A TEST FOR ESTIMATING THE NUMBER OF COMPONENT DISTRIBUTIONS IN A NORMAL MIXTURE DISTRIBUTION AND THE RESULTS OF A MONTE CARLO SIMULATION OF THE TEST

By
MARK WAYNE PALKO
Bachelor of Fine Arts
Arkansas Tech University
Russellville, Arkansas
1985

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 1993

OKLAHOMA STATE UNIVERSITY

A TEST FOR ESTIMATING THE NUMBER OF COMPONENT DISTRIBUTIONS IN A NORMAL MIXTURE DISTRIBUTION AND THE RESULTS OF A MONTE CARLO SIMULATION OF THE TEST

Thesis Approved:

Mark E Payton

Thesis Adviser

Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to express my sincere thanks to my report adviser, Dr. Mark Payton for his support, guidance, and, most of all, patience throughout my graduate program. I also wish to thank the members of my committee, Dr. David L. Weeks and Dr. William Warde, and my stand-by, Dr. Larry Claypool.

Of the many helpful people at OSU, I need to single out Robert Steiner and Bob Wilson, whose expertise, insights, and time saved me from a great deal of grief. Special thanks also go to Iris McPherson, who helped me adapt my simulation to the mainframe.

Off campus, I would like to thank Jennifer Gregory for her continued support. And finally, thanks to my family: my parents, Tom and Hilda, and my brothers, Carl and James, who kept faith in me despite all evidence to the contrary.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The effectiveness of certain drugs is strongly affected by qualitative factors such as gender, racial make-up or blood type. For this reason, experimenters studying these and similar drugs usually take these factors into account by analyzing the results from each group separately, partitioning out these factors in the analyses of variance of the results, or including appropriate indicator variables in their regression models.

Often, however, the importance of a qualitative factor is not suspected before or during an experiment, so no levels of the factor are recorded. This information is lost and the observations from the different levels of the factor are pooled together. This can produce a multimodal distribution, which leads to a variety of problems.

For example, the effects of a drug on different groups can cancel each other out in tests such as the paired-t. This occurs when a treatment has a positive effect on one group and a negative effect on another, so the overall mean effect is near zero, even though all of the subjects may have been strongly affected by the treatment. This illustrates how a Type II error may occur.

Even if the effects are not contradictory, multimodality can still lead to Type II errors by artificially inflating the estimated variance. Even if all of the groups show a significant effect at a specified level of a Type I error (α), the inflated error estimate can cause the pooled sample to be declared non-significant using the same α.

CHAPTER II

LITERATURE REVIEW

Finite Mixtures

The aforementioned examples are members of a special case of the class of distributions called *mixture* or *compound distributions.* A mixture distribution is a statistical distribution which can be expressed as a superposition of (usually simpler) component distributions.

One example of a mixture distribution is the quantity of rice on an acre of land, using the distribution of rainfall in that area as one component and the distribution of rice growth for a given amount of rainfall as the other component. The resulting distribution could be expressed as:

$$f(rice) = \int g(rice\ ;\ rain)\ h(rain)\ \mathrm{d}\ rain, \tag{1}$$

where $g(rice\ ;\ rain)$ is the conditional distribution of rice growth on rainfall and $h(rain)$ is the rainfall distribution in that area. Since $h(rain)$ is a continuous distribution, the probability density function of rice growth has been expressed here as an infinite superposition of conditional density functions.

If, instead of using rainfall, we had based the probability density function (p.d.f.) on whether or not the soil was fertilized, the marginal density function of rice growth would have been

$$f(rice) = g_1(rice\ ;\ nofert)Pr(nofert) + g_2(rice\ ;\ fert)Pr(fert). \tag{2}$$

The expression $g_1(rice\ ;\ nofert)$ is the density of the conditional distribution of rice growth given that fertilizer was not applied, $g_2(rice;\ fert)$ is the density of the distribution of growth given that fertilizer was applied, and $p_1(nofert)$ and $p_2(fert)$ are the probabilities of fertilizer not being or being applied, respectively, with the constraint that

$$Pr(nofert) + Pr(fert) = 1. \tag{3}$$

Because the number of component distributions in this expression was finite, distributions such as these are called *finite mixture distributions*, and they are the general focus of this paper. The specific focus is a special case of finite mixture distributions, *finite normal mixtures*.

Finite normal mixtures occur when the conditional distribution of a random variable, for any given value of a qualitative factor, is a normal distribution. Since most populations are affected by some qualitative factors, finite normal mixtures are almost as common as normal populations are. In fact, any time a valid F-test is performed in an analysis of variance, the distribution of samples pooled over treatments is a finite normal mixture.

There is no mathematical difference between the distribution of samples in a standard analysis of variance and those analyzed by finite mixture techniques. The difference is contextual. To perform an analysis of variance, the experimenter must have enough additional information about the experimental units to partition the down into subsamples, such that each subsample is drawn from a normal population. When finite mixture techniques are used, it is assumed that the information necessary to break down the distribution is not available, either because it was not observable, or was simply not gathered.

The finite normal mixture of K populations has the p.d.f.

$$\sum_{i=1}^{K} p_i \phi(x;\mu_i, \sigma_i), \tag{4}$$

where $\phi(x;\mu_i, \sigma_i)$ is the p.d.f. of a normal distribution with mean μ_i and standard deviation σ_i, and where p_i is the probability associated with the ith component population. In some contexts, it is more useful to define p_i as the proportion of the population contained in the ith component population. As a consequence of either of these definitions, the following two properties hold:

$$0 < p_i < 1, \qquad \text{and} \qquad \sum_{i=1}^{K} p_i = 1. \tag{5}$$

For a given K, there are (3K - 1) unknown parameters in the p.d.f. Specifically, K values for the μ's, K values for the σ's, and (K - 1) values for the p's (the last value being uniquely determined by the remaining values).

The value of K, however, is often not known *a priori*. In this case, the number of unknown parameters (3K) is itself dependent on the value of one of the parameters.

