) = [(i-j)(\theta_2-\theta_1) + nc](j\sigma_1^2 + (n-j)\ \sigma_2^2)^{1/2}/[(i+j)\ \sigma_1^2 + (2n-i-j)\sigma_2^2]$.

4B. Results for the Case of Large Sample Size

For convenience, we define some notation first. For a prefixed integer m, we define

(4.1) $\quad <0,1>^m = \{(\alpha_1,\alpha_2,\ldots,\alpha_m): \alpha_i > 0,\ \sum_1^m \alpha_i = 1\} \qquad (m \ge 2)$

Let $F_1(x;\theta)$, $F_2(x;\theta),\ldots,F_m(x;\theta)$ be m identifiable cdf's. We denote

(4.2) $\quad \underset{\sim}{F}(x;\theta) = (F_1(x;\theta),\ F_2(x;\theta),\ldots,F_m(x;\theta))$

(4.3) $\quad \underset{\sim}{\alpha}_i = (\alpha_{i1},\ \alpha_{i2},\ldots,\alpha_{im}),\ \ \underset{\sim}{\alpha}_i \in\ <0,1>^m.$

A finite mixture with m component $\underset{\sim}{F}(x;\theta)$ is defined to be the inner product of certain $\underset{\sim}{\alpha} \in <0,1>^m$ and $\underset{\sim}{F}(x;\theta)$ i.e.

$$(4.4)\quad G(x;\underset{\sim}{\alpha}) = \underset{\sim}{\alpha} \cdot \underset{\sim}{F}(x;\theta) = \sum_{i=1}^{m} \alpha_i F_i(x;\theta)$$

Let $\pi_1, \pi_2, \ldots, \pi_k$ be k populations such that π_i has cdf $G(x;\underset{\sim}{\alpha}_i)$ for some unknown but fixed parameter $\underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m$. Let $X_{i1}, X_{i2}, \ldots, X_{in}$ be n independent observations from π_i, $i = 1, 2, \ldots, k$. Let $G_{in}(x)$ denote the associated empirical distribution. Let λ denote a real-valued continuous function on $\langle 0,1\rangle^m$. Let $\lambda_{[1]}(\underset{\sim}{\alpha}) \leq \lambda_{[2]}(\underset{\sim}{\alpha}) \leq \ldots \leq \lambda_{[k]}(\underset{\sim}{\alpha})$ denote the ordered values of $\lambda(\underset{\sim}{\alpha}_1), \lambda(\underset{\sim}{\alpha}_2), \ldots, \lambda(\underset{\sim}{\alpha}_k)$.

Based on n independent observations from each population, we are interested in selecting t $(1 \leq t \leq k - 1)$ populations, say, $\pi_{r_1}, \pi_{r_2}, \ldots \pi_{r_t}$ such that $\lambda(\underset{\sim}{\alpha}_{r_1}), \lambda(\underset{\sim}{\alpha}_{r_2}), \ldots, \lambda(\underset{\sim}{\alpha}_{r_t})$ are the t largest namely, $\lambda_{[k]}(\underset{\sim}{\alpha}), \ldots, \lambda_{[k-t+1]}(\underset{\sim}{\alpha})$. We call these populations the t best.

We approach the problem using the indifference zone formulation. For given Δ (>0), we define

$$(4.5)\quad \Omega(\lambda;\Delta) = \{(\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_k) : \underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m,\ \lambda_{[k-t+1]}(\underset{\sim}{\alpha}) \geq \lambda_{[k-t]}(\underset{\sim}{\alpha}) + \Delta\}$$

Also, for convenience, we define the k-cartesian product

$$(4.6)\quad \Omega = \langle 0,1\rangle^m \times \langle 0,1\rangle^m \times \ldots \times \langle 0,1\rangle^m.$$

For specified $\underset{\sim}{F}(x;\theta)$ and λ, we consider our problem on the configuration $\Omega(\lambda;\Delta)$ for given Δ using the indifference zone approach.

Let $H(x)$ be some specified cdf. Let $\underset{\sim}{X}$ be a sample of size n from a population with density $\underset{\sim}{\alpha}_0 \cdot \underset{\sim}{F}(x;\theta)$ for some $\underset{\sim}{\alpha}_0 \in \langle 0,1\rangle^m$ and let $G_n(x)$ denote the associated empirical distribution. For $\underset{\sim}{\alpha} \in \langle 0,1\rangle^m$, we define

$$(4.7) \qquad S(\underset{\sim}{\alpha};H) = \int_{-\infty}^{\infty} (\underset{\sim}{\alpha} \cdot \underset{\sim}{F}(x;\theta) - G_n(x))^2 dH(x)$$

for some given value of θ.

a) Continuous Case

We assume that the parametric form of each component $F_i(x;\theta)$ is continuous in x for each θ and also that it is continuous in θ for each x. If n independent observations are drawn from a population with mixture density $G(x;\underset{\sim}{\alpha}_0)$ for unknown $\underset{\sim}{\alpha}_0 \in \langle 0,1\rangle^m$, the value $\hat{\underset{\sim}{\alpha}}_n$ which minimizes $S(\underset{\sim}{\alpha};H)$ seems a good estimate for $\underset{\sim}{\alpha}_0$ in the least squares sense. It is to be noted that $\hat{\underset{\sim}{\alpha}}_n$ is a statistic and is a function of H(x). A good choice in some sense for the weight function H(x) is not simple. Bartlett and Macdonald [2] study some special case for m = 2. For $m \geq 3$, the situation is complicated. A natural and reasonable choice of H(x) would be $G_n(x)$ which is the associated empirical function. This choice has been studied in [19] and [18]. For an alternative choice of H(x) consider $G(x;\underset{\sim}{\alpha}) = \underset{\sim}{\alpha} \cdot \underset{\sim}{F}(x;\theta)$ which has been studied in [70]. For a fixed p $(0 \leq p \leq 1)$, we take

$$(4.8) \qquad H(x) = p\, \underset{\sim}{\alpha} \cdot \underset{\sim}{F}(x;\theta) + (1-p)\, G_n(x).$$

Associated with each π_i, we define, analogous to (4.7),

$$(4.9) \qquad S_i(\underset{\sim}{\alpha};p) = \int_{-\infty}^{\infty} (\underset{\sim}{\alpha} \cdot \underset{\sim}{F}(x;\theta) - G_{in}(x))^2 dH(x)$$

where H(x) is defined by (4.8) and $G_{in}(x)$ is the empirical distribution function corresponding to π_i $(i = 1, 2, \ldots, k)$. Define $\hat{\underset{\sim}{\alpha}}_i$ to be such that

$$(4.10) \qquad S_i(\hat{\underset{\sim}{\alpha}}_i;p) = \inf_{\underset{\sim}{\alpha} \in \langle 0,1\rangle^m} S_i(\underset{\sim}{\alpha};p).$$

The existence of $\hat{\underset{\sim}{\alpha}}_i$ can be shown to hold. For a fixed p $(0 \leq p \leq 1)$, we define a selection rule R_p as follows.