This large, undetermined number of parameters can make the problem of estimating parameter values unmanageable. For this reason, the problem is usually simplified by limiting the possible cases.

One common solution is to set the value of K, based on inspection or some *a priori* knowledge of the population. Another is to assume that the component populations have equal variance. The solution used in this study is to assume equal variances and to place restrictions on the relative sizes of the μ's.

It is useful, at this point, to define a new parameter which focuses more narrowly on the properties of interest. Since the modality of a population (and the test statistic proposed in this report) is invariant to location, a variable is needed that gives the relative positions of the component means and is also invariant with respect to location. Such a variable can be defined using the differences between the means.

Given that, for any i from 1 to (K - 1):

$$\mu_i < \mu_{i+1}; \tag{6}$$

define the difference variable (d_i) as

$$d_i = \mu_i - \mu_{i+1}. \tag{7}$$

The p.d.f. can now be rewritten:

$$\sum_{i=1}^{K} p_i \phi(x; \quad \{\mu_1 + \sum_{j=1}^{i} d_j\}, \quad \sigma_i). \tag{8}$$

Assuming the following

$$d_i = d_j = d, \text{ for all i and j}, \tag{9}$$

$$p_i = p_j = \tfrac{1}{K}, \text{ for all i and j}, \tag{10}$$

$$\sigma_i = \sigma_j = \sigma, \text{ for all i and j,} \tag{11}$$

gives us the following p.d.f.:

$$\frac{1}{K} \sum_{i=1}^{K} \phi(x;\{\mu_1 + (id)\}, \sigma). \tag{12}$$

Existing Methods for Estimating K

The problem of estimating values for μ_i and σ_i has parallels in the analysis of variance context. The values p_i and K, on the other hand, have to be known before an analysis can be performed. The problem of estimating p_i occurs indirectly in discriminate analysis and the problem of estimating K also occurs (in a multivariate context) in cluster analysis.

The hypotheses for testing modality in finite mixtures are not, in most cases, well defined. This is due to the number of combinations of parameters that can produce the same distribution. In the case of a finite normal mixture, a k-modal distribution approaches a (k - 1)-modal distribution when one of the differences between the means, or the proportion of one of the component distributions approaches zero.

For the purpose of developing a test for modality, equal differences, proportions and variances were used. The robustness of this procedure to deviations from these assumptions (component normality, equal component variance and equal spacing of component means) will have to be answered in later research.

A number of techniques have been proposed to estimate the smallest value of K that satisfies the normality conditions of the finite normal mixture. Of these, the most important are inspection (graphic) techniques, the method of moments, and the likelihood ratio test.

Inspection techniques are not designed to test hypotheses about K. Their purpose is to present the data in a way that helps the experimenter recognize signs of a mixture distribution. These techniques fall into two categories: *histograms* and *probability plots.*

Histograms are an obvious choice when trying to determine the nature of a distribution, being easy to set up and to explain, and familiar to almost everyone. Unfortunately, multimodality is often difficult to detect from a histogram. Murphy (1964) gave examples of samples of size 50 taken from a single normal ($K=1$). Many of these gave the impression of multimodality.

The simplest type of probability plot useful for detecting mixtures of finite normals is a plot of the sample quantiles against the theoretical quantiles of the standard normal curve (Everitt and Hand, 1981). If samples are taken from a single ($K=1$) standard normal distribution, their plots should tend to be approximately linear, while mixtures will tend to produce curves. A more sensitive plotting technique was developed by Fowlkes (1979). His technique plots the standardized sample quantiles against

$$\Phi((y_{(i)} - \bar{y})/s) - p_i, \tag{13}$$

where

$$p_i = (i - 1/2)/n, \tag{14}$$

$$y_{(1)} < y_{(2)} < \ldots < y_{(n)}, \tag{15}$$

and $\Phi((y_{(i)} - \bar{y})/s)$ is the probability of $z < (y_{(i)} - \bar{y})/s$, where z is a standard normal variate.

The method of moments was first used by Karl Pearson in 1894 for the case when $K=2$ (Fowlkes, 1979) and has since been extended to $K=K_0 \geq 2$.

The likelihood ratio is probably the best known formal test for determining the number of components in a mixture, or, more formally, comparing the hypothesis $K=K_0$ against the alternative hypothesis $K=K_1$ $(K_0 < K_1)$.

These hypotheses are tested by computing the likelihood ratio, λ, given by

$$\lambda = \mathscr{L}_{k1} / \mathscr{L}_{k2}, \tag{16}$$

where $\mathscr{L}_{k1}$ and $\mathscr{L}_{k2}$ are the likelihoods, respectively, of the sample coming from the distribution described in the null hypothesis and from the distribution described in the alternative hypothesis. Given this ratio, the power and the probability of Type I error of the test can be calculated, provided the sampling distribution of λ under the null hypothesis is known. This, unfortunately, is seldom the case. Researchers have been searching for the asymptotic distribution of λ since the 1930's and, as yet, have only found special cases (Everitt and Hand, 1981).

From a practical point of view, likelihood ratios have an even more serious problem: time. Since the likelihood equations for normal (and many non-normal) mixture distributions are typically algebraically complex, non-linear functions of the mixture parameters, iterative procedures that require extensive computing are needed to find the maximum likelihood estimators (MLE's). Furthermore, the likelihood surfaces may have multiple local maxima or saddle points. This means that the algorithms may converge very slowly, or perhaps not converge at all. Even if they do converge, there is no guarantee that the resulting estimate is a global maximum, only that a maximum is found from that particular starting value (Furman and Lindsay, 1992).

A third approach is the method of moments. This method uses the moments (and functions of the moments) to estimate parameters of the distribution from which the sample was drawn. The currently used version of this test involves a matrix of moments, which is constructed by putting 1 (the expected value of X^0) in the upper left hand corner, using the first moment as the next diagonal element, using the second moment as the next diagonal element, and continuing until the $2p^{th}$ moment is placed in the lower right hand corner. The structure of the resulting matrix "reveals information about the number and location of the support points for a discrete distribution..." (Lindsay, 1989). The term, "number of support points," in Lindsay's paper is analogous to the number of component distributions in this paper. The determinants of the matrices are used to analyze the results.

W. David Furman and Bruce G. Lindsay (1992) offered two procedures based on the method of moments and argued for the general superiority (or, at least, equivalence) of these methods over the traditional likelihood ratio procedures. They work with the case of normal subpopulations with equal variance, and use the following hypotheses:

$$H_0{:}K = K_0 \qquad \text{vs.} \qquad H_1{:}K = K_0 + 1.$$

In order to find a statistic to test these hypotheses, Furman and

Lindsay start by finding a moment matrix for the observed data and use this matrix to estimate the variances of the component distributions (σ_i^2) under both the null and alternative hypotheses. The estimated values are written $\hat{\sigma}_p^2$ and $\hat{\sigma}_{p+1}^2$ respectively. The test statistic, called a pseudo F, is

$$F_p^* = ln(\hat{\sigma}_p^2/\hat{\sigma}_{p+1}^2). \tag{17}$$

The logarithm of F^* was used "with the goal of stabilizing the limiting distribution." Note that this statistic is location and scale invariant (Furman and Lindsay, 1992).

Under the null hypothesis of p subpopulations, the expected value of the ratio $(\hat{\sigma}_p^2/\hat{\sigma}_{p+1}^2)$ is one. Under the alternative, however, the expected value of the ratio is greater than one. The value of $\hat{\sigma}_p^2$ is larger in this case because the difference between two of the means is included in this estimate. This idea is examined more fully in the next chapter.

Furman and Lindsay draw an analogy between this procedure and the more familiar analysis of variance. Both tests measure the reduction of variance caused by reassigning the observed data points to new groups and pooling the resulting variances. Unlike the analysis of variance, however, there are not pre-existing criteria for assigning a given observation to a given group, so the estimation of the factors determining these criteria adds another level of difficulty to the problem.

Though the likelihood ratio test is usually slightly more powerful than the methods of moments test, it involves a slower procedure, making the method of moments test a more practical choice in most situations.

Engleman and Hartigan (1969) developed a variation of the likelihood ratio to test for bimodality. This variation is of special relevance to the method proposed in the next section. Their method finds the grouping that maximizes the ratio of variance between groups to variance within groups for k groups by calculating each of the different combinations that satisfy the requirement that all groupings be contiguous. They refer to the maximum ratio as the F-max. This statistic is not actually F-distributed except in the extreme case where the samples from each component distribution have no overlap.

F-max is used as a statistic for testing the hypothesis of unimodal normality against the alternative hypothesis of bimodal normality. The possibility of a compound alternative hypothesis, which would have included trimodal normality, was not considered.

CHAPTER III

A PROPOSED METHOD FOR ESTIMATING THE NUMBER OF COMPONENT POPULATIONS

This paper proposes a method which (after limited Monte Carlo simulations) seems to have power comparable to the two established tests and with potentially greater efficiency than either. This test has a theoretical basis similar to that of the method of moments, but approaches the problem from a different direction.

Variance Inflation

Variance, as a measure of dispersion, is affected by such distributional properties as skewness and multimodality. Given this, the use of variance when distributions are multimodal may lead to questionable results. In a sense, variance estimates in these situations are *inflated* when compared to other measures of dispersion such as range.

This property of the class of multimodal distributions considered here can cause serious problems for experimenters, particularly when the difference between component means is large relative to the component variances. It can also be used as the basis for a test of multimodality.

The Bimodal Normal

The simplest class of multimodal normal distributions is the bimodal normal, which has the p.d.f.:

$$f(x) = p\ \phi_1(\mu_1, \sigma_1) + q\ \phi_2(\mu_2, \sigma_2), \tag{18}$$

$$\text{where } p + q = 1. \tag{19}$$

If each component has equal proportion and variance, the p.d.f. simplifies to:

$$f(x) = \frac{\phi_1(\mu_1, \sigma) + \phi_2(\mu_2, \sigma)}{2}. \tag{20}$$

Let

$$f_1 = \phi_1(\mu_1, \sigma); \tag{21}$$

$$f_2 = \phi_2(\mu_2, \sigma). \tag{22}$$

All simulations were performed on this simplified case and, unless otherwise noted, all further discussion about the bimodal normal assumes equal variances and proportions.

Compared to a unimodal distribution with the same variance, a bimodal distribution with approximately equal component weights can have a range less than one half as large, depending on the variance of the component populations and the difference between the means of the subpopulations.

This property can be confirmed by considering the degenerate case of the bimodal distribution. If the variances of the component populations are negligible relative to the difference between the component means, the bimodal normal degenerates into a binomial distribution with the appropriate scaling parameter, which has a range equal to twice the standard deviation assuming equal proportions.

The variance of a symmetric bimodal is

$$\frac{d^2}{4} + \sigma^2 \tag{23}$$

where

$$d^2 = \left|(\mu_1 + \mu_2)\right|. \tag{24}$$

The formula for variance is somewhat more complicated when equal proportions and variances are not assumed.

If a bimodal mixture of normal distributions is divided at the point (a), the following condition holds:

$$f_1(a) = f_2(a). \tag{25}$$

The resulting subdistributions are roughly mound shaped and, in some cases, almost symmetric. The pooled variance of the two subdivisions will be less than or equal to the variance of the component distributions (σ^2), depending on the degree of overlap. The squared difference terms drop out entirely. A large difference between the variance of the original distribution and the pooled variance of the subdivisions indicates a large difference between the component means. Consequently, the difference between the estimated variance of a whole sample and the estimated pooled variance of subgroups of that sample could form a reasonable basis for a test of modality.

We also see a similar reduction in pooled variance when going from a trimodal normal divided into two subdistributions to three subdistributions.

Using this principle, a test statistic may be constructed for testing the hypothesis of H_0:K = K_0, vs. H_a:K > K_0. The test statistic is

$$M = \frac{\text{Pooled Variance (K)}}{\text{Pooled Variance (K+1)}},$$

where K is the smallest number of normal component distributions with which the original distribution can be written.