Take n independent observations from each π_i and compute $\underset{\sim}{\alpha}_i = \underset{\sim}{\alpha}_i(X_{i1}, X_{i2},\ldots,X_{in})$ which is defined by (4.10) and (4.9). Let $\lambda_{[1]}(\hat{\underset{\sim}{\alpha}}) \le \lambda_{[2]}(\hat{\underset{\sim}{\alpha}}) \le \ldots \le \lambda_{[k]}(\hat{\underset{\sim}{\alpha}})$ denote the ordered values of $\lambda(\hat{\underset{\sim}{\alpha}}_1), \lambda(\hat{\underset{\sim}{\alpha}}_2),\ldots,\lambda(\hat{\underset{\sim}{\alpha}}_k)$.

R_p: Select π_i if, and only if $\lambda(\hat{\underset{\sim}{\alpha}}_i) \ge \lambda_{[k-t+1]}(\hat{\underset{\sim}{\alpha}})$.

A random mechanism is used to break the ties. By a correct selection (CS) we mean a set of t populations associated with the t largest values $\lambda(\underset{\sim}{\alpha}_1), \lambda(\underset{\sim}{\alpha}_2),\ldots,\lambda(\underset{\sim}{\alpha}_k)$ is selected.

Definition 1 A selection procedure R is consistent with respect to λ if

$$\lim_{\Delta\to 0}\lim_{n\to\infty} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R\} = 1$$

Definition 2 A selection procedure R is asymptotically strongly monotone with respect to λ if $\lambda(\underset{\sim}{\alpha}_i) < \lambda(\underset{\sim}{\alpha}_j)$ and for any $\in > 0$ implies

$$\lim_{n\to\infty} \sup_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_i \text{ is selected}|R\} - \in < \lim_{n\to\infty} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_j \text{ is selected}|R\}$$

Theorem 4 R_p is consistent and asymptotically strongly monostone with respect to a continuous λ.

Proof: (a) We show that $\hat{\underset{\sim}{\alpha}}_i \to \underset{\sim}{\alpha}_i$ with probability one for each i = 1, 2,...,k. Now, by the Glivenko-Cantelli theorem, for $\in > 0$, $\exists\, N(\in)$ such that, whenever $n \ge N(\in)$,

$$P_r\{|[p\,\underset{\sim}{\alpha}_i \cdot \underset{\sim}{F}(x;\theta) + (1-p)\,G_{in}(x)] - G_{in}(x)| < \in\} = P_r(p|\underset{\sim}{\alpha}_i \cdot \underset{\sim}{F}(x;\theta) - G_{in}(x)| < \in) = 1.$$

Replacing $dF_n(x)$ by $d(p\,\underset{\sim}{\alpha}_i \cdot \underset{\sim}{F} + (1-p)\,G_{in}(x))$ and follow the same argument as given in the proof of Theorem 2 in [19] the result follows.

(b) Consistency of R_p

Since λ is continuous it follows thus $\lambda(\hat{\underset{\sim}{\alpha}}_i) \to \lambda(\underset{\sim}{\alpha}_i)$ with probability one.

Now, by the Egoroff's theorem, for $\in > 0$ and $\delta > 0$ there exists $N_i(\in,\delta)$, A_i and B_i such that the sample space is decomposed to be $A_i \cup B_i$ with B_i the complement of A_i and $P(B_i) > 1-\in$ and on B_i, $|\lambda(\hat{\underset{\sim}{\alpha}}_i) - \lambda(\underset{\sim}{\alpha}_i)| < \delta$ whenever $n \geq N_i(\in,\delta)$ uniformly in $\underset{\sim}{\alpha}_i \in <0,1>^m$, i.e. $N_i(\in,\delta)$ is independent of $\underset{\sim}{\alpha}_i$. Note that $\lambda(\hat{\underset{\sim}{\alpha}}_i)$ depends on n. Set $N = N_1(\in,\delta) + \ldots + N_k(\in,\delta)$ and set $B = \bigcap_{i=1}^{k} B_i$. Then, $P(B) > 1-\in$, and on B, whenever $n \geq N$,

$\max_{1\leq i\leq k} |\lambda(\hat{\underset{\sim}{\alpha}}_i) - \lambda(\underset{\sim}{\alpha}_i)| < \delta$ uniformly for each $(\alpha_1, \alpha_2, \ldots \alpha_k) \in \Omega$. Now, for any $P^* \in (0,1)$, and any given $\Delta > 0$, choose $\delta = \frac{\Delta}{3}$ and $\in = 1-p^*$. Since on $\Omega(\lambda;\Delta)$, $\lambda_{[k-t+1]} - \lambda_{[k-t]} \geq \Delta = 3\delta$. Hence, we conclude that

$$P_{\underset{\sim}{\alpha}}\{\lambda(\hat{\underset{\sim}{\alpha}}_{r_i}) > \lambda_{[k-t]}(\hat{\underset{\sim}{\alpha}}),\ i = 1, 2, \ldots, t \mid \lambda(\underset{\sim}{\alpha}_{r_i}) > \lambda_{[k-t]}(\underset{\sim}{\alpha})\} > p^*$$

$\forall\, \underset{\sim}{\alpha} \in \Omega(\lambda,\Delta)$. Hence, we have shown that for every $\Delta > 0$,

$\lim_{n\to\infty} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\alpha}(CS|R_p) = 1$. This is the consistency of R_p.

(c) Suppose $\lambda(\underset{\sim}{\alpha}_i) < \lambda(\underset{\sim}{\alpha}_j)$.