Dividing the Distribution

In the bimodal case, the optimum point for dividing a distribution into two subgroups occurs at the point of equal likelihood, defined as the point that has an equal likelihood value for each of the normal subcomponents (which occurs at the point $f_1(a) = f_2(a)$, described earlier in this chapter). Since the normal distribution is continuous, the point of equal likelihood will occur somewhere between the observed values of the sample.

The likelihood of an observation being from a certain component is a decreasing function of the distance between a given point and the mean of that component. For that reason, the values falling in the group with the smaller mean will all be smaller than any value from the group with the larger mean.

W.D. Fisher uses this fact to develop a system for grouping observations for maximum homogeneity. His method finds the grouping that minimizes the pooled variance of the groups. Fisher showed that any grouping that satisfies this condition must consist of nonoverlapping groups, so only partitions of the set of ranked observations need to be considered. Without this property, the pooled variance of all possible combinations would have to be calculated, making the procedure impractical for most cases.

In order to use Fisher's procedure, the number of groups has to be decided before the calculations commence. The procedure cannot be used to test hypotheses about the optimum number of groupings. It does, however, provide the basis for Engleman and Hartigan's test (1969), described in the previous chapter, and that test forms an important part of the test proposed in this chapter.

Engelman and Hatigan's test provides a simple method for locating the the region between two consecutive observed values that contains the point of equal likelihood.

Engleman and Hatigan limited their work to K=1, but the basis of the work, established by Fisher, is valid for all values of K. The method proposed in this chapter uses F-max to estimate the points of equal likelihood, then, unlike Engleman and Hartigan's test, compares the behavior of certain test statistics for different numbers of groups.

Having divided the sample into k groups, the pooled variance of the k groups (s_k^2) can be compared to the pooled variance of k-1 groups (s_{k-1}^2). This comparison is achieved by the ratio

$$R = \frac{s_k^2}{s_{(k-1)}^2} \quad . \tag{27}$$

The Procedure

The first step in this procedure is to rank the observed values from smallest to largest and apply Engleman and Hartigan's method. Starting with the smallest observation in group one and the remaining observations in group two, calculate the error mean squared (MSE) for the two groups. Then find the mean of each group and calculate the between group variance. Calculate the ratio of these two numbers. Repeat the process with the first two observations in the first group and the remaining observations in the second. Repeat for all (n-1) combinations.

Consider the following example. A sample of size 24 is taken. Only one variable is recorded, but on inspection of the data, the experimenter suspects that the population sampled may not be unimodal. If this is the case, he may have to conduct another experiment, this time including additional variables that might explain the multimodality.

Here are the results of the first experiment ranked in ascending order:

Observed values: -2.48, -1.51, -0.97, -0.83, -0.37, -0.18
-0.05, 0.2, 0.25, 0.32, 0.7, 1.28, 1.3,
1.37, 2.08, 2.19, 2.66, 2.72, 3.11, 3.83,
3.83, 4.0, 4.11, 4.24.

The variance of this sample is 3.805 (s_1^2). The pooled variances of the groups will be less than 3.805. If one of the pooled variances is greater than the variance of the sample taken as a whole, a mistake has been made in the calculations.

The first step is to calculate the F statistic for each nonoverlapping partition. The first partition is:

First group: -2.48.

Second Group: -1.51, -0.97, -0.83, -0.37, -0.18, -0.05, 0.2, 0.25, 0.32, 0.7, 1.28, 1.3, 1.37, 2.08, 2.19, 2.66, 2.72, 3.11, 3.83, 3.83, 4.0, 4.11, 4.24.

The F statistic for these two groups is 4.608. The pooled variance of the two groups is 3.29, only slightly lower than the variance of the sample taken as whole. This is to be expected because pooled variance is a weighted average and, though the first group yields no estimate of variance and is treated as zero in the calculations, the larger weight of the second group largely cancels out the reduction in variance caused by the first group.

If 4.608 had been the largest F value, the statistic for testing bimodality would be the ratio of 3.29 and 3.805. Other partitions, however, prove to have considerably larger values.

When -1.51 is taken from the second group and included in the first group, the value of F increases to 8.368 and the pooled variance decreases to 2.88.

The maximum value of F is reached when the sample is partitioned as follows:

First Group: -2.48, -1.51, -0.97, -0.83, -0.37, -0.18
-0.05, 0.2, 0.25, 0.32, 0.7, 1.28, 1.3, 1.37.
Second Group: 2.08, 2.19, 2.66, 2.72, 3.11, 3.83,
3.83, 4.0, 4.11, 4.24.

The value of F for this partition is 64.88. The pooled variance of the groups is 1.01.

The grouping that produces the largest ratio is used to estimate the pooled variance. This doesn't involve any additional calculation since the pooled variance had already been used as the denominator of the ratio.

In the case of the earlier example, the statistic for testing the hypothesis of unimodality against the hypothesis of bimodality is:

$$R = \frac{1.01}{3.805} = 0.265 \tag{28}$$

A test statistic value this low would indicate that the experimenter probably did take his sample from a multimodal distribution.

From this point, the process is similar to the method of moments test, but with one important difference in interpretation. Both tests use ratios of variance estimates corresponding to the hypothesized number of components (k) and one more than the hypothesized number (k+1), and in both tests, a large difference from k to k+1 is significant. In the method of moments test, however, the expected value of the ratio under the null is one, while in the proposed test, the ratio will always be less than one, regardless of the distribution.

This property is a result of the contiguous arrangement of the groups. With this arrangement, each cut creates an additional group, and since the groups are nonoverlapping, more groups mean smaller pooled variance.

Since the expected value of the proposed test statistic is a function of the number of cuts, p-values will have to be calculated individually for each k. This problem may also occur with the method of moments test. Furman and Lindsay only ran simulations for $k=1$ and $k=2$. Other factors, such as sample size and possibly proportion, will also have to be taken into account.