(i) If $\lambda(\underset{\sim}{\alpha}_i) \leq \lambda_{[k-t]}(\underset{\sim}{\alpha})$ and $\lambda(\underset{\sim}{\alpha}_j) \geq \lambda_{[k-t+1]}(\underset{\sim}{\alpha})$. Then, take $p^* \geq \frac{2}{3}$ and go through the arguments given in the previous part (b), we can conclude that $\inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_j \text{ is selected}|R_p\} \geq \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\alpha}(CS|R_p) \geq \frac{2}{3}$ whenever $n \geq N_0 = N_0(\Delta)$ for some N_0. On the other hand, for each $n \geq N_0$, $\{\pi_i$ is selected$|R_p\} \subset \{$selection is not correct$|R_p\}$. Hence, $P_{\underset{\sim}{\alpha}}\{\pi_i \text{ is selected}|R_p\} \leq 1-p_{\underset{\sim}{\alpha}}(CS|R_p) \leq \frac{1}{3}$ $\forall\, \underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)$, i.e. $\sup_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{\pi_i \text{ is selected}|R_p\} \leq \frac{1}{3}$ for each $n \geq N_0$.

(ii) Suppose both $\lambda(\underset{\sim}{\alpha}_i)$ and $\lambda(\underset{\sim}{\alpha}_j)$ are no larger than $\lambda_{[k-t]}(\underset{\sim}{\alpha})$. Then, for $\epsilon > 0$ and by the arguments in (b), there exists a subset of sample space B and an integer N_0 such that $P\{B\} > 1-\frac{\epsilon}{2}$ and for $n \geq N_0$ and on B, $\max_{1\leq i\leq k} |\underset{\sim}{\alpha}_i - \hat{\underset{\sim}{\alpha}}_i| < \frac{\Delta}{3}$. Let E denote the event $\{\pi_i \text{ is selected} | R_p\}$. Then $E = E \cap B + E \cap B^c$. Hence $\sup_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}}(E) \leq \sup_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}}\{E \cap B\} + \sup_{\underset{\sim}{\alpha}} P_{\underset{\sim}{\alpha}}\{E \cap B^c\}$

$\leq \sup_{\underset{\sim}{\alpha}} P_{\alpha}\{E \cap B\} + \frac{\epsilon}{2}$ since $P_{\underset{\sim}{\alpha}}\{E \cap B^c\} \leq P_{\alpha}\{B^c\} < \frac{\epsilon}{2}$ $\forall\ \underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)$. Noting that, for any $\alpha\in\Omega(\lambda;\Delta)$, $P_{\underset{\sim}{\alpha}}\{E \cap B\} = 0$ since, on B $\hat{\underset{\sim}{\alpha}}_i < \alpha_{[k-t+1]} - \frac{\Delta}{3}$.

(iii) If $\lambda(\underset{\sim}{\alpha}_i)$ and $\lambda(\underset{\sim}{\alpha}_j)$ are both no less than $\lambda_{[k-t+1]}(\underset{\sim}{\alpha})$, the argument is analogous to the case of (ii). The proof is complete.

<u>Remark 1</u> Let $t_1, t_2, \ldots, t_m$ be positive integers such that each t_i is no larger than k-1. Let $\Omega(t_1, t_2, \ldots, t_m) \equiv \{(\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_k): \alpha^{(i)}_{[k-t_i+1]} > \alpha^{(i)}_{[k-t_i]}, i = 1, 2, \ldots, m\}$ where $\alpha^{(i)}_{[j]}$ denotes the j-th largest value of the i-th component of $\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, \ldots, \underset{\sim}{\alpha}_k$ and we denote $\underset{\sim}{\alpha}_r = (\alpha^{(1)}_r, \alpha^{(2)}_r, \ldots \alpha^{(m)}_r)$. If for each i we are desired to select the t_i largest in the i-th component simultaneously, then, using the statistics $(\hat{\underset{\sim}{\alpha}}_1, \hat{\underset{\sim}{\alpha}}_2, \ldots \hat{\underset{\sim}{\alpha}}_k)$ which are defined by (4.10), associated with the i-th component, we select these populations which have the t_i largest values in the i-th component of $(\alpha^{(i)}_1, \alpha^{(i)}_2, \ldots, \alpha^{(i)}_k)$ $(i = 1, 2, \ldots, m)$. It can be shown that the simultaneous selections are also consistent and asymptotically strongly monotone on $\Omega(t_1, t_2, \ldots t_k)$.

<u>Definition 3</u> A selection procedure R is consistent of order $O(A(\Delta))$ $(o(A(\Delta)))$ with respect to λ if $\lim_{\substack{\Delta\to 0 \\ n=O(A(\Delta))}} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R\} = 1$ $\left(\lim_{\substack{\Delta\to 0 \\ n=o(A(\Delta))}} \inf_{\underset{\sim}{\alpha}\in\Omega(\lambda;\Delta)} P_{\underset{\sim}{\alpha}}\{CS|R\} = 1\right)$.

Theorem 5. R_p is consistent of order $O(\Delta^\delta)$ with respect to λ if λ satisfies Lipschitz condition. $(-\frac{1}{2} < \delta < 0)$.

Proof: We note that, by the Glivenko-Cantelli theorem that $\sup_x |G_i(x) - G_{in}(x) + o(1)| \to 0$ WP1 as $n \to \infty$ for each i. For any fixed i, let $\dot{S}(\underset{\sim}{\alpha}_i;p)$ denote the m-1 equations for which each equation is differentiated with respect to α_{ij}, $j = 1, 2, \ldots m-1$, where $\underset{\sim}{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{im-1}, 1 - \sum_{j=1}^{m-1} \alpha_{ij})$. Then, the first element of $\dot{S}(\underset{\sim}{\alpha}_i;p)$ for $j = 1$ becomes

$$\frac{1}{n} \sum_{j=1}^{n} F_1(X_{i[j]};\theta_1)\{\sum_{r=1}^{m} \alpha_{ir} F_r(X_{i[j]};\theta_r) \frac{j}{n} + \frac{1-p}{2n}\}$$

$$\leq \sup_x |G_i(x) - G_{im}(x) + o(1)| \frac{1}{n} \sum_{j=1}^{n} F_1(X_{i[j]};\theta_1)$$