CHAPTER IV

THE SIMULATION

In order to test the effectiveness of this new test (and judge the merits of other test statistics), a series of small scale Monte Carlo simulations were performed using different combinations of values for sample size, difference in means, and modality. Each component of a given sample had equal variance and were in proportion. Each sample was standardized for convenience of interpretation. This was accomplished by dividing by the sample variance. This had no effect on the test statistic's values because the common multiplier canceled out in the ratio.

The program was written in SAS/IML and is included in the appendix. It tested for bimodality and trimodality. An additional subroutine was written to test for higher modality, but was not used in this simulation.

The program accomplished three primary tasks: it generated random samples from suitable distributions; it performed Engelman and Hartigan's method on the sample; it calculated the test statistic for the bimodality test. The last two tasks were repeated to find the test statistic for the trimodality test.

Generating the Sample

The distributions were broken into three categories according to modality (unimodal, bimodal and trimodal), all based on normal component distribution of equal variance (unity). Separate simulations were performed for each of the categories. The results of these three simulations were used to determine the power of the proposed test.

Each of the component distributions was generated individually using the random normal function of SAS/IML. For the univariate case, the component distribution was normal with mean zero and variance one (N(0,1)). For the bimodal case, the distributions were N(0,1) and N(Diff,1), where the values of Diff varied from 1 to 4 standard deviations. For the trimodal case, the component distributions were N(0,1), N(Diff,1) and N(2Diff,1), with the same range for the parameter Diff. Equal differences between means were chosen to make the analysis easier.

Each component was assigned an equal number of data points. The number of combined samples for each run was a multiple of 6 so results from the three distributions could be compared. The number of combined samples varied from 12 to 180.

Standardizing and Sorting the Sample

Though the variances of the component distributions were the same in all the simulations, the variances of the finite mixtures were not. The variance of a finite mixture density was much larger.

To accommodate for the discrepancy in variances, the distributions were standardized by dividing by sample standard deviation. This had no effect on the test statistic which, due to the properties of the ratio, was not affected by a multiplier. The procedure, however, was easier with the standardized samples.

After standardizing, the observations were sorted using a bubble sort. This algorithm was chosen for simplicity, not for speed.

Defining the Sum of Squares (Total) and the Correction Factor

Since each grouping of the sample will have the same total sum of squares, the same total degrees of freedom and the same correction factor, these quantities are calculated first and used throughout the remaining calculations regarding this sample. In the original version of this program, these quantities were calculated using matrix multiplication (in the form y'Ay). Unfortunately, the memory of the PC's on which these programs were written limited the number of matrices that could be declared in a single program. To avoid this limitation, the final program used the sum and sum of squares functions of SAS/IML.

Defining the Subgroups and Calculating the Pseudo-F

The next step in the test is to divide the sample into two subgroups and find the ratio of the variance between these two groups to the pooled variance within the two groups, then repeat the procedure for all contiguous combinations.

The subgroups are defined by partitioning the ordered sample vector into two smaller vectors. The sum of squares from these two vectors is called sstrt, which stands for treatment sum of squares. This is not a treatment sum of squares in the sense of a true F test. In that context, the samples would have to come from two independent distributions. That condition is not met by these two nonoverlapping samples. However, the calculation of this sstrt and its role in calculating the pseudo-F is analogous to that of treatment sum of squares for the true F.

The pooled variance of these two groups (mserr) is calculated by subtracting the treatment sum of squares from the total sum of squares and dividing by the appropriate degrees of freedom. The error mean square is more commonly calculated by taking the difference between corrected sums of squares, but since the same correction term is used for both total and treatment sum of squares, the term cancels out and the end result is the same.

The correction term is then subtracted from the treatment sum of squares and the difference is divided by the appropriate degrees of freedom to produce the mean square of the treatment (mstrt). The ratio of treatment mean square and the error mean square is the pseudo F for that particular combination.

Finding F Max

The proposed method uses the mean square error term associated with the largest value of the pseudo F, so, for each new combination, the F is compared to the largest previous F. If the new F is larger, its value and the corresponding error mean square are substituted for the previous values.

This algorithm was written for simplicity rather than speed and many streamlining steps have been ignored. One such step is to start the procedure at some point in the middle of the sample and work toward the side with the larger variance. This step alone would reduce the time required by this procedure by about half.

The First Test Statistic

The generalized form of the proposed test statistic for testing the null hypothesis ($K \leq K_0$) against the alternative hypothesis ($K > K_0$) is

$$\frac{MSE(B)}{MSE(A)} , \tag{29}$$

where MSE(A) is the error mean square of the sample when partitioned into K_0 subgroups and MSE(B) is the error mean square of the sample when partitioned into (K_0+ 1) subgroups.

In this case, MSE(B) is the error mean square of the sample when partitioned into two subgroups and MSE(A) is the error mean square of the sample before any cuts were made. In other words, MSE(A) is the sample variance. However, since the sample had already been standardized with respect to the sample variance, MSE(A) is unity and the first test statistic is simply MSE(B).

Dividing the Sample into Three Subgroups

The procedure for partitioning the sample into three subgroups uses the same steps used to partition the sample into two subgroups, but adds a subroutine. For each combination of two subgroups except (n - 1) and 1, a nested loop divides the second subgroup into two additional subgroups, running through all possible combinations of three subgroups.

This step is the most time consuming part of the program, due to the sheer number of combinations of three subgroups. All possible combinations are tried. No information from the first series of divisions is carried over to the second series, though only the variance of the middle subgroup is being calculated for the first time. This time-saving information was ignored to keep the program simple and to stay within the memory constraints of the PC.

Other than the additional number of cuts, the only differences in this part of the program are the degrees of freedom and the number of variance terms pooled in the error mean square. The same algorithm is used to find the largest F and the corresponding error mean square.

The Second Test Statistic

The second test statistic is the ratio of the two error mean squares found by this program. This ratio is labeled <u>ereduct</u>, which stands for error reduction. The difference between the two variances was also included in the output under the name <u>gain</u>.