where $X_{i[1]} \leq \cdots \leq X_{i[m]}$ are order statistics from π_i. Apply the analogous arguments in [19], we have $|\lambda(\hat{\underset{\sim}{\alpha}}_i) - \lambda(\underset{\sim}{\alpha}_i)| < O(n^{\delta-1/2})$ for all but finite n with probability 1 $(0 < \delta < 1/2)$ since λ satisfies Lipschitz condition. Now, take $|\lambda(\hat{\underset{\sim}{\alpha}}_i) - \lambda(\underset{\sim}{\alpha}_i)| = \Delta$ and let $\Delta \to 0$. Then, as $n \to \infty$, $\Delta \to 0$ and we have $n = O(\Delta^{\frac{-2}{1-2\delta}})$. This means as $\Delta \to 0$, the rate of divergence of n is to the order $(\frac{1}{\Delta})^{\frac{2}{1-2\delta}}$. In order that $\inf_{\alpha\ \Omega(\lambda;\Delta)} P_\alpha(CS|R) \to 1$ it suffices to take $n = (\frac{1}{\Delta})^{\frac{1}{2} - \delta}$ as $\Delta \to 0$.

Let $\bar{\underset{\sim}{\alpha}}_i$ denote the arithmetic mean of r independent estimates of $\hat{\underset{\sim}{\alpha}}_i$ where r is some integer. This means rn samples are drawn from each population. And for each subgroup of n samples, we obtain an estimate $\hat{\underset{\sim}{\alpha}}_i$ for the population π_i. If n is large, $\lambda(\underset{\sim}{\alpha}_i) = \alpha_{i1}$, and $t = 1$, we propose the following rule R_p^*.

R_p': Select π_i if $\bar{\alpha}_{i1} \geq \bar{\alpha}_{j1}$ for all $j \neq i$, where $\bar{\alpha}_{i1}$ is the first component of $\bar{\alpha}_i$.

Theorem 5. If n is large, $t = 1$ and $\lambda(\alpha_i) = \alpha_{i1}$, the projection function, then we have

$$\inf_{\alpha\in\Omega(\lambda;\Delta)} P_{\alpha}\{CS|R_p'\} \geq \int_{-\infty}^{\infty} \prod_{j=2}^{k} \Phi(\delta_j z + \frac{\sqrt{r}\Delta}{\sigma_{[j]}})d\Phi(z)$$

where $\Phi(x)$ denotes the standard normal distribution and

$$\sigma_j^2 = 2 \int\int_{-\infty<x<y<\infty} G_j(x)[1-G_j(y)]dB_j(x)dB_j(y)$$

where

$$B_j(x) = F_1(x;\theta_1)G_j(x) - \int_{-\infty}^{\infty} F_1(x;\theta_1)dG_j(x)$$

for $j = 1, 2, \ldots, k$ and

$$\sigma_{[1]} \leq \sigma_{[2]} \leq \cdots \leq \sigma_{[k]}, \ \delta_j = \sigma_{[1]}/\sigma_{[j]}$$

Proof: It has been shown in [18] that $\hat{\alpha}_i$ is asymptotically normal and hence, the first component of $\hat{\alpha}_i$, say $\hat{\alpha}_{i1}$ is asymptotically normal with mean α_{i1} and variance

$$\sigma_i^2 = 2 \int\int_{-\infty<x<y<\infty} G_i(x)\ [1-G_i(y)]dB_i(x)dB_i(y)$$

where

$$B_i(x) = F_1(x;\theta_1)G_i(x) - \int_{-\infty}^{\infty} F_1(x;\theta_1)dG_i(x).$$

Hence, when n is large and t = 1, we have for $\alpha \in \Omega(\lambda;\Delta)$

$$P_{\underset{\sim}{\alpha}}\{CS|R_p'\} = P_{\underset{\sim}{\alpha}}\{\bar{\alpha}_{k1} \geq \bar{\alpha}_{j1},\ j = 1, 2, \ldots, k-1 | \alpha_{k1} = \max_{1\leq j\leq k} \alpha_{j1}\}.$$

$$= P_{\underset{\sim}{\alpha}}\{\frac{\sqrt{r}(\bar{\alpha}_{k1} - \alpha_{k1})}{\sigma_k} \geq \frac{\sqrt{r}\ \bar{\alpha}_{j1} - \alpha_{j1})\sigma_j}{\sigma_j\ \sigma_k} + \frac{\sqrt{r}(\alpha_{j1} - \alpha_{k1})}{\sigma_k}\}$$

$$\geq P_{\underset{\sim}{\alpha}}\{Z_k \geq Z_j(\frac{\sigma_j}{\sigma_k}) - \frac{\sqrt{r}\Delta}{\sigma_k}\ \ j = 1, 2, \ldots, k\}$$

(where $Z_1, Z_2, \ldots Z_k$ are iid standard normal)

$$= \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(\frac{\sigma_k}{\sigma_j} z + \frac{\sqrt{r}\Delta}{\sigma_j}) d\Phi(z)$$

$$\geq \int_{-\infty}^{\infty} \prod_{j=1}^{k-1} \Phi(\sigma_j z + \frac{\sqrt{r}\Delta}{\sigma_{[j+1]}}) d\Phi(z) \text{ (by a lemma in [45])}$$

where $_j = \sigma_{[1]}/\sigma_{[j+1]}$, $\sigma_{[1]} \leq \sigma_{[2]} \leq \ldots \leq \sigma_{[k]}$. This completes the proof.

Asymptotic relative efficiency of R_p with respect to R_B

We assume m = 2, t = 1 and λ is the projection function. In this case we have $G_i(x) = \alpha_i\ F_1(x;\theta_1) + (1-\alpha_i)\ F_2(x;\theta_2)$ for i = 1, 2, ..., k and we denote α_i instead of $\underset{\sim}{\alpha}_i$. Suppose $F_1(x;\theta_1)$ and $F_2(x;\theta_2)$ are not specified, however, we assume there exists some point X_0, known, such that $F_1(x_0;\theta_1) \neq F_2(x_0;\theta_2)$. Assume $F_1(x_0;\theta_1) > F_2(x_0;\theta_2)$. Then, we see that $\alpha_i > \alpha_j$ if, and only if $G_i(x_0) > G_j(x_0)$. Hence, selecting the best is equivalent to selecting the population associated with the largest $G(x_0;\alpha_i)$ value.