CHAPTER V

ANALYSIS OF THE RESULTS

Methods of Analysis

Three primary methods were used to analyze the results of the Monte Carlo simulation: a number of regression models were written (using PROC REG on SAS) to check which variables or combinations of variables showed a significant relationship to modality; graphs and quantiles were used to provide an idea of the usefulness and power of different tests.

Regression Models

Multiple regression models were constructed for a number of the potential test statistics. The primary purpose was to see if the test statistics of samples taken from unimodal, bimodal and trimodal populations behaved differently. To test this, dummy variables were included corresponding to the different modalities.

Initially, these models were applied to all of the results of the simulation. In later stages, the results were grouped into pairs of populations to determine how well the test statistics differentiated between unimodal, bimodal and trimodal distributions.

The secondary purpose of the models was to determine the relationship of the other parameters (sample size and difference in means) on the test statistic. This was complicated by the fact that difference was related to population modality. The samples from unimodal populations all had a difference of means equal to zero.

Results of the Regression Analysis

The test statistics that were tried included the two error terms, the F statistics, the ratio of the errors, the sum of the errors, and various combinations of these quantities. As expected the dummy variable differentiating between the unimodal and multimodal populations was most significant for the test statistics derived from the first series of cuts (those dividing the population into two subpopulations). The dummy variable distinguishing unimodal and bimodal populations from trimodal populations was most significant for test statistics that used information from both series of cuts. The results were particularly encouraging for the ratio of the two error terms.

Sample size (n) had a comparatively small effect on the means of all the test statistics considered, but it did have a significant effect on the variance of the statistics. All of the test statistics showed their highest variance at n = 12, and their smallest variance at n = 240. For this reason, further analysis (and some additional simulations) omitted the smaller sample sizes and concentrated on samples of size greater than or equal to 60.

All test statistics showed a significant relationship to the difference in means of the component populations. This indicated that the larger the difference between component means, the easier the difference in modality is to detect.

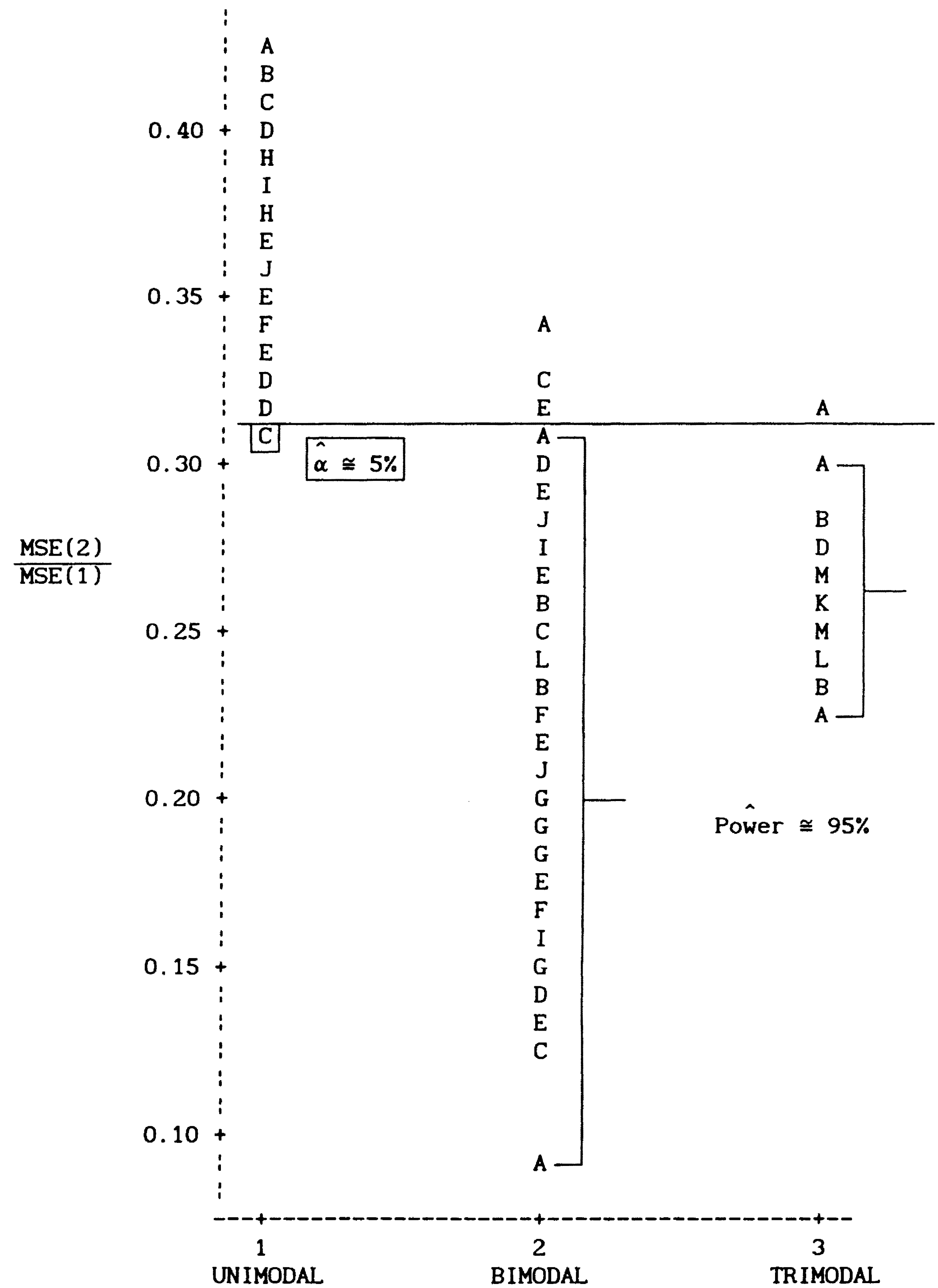
Graphs and Quantiles

A number of graphs were plotted based on the results of the regression analysis. In these graphs, the value of the given statistics was plotted against the modality of the sample (see figures 1 and 2). Since the object of the test is to use these values to determine modality, this type of graph summarizes the most important aspect of the data in the simplest fashion.