For a given i, $1 \leq i \leq k$, and j, $1 \leq j \leq n$, define

$$Y_{ij} = \begin{cases} 1 & \text{if } X_{ij} \leq x_0 \\ 0 & \text{otherwise} \end{cases}$$

and define

$$\hat{G}_i(x_0) = \sum_{j=1}^{m} Y_{ij}.$$

Then, it is obvious that $\hat{G}_i(x_0)$ is a binomial random variable with cdf $B(n;\hat{G}(x_0))$.

We define a selection procedure R_B as follows:

R_B: Select the population π_i which is associated with the largest $\hat{G}_i(x_0)$.

When n is large, we use the normal approximation. Let $F_1(x_0;\theta_1) - F_2(x_0;\theta_2) = d > 0$. Then, by the result of [114], we have asymptotically $n \approx c^2(p^*)(1-\Delta^2 d_0^2)/2\Delta^2 d_0^2$ when $\Delta \to 0$ and $p^* \to 1$. Again, by the Feller's inequality, we see that $\Phi(z) \approx 1- \frac{1}{\sqrt{2\pi}\, z} e^{-\frac{z^2}{2}}$. We obtain $c^2(p^*) = (\frac{1}{1-p^*})^2$. Let n_1 and n_2 denote, respectively, the sample sizes associated with R_p and R_B when $\inf_{\underset{\sim}{\alpha} \in \Omega(\lambda;\Delta)} P_\alpha\{CS\} = P^*$ is satisfied for both rules. We define the asymptotic relative efficiency of R_p with respect to R_B by $ARE(R_p;R_B) = \frac{n_1(P^*,\Delta)}{n_2(P^*,\Delta)}$ as $P^* \to 1$ and $\Delta \to 0$. It follows from the previous result and the result in Theorem 4 we have

$$ARE(R_p;R_B) = \lim_{\substack{\Delta\to 0\\ p^*\to 1}} \frac{2(1-p^*)^2\, \Delta^{1.5+\delta}\, d_0^2}{1-\Delta^2\, d_0^2} = 0$$

However, if we take $1-P^* = \Delta \to 0$, we have an alternative efficiency defined by

$$ARE'(R_p;R_B) \equiv \lim_{\substack{\Delta\to 0\\ \Delta=1-p^*}} \frac{n_1(P^*,\Delta)}{n_2(P^*,\Delta)} = \lim_{\Delta\to 0} \frac{2\Delta^{\delta+3.5}\, d_0^2}{1-\Delta^2\, d_0^2} = 0$$

This shows that R_p is good compared to R_B.

b) Discrete case

In this case, we denote F_1, $F_2,\ldots,F_m$ as discrete distributions such that the outcomes from each distribution with cdf F_i, for some i, can be classified into $s(\geq 2)$ states. Let the probability that an outcome from F_i belongs to state ℓ denoted by $p_{i\ell}$. We assume F_1, $F_2,\ldots,F_m$ are all specified and $p_{i\ell}$ are all given.

For $\underset{\sim}{\alpha}_i \in \langle 0,1\rangle^m$ we define a mixture distribution G_i by $G_i = \alpha_{i1} F_1(x) + \alpha_{i2} F_2(x) +\ldots+\alpha_{im} F_m(x)$. Then, $G_i(x)$ is also a discrete distribution such that the probability of an outcome belonging to state j is given by

$$g_{ij} = \alpha_{i1} p_{1j} + \alpha_{i2} p_{2j} +\ldots+\alpha_{im} p_{mj} \quad \text{for } j = 1, 2,\ldots,s.$$

We assume that there exists a lower bound g_0 such that $g_{ij} \geq g_0 \geq 0$ for all $i = 1, 2,\ldots,k$, $j = 1, 2,\ldots,s$. Let n samples be drawn from π_i and let n_j denote the number of outcomes which belong to state j. For any $\underset{\sim}{\alpha} = (\alpha_1, \alpha_2,\ldots,\alpha_m)$ we define the Matusita distance (see [71]) as follows.

$$(4.11) \qquad S_i(\underset{\sim}{\alpha}) = \{\sum_{j=1}^{s} (\sqrt{g_i} - \sqrt{\frac{n_j}{n}})^2\}^{1/2}$$

where $g_i = \sum_{i=1}^{m} \alpha_i p_{ij}$. $S_i(\underset{\sim}{\alpha})$ is thus a function on $\langle 0,1\rangle^m$.

Let $\hat{\underset{\sim}{\alpha}}_i$ denote a value in $\langle 0,1\rangle^m$ such that $S_i(\hat{\underset{\sim}{\alpha}}_i)$ attains its infimum. For given π and λ, to select the t best with respect to λ, we propose the following selection procedure.

R: Select $\pi_{r_1}, \pi_{r_2},\ldots,\pi_{r_t}$ if, and only if

$\lambda(\hat{\underset{\sim}{\alpha}}_{r_1})$, $\lambda(\hat{\underset{\sim}{\alpha}}_{r_2}),\ldots,\lambda(\hat{\underset{\sim}{\alpha}}_{r_t})$ are the t largest values of

$\lambda(\hat{\underset{\sim}{\alpha}}_1)$, $\lambda(\hat{\underset{\sim}{\alpha}}_2),\ldots,\lambda(\hat{\underset{\sim}{\alpha}}_k)$, which are defined by (4.11).

We use a random mechanism in case of ties.

Theorem 6. The selection procedure R is consistent and asymptotically strongly monotone with respect to λ if λ is continuous.

Proof: It has been shown in [71] that for our case $\hat{\underset{\sim}{\alpha}}_i \to \underset{\sim}{\alpha}_i$ with probability one in the usual sense of convergence of a sequence of vectors. Therefore, $\lambda(\hat{\underset{\sim}{\alpha}}_i) \to \lambda(\underset{\sim}{\alpha}_i)$ WP1. Appling the analogous arguements given in the proofs of Theorem 4 we can conclude the same results. This completes the proof.

References

1. Andrews, D.F. and Mallows, C.L. (1974) Scale mixtures of normal distributions J.R. Statist. Soc. B, 36, 11-102.

2. Bartlett, M.S. and MacDonald, P.D.M. (1968) "Least-squares" estimation of distribution mixtures. Nature, Lond., 217, 195-196.

*3. Beeker, N.G. (1970) Mixture designs for a model linear in the proportions. Biometrika 57, 329-338.

4. Behboodian, Javad (1970) On a mixture of normal distributions. Biometrika 57, 215-217.

5. Blischke, W.R. (1962) Moment estimators for the parameters of a mixture of two binomial distributions. Ann. Math. Statist. 33, 444-454.

6. Blischke, W.R. (1963) Mixtures of discrete distributions. Proceedings of the International Symposium on Classical and Contagious Discrete Distributions, Statistical Publishing Society, Calcutta, 351-372.

7. Blischke, W.R. (1964) Estimating the parameters of mixtures of binomial distributions. J. Amer. Stat. Assoc. 59, 510-528.

8. Bliss, C.I. and Fisher, R.A. (1953) Fitting the negative binomial distribution to biological data. Note on the efficient fitting of the negative fitting of the negative binomial. Biometrics 9, 176-200.

9. Blum, J.R. and Susarla, V. (1977) Estimation of a mixing distribution function. Ann. Prob. 5, 200-209.

10. Blumenthal, S. and Govindarajulu, Z. (1977) Robustness of Stein's two-stage procedure for mixtures of normal populations. J. Amer. Statist. Assoc. 72, 192-196.

11. Boes, D.C. (1966) On the estimation of mixing distributions. Ann. Math. Statsit. 37, 177-188.

12. Boes, D.C. (1967) Minimax unbiased estimator of mixing distribution for finite mixtures. Sankhyā A 29, 417-420.

13. Burrau, Carl (1934) The half-invariants of the sum of two typical laws of errors, with an application to the problem of dissecting a frequency curve into components. Skand. Aktuarietidskr., 17, 1-6.

14. Chandra, Satish (1977) On the mixtures of probability distributions. Scand. J. Statist. 4, 105-112.

15. Chang, W.C. (1971) Resolution of mixtures of normal distributions Ph.D. dissertation, University of Wisconsin.

* denotes a reference dealing with the topic of experimental designs which is not discussed in this paper.

16. Charlier, C.V.L. (1960) Researches into the theory of probability. Meddelanden frau Lunds Astron. Observ., Sec. 2, Bd. 1.

17. Charlier, C.V.L. and Wicksell, S.D. (1924) On the dissection of frequency functions. Arkiv for Mathematik, Astronomi Och Fysik, Bd. 18, No. 6.

18. Choi, K. (1969) Estimators for the parameters of a finite mixture of distributions. Ann. Inst. Statist. Math. 21, 107-116.

19. Choi, K. and Bulgren, W.G. (1968) An estimation procedure for mixtures of distribution. J. Roy. Statist. Soc. B 30, 444-460.

20. Cohen, A. Clifford (1963) Estimation in mixtures of discrete distributions. Proceedings of the International Symposium on Classical and Contagious Discrete Distributions, Statistical Publishing Society, Calcutta, 373-378.

21. Cohen, A. Clifford (1967) Estimation in mixtures of two normal distributions. Technometrics 9, 15-28.

*22. Cornell, J.A. and Gared, I.J. (1970) The mixture problem for categorized components. J. Roy. Statist. Soc. 65, 339-355.

*23. Cornell, J.A. (1973) Experiments with mixtures: A review Technometrics 15, 437-456.

*24. Cornell, J.A. (1975) Some comments on designs for Cox's mixture of polynomial. Technometrics 17, 25-35.

*25. Cornell, J.A. (1977) Weighted versus unweighted estimates using Scheffé's mixture model for symmetric error variances patterns. Technometrics 19, 237-247.

*26. Cox, D.R. (1971) A note on polynomial response functions for mixture. Biometrika 58, 155-159.

27. Day, N.E. (1969) Estimating the components of a mixture of normal distributions. Biometrika 56, 463-474.

28. Deely, J.J. and Kruse, R.L. (1968) Construction of sequences estimating the mixing distribution. Ann. Math. Statist. 39, 286-288.

29. Desu, M.M. (1970) A selection problem. Ann. Math. Statist. 41, 1569-1603.

30. Doetsch, G. (1936) Zerlegung einer Function in Gauss'sche Fehlerkurven. Math. Zeit. 41, 283-318.

*31. Draper, N.R. and John, R.C. St. (1977) A mixture model with inverse terms. Technometrics 19, 37-46.

*32. Draper, N.R. and John, Ralph C. St. (1977) Designs in three and four components for mixture models with inverse terms. Technometrics 19, 117-130.

*33. Draper, N.R. and Lawrence, N.L. (1965) Mixture designs for three factors. J. Roy. Statist. Soc. B 27, 450-465.

*34. Draper, N.R. and Lawrence, W.I. (1965) Mixture designs for four factors. J. Roy. Statist. Soc B 27, 473-478.

35. Efron, Bradley and Olshen, Richard (1978) How broad is the class of normal scale mixture? Ann. Statist. 6, 1159-1164.

36. Falls, Lee W. (1970) Estimation of parameters in compound Weibull distributions. Technometrics 12, 399-407.

37. Feller, W. (1971) An Introduction to Probability Theory and Its Applications, vol. 2 (2nd ed.) John Wiley & Sons, Inc., New York.

38. Fisher, L. and Yakowitz, S.J. (1970) Estimating mixing distributions in metric spaces. Sankhyā, Ser. A, 32, 411-418.

39. Fryer, J.G. and Robertson, C.A. (1972) A comparison of some methods for estimating mixed normal distributions. Biometrika 59, 639-648.

*40. Galil, Z. and Kiefer, J. (1977) Comparison of simplex designs for quadratic mixture models. Technometrics 19, 445-453.

41. Goldie, C.M. (1967) A class of infinitely divisible distributions. Proc. Cambridge philos. Soc. 63, 1141-1143.

42. Gnedenko, B.V. and Kolmogorov, A.N. (1954) Limining Distributions for Sums of Independent Random Variables, translated from Russian by K.L. Chang, Addison-Wesley Publishing Co., Inc., Reading, Mass.

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

44. Gupta, Shanti S., Huang, D.Y. and Huang, W.T. (1976) On ranking and selection procedures and tests of homogeneity for binomial populations. In Essays in Probability and Statistics (S. Ikeda and others, eds., Tokyo), 501-533.

45 Gupta, S.S. and Huang, W.T. (1974) A note on selecting a subset of normal populations with unequal sample sizes. Sankhyā, Ser. A, 36, 389-396.