The graphs included in this thesis are the two test statistics of the proposed method (ratios of consecutive mse's). The graphs of the two proposed statistics suggest that the proposed test shows sufficient power to merit further study.

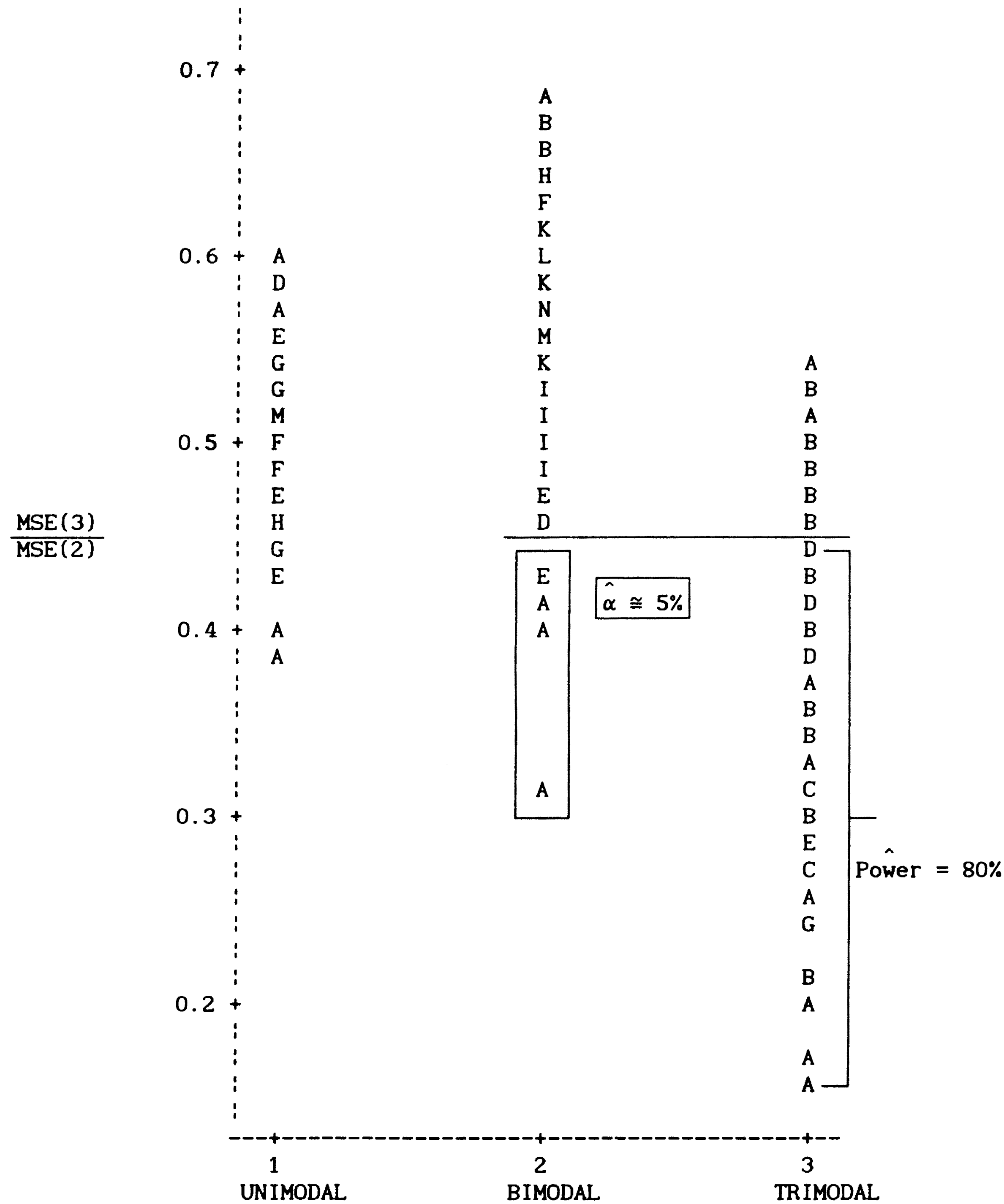
The graphs, though useful for suggesting relationships and power, can only approximate the actual quantities. For a more accurate estimate of power, the quantiles of the different modalities need to be compared. Regardless of the technique, the power estimates based on this data will have limited accuracy, due to the small number of trials in this study.

Based on the results of the regression analysis and the graphs, additional simulations were run concentrating on certain distributions. These included unimodal normals of size $n = 60$, and bimodal normals of size $n = 60$ and difference $= 2.5$ or 3.5.



Legend: A = 1 obs, B = 2 obs, etc.
n = 72 to 144 and difference between means>1.5

Figure 1. Approximate Power of the Proposed Method for Testing the Null Hypothesis of Unimodality against the Alternative Hypothesis of Multimodality.



Legend: A = 1 obs, B = 2 obs, etc.
n = 72 to 144 and difference between means>1.5

Figure 2. Approximate Power of the Proposed Method for Testing the Null Hypothesis of Bimodality against the Alternative Hypothesis of Trimodality.

The quantiles of interest are those in the overlap between the sample which was taken from population(s) in the null hypothesis and the sample taken from the population(s) in the alternative hypothesis. For the first test statistic, the unimodal population is in the null and the bimodal and trimodal populations are in the alternative. For the second test statistic, the bimodal population is added to the null, leaving the trimodal population in the alternative.

By comparing the percentile in the null sample of a given value with the percentile that value would have in the alternative sample, the power $(1 - \beta)$ is estimated for the α that corresponds to the percentile of the null sample.

For example, to estimate the power of a one sided, $\alpha = 5\%$ test, find either the 5th percentile or the 95th percentile, depending on the direction of the test, of the sample representing the null hypothesis (in this case, the 5th percentile). Find the percentile (call it the r^{th} percentile) in the sample representing the alternative hypothesis that corresponds with the percentile from the null. The estimate of the power for $(\alpha = 5\%)$ is r% if the 5th percentile of the null sample was used and (100 - r)% if the 95th percentile was used (power = r%, in this case).