46. Gupta, Shanti S. (1977) Selection and ranking procedures: a brief introduction, Commun. Statist. Theor. Math. A 6, 993-1001.

47. Gupta, S.S. and Sobel, M. (1960) Selecting a subset containing the best of several binomial populations. Chapter XX in Contributions to Probability and Statistics. Ed. by I. Olkin, Standrod Univ. Press, Stanford, Calif., 224-248.

48. Hasselblad, V. (1966) Estimation of parameters for a mixture or normal distributions. Technometrics 8, 431-446.

49. Hasselblad, V. (1969) Estimation of finite mixtures of distributions from the exponential family. J. Amer. Statist. Assoc. 64, 1459-1471.

50. Hill, B.M. (1963) Information for estimating the proportions in mixtures of exponential and normal distributions. J. Amer. Statist. Assoc. 58, 918-932.

51. Hawkins, R.H. (1972) A note on multiple solutions to the mixed distribution problem. Technometrics 14, 973-976.

52. Hosmer, David W. Jr. (1973) On MLE of the parameters of a mixture of two normal distributions when the sample size is small. Commun. Statist. A1(3), 217-227.

53. Hosmer, David W. Jr. (1973) A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics 29, 761-770.

54. Ifram, A. (1970) On mixtures of distributions with applications to estimation. JASA 65, 249-254.

55. Isaenko, O.K. and Urbakh, V. Yu. (1977) Partitioning mixed probability distribtuions into their constituents. J. Soviet Math. 7, 148-161.

56. John, S. (1970) On identifying the population of origin of each observation in a mixture of observations from two normal populations. Technometrics 12, 553-563.

57. John, S. (1970) On identifying the population of origin of each observation in a mixture of observations from two Gamma populations. Technometrics, 12, 565-568.

58. Johnson, N.L. (1973) Some simple tests of mixtures with symmetrical components. Communications in Statist. 1, 17-25.

59. Kale, B.K. (1962) On the solution of likelihood equations by iteration processes. The multiparametric case. Biometrika 49, 479-486 .

60. Katti, S.K. and Gurland, John (1961) The Poisson Pascal distribution. Biometrics 17, 527-538.

61. Katti, S.K. and Gurland, John (1962) Some methods of estimation for the Poisson binomial distribution. Biometrics 18, 42-51.

62. Katti, S.K. and Gurland, John (1962) Efficiency of certain methods of estimators for the negative binomial and the Neyman Type A distributions. Biometrika 49, 215-226.

63. Keilson, J. (1965) A review of transient behavior in regular diffusion and birth-death processes, Part II. J. Appl. Probability 2, 405-428.

64. Keilson, J. and Steutel, F.W. (1972) Families of infinitely divisible distributions closed under mixing and convolution. Ann. Math. Statist. 43, 242-250.

65. Keilson, Julian and Steutel, F.W. (1974) Mixtures of distributions, moment inequalities and measures of exponentiality and normality. Ann. Prob. 2, 112-130.

66. Kelker, Douglas (1971) Infinite divisibility and variance mixtures of the normal distribution. Ann. Math. Statist. 42, 802-808.

67. Kingman, J.F.C. (1966) The algebra of quenes. J. App. Probability 3, 285-326.

*68. Kurotori, I.S. (1966) Experiments with mixtures of components having lower bounds. Industrial Quality Control 22, 592-596.

69. LeCam, L. (1956) On the asymptotic theory of estimation and testing hypotheses. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press, 129-156.

70. MacDonald, P.D.M. (1971) Comment on "An estimation procedure for mixtures of distributions" by Choi and Bulgren. J.R. Statsit. Soc. Ser. B, 33, 326-329.

71. Matusita, K. (1954) On the estimation by the minimum distance method. Ann. Inst. Statist. Math. 5, 59-65.

*72. McLean, R.A. and Anderson, V.L. (1966) Extreme vertices design of mixture experiments. Technometrics 8, 447-454.

73. Medgyessi, P. (1961) Decomposition of Superpositions of Distribution Functions. Publ. House of the Hungarian Academy of Sciences, Budapest.

74. Meeden, G. (1972) Bayes estimation of the mixing distribution, the discrete case. Ann. Math. Statist. 43, 1993-1999.

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

76. Mohanty, N.C. (1972) On the identifiability of finite mixtures of Laguerre distributions. IEEE Trans. Inform. Theory, 18, 514-515.

77. Mosimann, James E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika 49, 65-82.

78. Mudholker, Govind S. and Chaubey, Yogendra P. (1976) On the distribution of Fisher's transformation of the correlation coefficient. Comm. in Statist. Part B, 5, 163-172.

79. Murty, J.S. and Dao, M.N. (1968) Design and analysis of experiments with mixtures. Ann. Math. Statist. 39, 1517-1539.

80. Neyman, J. (1939) On a new class of 'contagious' distributions applicable in entomology and bacteriology. Ann. Math. Stat. 10, 35-57.

*81. Nigam, A.K. (1970) Block designs for mixture experiments. Ann. Math. Statist. 41, 1861-1869.

82. Odell, P.L. and Basu, J.P. (1976) Concerning several methods for estimating crop acreages using remote sensing data. Commun. Statist. A 5(12), 1091-1114.

83. Parthsarthy, K. (1967) Probability measures on metric spaces Academic Press, London and New York.

84. Pearson, Karl (1894) Contributions to the mathematical theory of evolution. Philos. Trans. Roy. Soc. London 185, 71-110.

85. Peters, Charles and Coberly, William A. (1976) The numerical evaluation of the maximum-likelihood estimate of mixture proportions. Commun. Statist. A, 5, 1127-1135.

86. Pollard, Harry S. (1934) On the relative stability of the median and arithmetric mean, with particular reference to certain frequency distributions which can be dissected into normal distributions. Ann. Math. Statist. 5, 227-262.

87. Preston, E.J. (1953) A graphical method for the analysis of statistical distributions into two normal components. Biometrika, 40, 460-464.

88. Preston, P.F. (1971) Estimating the mixing distribution by piece-wise polynomial arcs. Austral. J. Statist. 13, 64-76.

89. Ramachandran, D. (1974) Mixtures of perfect probability measures. Ann. Prob. 2, 495-500.

90. Rao, Bhaskara, K.P.S. and Bhaskara Rao, M. (1972) Mixtures of nonatomic measures. Proc. Amer. Math. Soc. 33, 507-510.

91. Rao, C.R. (1948) The utilization of multiple measurements in problems of biological classification. J. Roy. Stat. Soc. 10 B, 159-203.

92. Rao, C.R. (1952) Advanced Statistical Methods in Biometric Research. John Wiley, New York, pp. 300-304.

93. Rider, Paul R. (1961) Estimating the parameters of mixed Poisson, binomial, and Weibull distributions by the method of moments. Bull. Inst. Inter. Stat. 39, 143-147.

94. Rider, Paul R. (1961) The method of moments applied to a mixture of two exponential distributions. Ann. Math. Statist. 32, 143-147.

95. Robbins, H. (1948) Mixture of distributions. Ann. Math. Statist. 19, 360-369.

96. Robbins, H. (1969) The empirical Bayes approach to statistical decision problems. Ann. Math. Statist. 35, 1-20.

97. Rodine, R.H. (1966) Perfect probability measures and regular conditional probabilities. Ann. Math. Statist. 37, 1273-1278.

98. Rolph, J.E. (1968) Bayesian estimation of mixing distributions. Ann. Math. Statist. 39, 189-1302.

99. Saxena, S.K. and Nigam, A.K. (1973) Symmetric simplex block designs for mixtures. J. Roy. Statist. Soc. B 35, 466-472.

100. Saxena, S.K. and Nigam, A.K. (1977) Restricted exploration of mixtures by symmetric-simplex design. Technometrics 19, 47-52.

101. Sazonov, V. (1965) On perfect measures. Amer. Math. Soc. Transl. Ser. 2, 48, 229-254.

102. Scheffé (1958) Experiments with mixtures. J. Roy. Statist. Soc. B 20, 334-360.

103. Scheefé (1963) The simplex-centroid design for experiments with mixtures. J. Roy. Statist. Soc. B 52, 235-263.

104. Schoenberg, I.J. (1938) Metric spaces and completely monotone functions. Ann. Math. 39, 811-841.

105. Shenton, L.R. (1949) On the efficiency of the method of moments and Neyman's Type A distribution. Biometrika 36, 450-454.

106. Shenton, L.R. (1950) Maximum likelihood and the efficiency of the method of moments. Biometrika 37, 111-116.

107. Shenton, L.R. and Wallington, P.A. (1962) The bias of moment estimators with an application to the negative binomial distribution. Biometrika 49, 193-204.

108. Skellam, J.G. (1948) A probability distribution derived from the binomial distribtuion by regarding the probability of success as variable between the sets of trials. J. Roy. Stat. Soc. 10 B, 257-261.

*109. Snee, R.D. (1971) Design and analysis of mixture experiments. J. Quality Technology 3, 159-169.

*110. Snee, R.D. (1973) Techniques for the analysis of mixture data. Technometrics 15, 517-528.

*111. Snee, R.D. and Marquard, D.W. (1974) Extreme vertices designs for linear mixture models. Technometrics 16, 399-408.

*112. Snee, R.D. (1975) Experimental designs for quadratic models in constrained mixture spaces. Technometrics 17, 149-159.

*113. Snee, R.D. and Marquard, D.W. (1976) Screening concepts and designs for experiments with mixtures. TEchnometrics 18, 19-29.

114. Sobel, M. and Huyett, M. (1957) Selecting the best one of several binomial populations. Bell system Tech. J. 36, 537-576.

115. Sprott, D.A. (1958) The method of maximum likelihood applied to the Poisson binomial distribution. Biometrics 14, 97-106.

116. Steutel, F.W. (1968) A class of infinitely divisible mixtures. Ann. Math. Statist. 39, 1153-1157.

117. Strömgren, Bengt (1934) Tables and diagrams for dissecting a frequency curve into components by the half-invariant method. Skand. Aktnarietid. Skr. 17, 7-54.

118. Tan, W.Y. and Chang, W.C. (1971) Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of normal densities. Biometrics 27, 489.

119. Tan, W.Y. and Chang, W.C. (1972) Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. J. Amer. Statist. Assoc. 67, 702-708.

120. Teicher, H. (1960) On the mixture of distributions. Ann. Math. Statist. 31, 55-73.

121. Teicher, H. (1961) Identifiability of mixtures. Ann. Math. Statist. 32, 244-248.

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

*123. Thompson, W.C. and Myers, R.M. (1968) Response surface designs for experiments with mixtures. Technometrics 10, 739-756.

124. Thomas, E.A.C. (1969) Distribution free tests for mixed probability distributions. Biometrika 56, 475-484.

125 Tubbs, J.D. and Coberly, W.A. (1976) An empirical sensitivity study of mixture proportion estimators. Commun. Statist. A 5(13), 1115-1125.

126. Wolfe, J.H. (1970) Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Res. 5, 329-350.

127. Yakowitz, S. and Spragins, J. (1968) On the identifiability of finite mixtures. Ann. Math. Statist. 39, 209-214.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Mimeograph Series #79-22	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) On Mixtures of Distributions: A Survey and Some New Results on Ranking and Selection		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER Mimeo. Series #79-22
7. AUTHOR(s) Gupta, S. S. and Huang, W. T.		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS Purdue University Department of Statistics West Lafayette, IN 47907		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC		12. REPORT DATE August 1979 13. NUMBER OF PAGES 64
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release, distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Mixture of distributions, identifiability, infinite divisibility, atomic, perfect mixtures, moment estimates, maximum likelihood estimates, Bayes estimates, least square estimates, uniformly minimum variance estimates, minimax unbiased estimates, testing for mixture of distributions, ranking and selection for the mixture of distributions.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper surveys some of the literature dealing with mixtures of distributions. The topics covered relate to probabilistic properties, estimation, hypotheses testing and multiple decision (selection and ranking) procedures. The results reviewed concerning probabilistic properties of mixture distributions include the identifiability, scale mixture, infinite divisibility, atomicity and perfectness. The results on estimation theory reviewed include the method of moments, method of maximum likelihood estimation, method of least squares, Bayesian estimation, and the method of curve fitting. The

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered) cont

results for hypotheses testing provide tests for hypothesis whether an observed sample is a mixture from two samples with certain unknown proportion and also provide test if the mean of the mixture population is equal to some known value. In the last section, we give some new results for selection and ranking procedures for mixtures of distributions.