Based on the quantiles of the later simulations, an $(\alpha = 5\%)$ test would have a power of almost 50% against an alternative (bimodal) distribution with difference = 2.5 σ and a power of more than 90% for a difference of 3.5 σ.

CHAPTER VI

SUMMARY AND CONCLUSIONS

Advantages and Disadvantages of the Proposed Test

Multimodal normal mixtures occur frequently in a wide variety of practical contexts. The failure to recognize multimodality can lead to a number of problems for the experimenter and data analyst, including a greatly increased probability of Type II error. Multimodality also causes problems in multivariate situations (particularly clustering) where variables of different units are standardized using the standard deviations. All of which indicates the need for a practical test of modality.

None of the existing tests (including the proposed method) have a known asymptotic distribution. Many other questions such as the sufficiency of the statistic are yet unanswered.

Although the exact distribution of the proposed test statistic is not known, a general relationship between partitioning and variance has been defined, and this relationship supports the observed behavior of the test statistic.

This method is also the only one described which addresses the problem of a composite null. When testing for three or more modes (K), the null hypothesis is $H_0\!:\!K \leq 2$ or, equivalently $K = 1$ or $K = 2$. The proposed method uses sequential testing: the test for K = 1 vs. K = 2 is performed first and the test for K = 2 vs K = 3 is applied only if the statistic from the first test was in the critical region. Even with the first test statistic in the critical region, there is still a chance of the population being unimodal (the p-value of the first test), but since this probability is known, its potential impact can be assessed.

Another potential advantage of the proposed method is its focus. This method measures the increase in variance (or error) caused by multimodality. The increase in variance causes a number of problems for the experimenter, including excessively wide confidence intervals, prohibitively large sample size requirements and a greatly increased probability of Type II error. This method, in a sense, "measures the damage" to an experiment caused by unrecognized multimodality. As a result, the multimodal distributions that this test fails to recognize are those distributions where the multimodality causes the fewest problems.

Though the small scale of this simulation limits the inferences that should be drawn about the proposed test, the initial results indicate a potentially useful test meriting further study.

The next step is a larger scale simulation sufficient to construct probability tables for the test statistics and describe power as function of the difference between means of component distributions. Additional simulations will be needed to study the effect of unequal component variances and proportions, as well as unequal differences in means in the trimodal case. Additional simulations are also needed to investigate the behavior of the test statistics under other distributions, such as quadmodal and uniform (which can be considered the limiting distribution as k approaches infinity).

Suggestions for Further Study

Investigate the possibility of a new test based on comparisons of variance and other measures of dispersion, such as range, quantiles, or interquartile range. Such a test could be investigated with a small modification to the program described in this paper and, if feasible, would be easier and quicker than any existing method.

Reexamine the Engelman and Hartigan test. The research done for this report has suggested modifications to the test which could increase the speed of the test.

Investigate the behavior of the proposed test statistic working directly from the finite normal mixture distribution. The goal here would be to define the distribution of this test statistic explicitly, rather than inferring the distribution from the results of simulations.

BIBLIOGRAPHY

[1] Engleman, L., and Hartigan, J.A. (1969). Percentage points of a test for clusters. *J.Amer.Statist.Assoc.* 64, 1647-48.

[2] Everitt, B.S., and Hand, D.J. (1981). *Finite Mixture Distributions.* Chapman and Hall. London.

[3] Fisher, W.D., (1958). On grouping for maximum homogeneity. *J.Amer.Statist.Assoc.* 53, 789-98.

[4] Fowlkes, E.B., (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *J.Amer.Statist.Assoc.* 74, 561-75.

[5] Furman, W.D., and Lindsay, B.G., (1992). Testing for the number of components in a mixture of normal distributions using moment estimators. *American Statistical Association Annual Meeting.* Boston.

[6] Lindsay, B.G., (1989). Moment matrices: applications in mixtures. *Annals of Statistics.* 17, 722-40.

[7] Murphy, E.A., (1964). One cause? Many causes? The argument from the bimodal distribution. *J.Chron.Dis.* 17, 301-24.

APPENDICES

APPENDIX A

SIMULATION PROGRAM

```
* F-max Master Program, Simplified (name output.prg);

* Corrected for Mean Square Mistake;

* With MSE tabulation;

*        and ratio gain in error;

* Standardized Variance;


proc iml;  reset noname; start main;

filename uuuu 'c:mwp2.dat'; file uuuu;
  n=18; seed=51; differ=2.5;
  dummy=0; dummy2=0;

* Two Populations partitioned by Two and Three Subpops;

* Create New Data Sets;

y=j(n,1,0);
do L=1 to 2;


    Do I=1 to (n/2);
    Y(|i|)=normal (seed);
    end;

    Do I=(n/2+1) to n;
    Y(|i|)=normal (seed) + differ;
    end;


* Standardize the variance;
    sd=sqrt((ssq(y) - (((sum(y))**2)/n))/(n-1));
    y=(y/sd);
    print '                                        ';
    print 'sd=' sd;
    print '                                        ';

* Sort the Observations;

  k=1;
      do until (k=n);
        if y(|k|)>Y(|k+1|) then do;
           d=y(|k|);
           y(|k|)=y(|k+1|);
           y(|k+1|)=d;
           k=1;
          end;
```

```
         else k=k+1;
       end;

* Define the Sum of Squares (total) and the Correction Factor;
  sstot  =  y`*y;
  cfact  =  sum(y)**2/n;

* Define Degrees of Freedom (total);
  dftot  =  n - 1;

* Make the cuts;

  do i=1 to (n-1);
    n1=i; n2=n-n1;
    y1=y[1:i];
    y2=y[(i+1):n];

* Define Sum of Squares (treatment);

    sstrt=(sum(y1)**2)/n1  +  (sum(y2)**2)/n2;

* Define Degrees of Freedom for treatment and error;
  dftrt=1;
  dferr=dftot - dftrt;

* Define Mean Squares for Treatment and Error;
  mstrt=(sstrt - cfact)/dftrt;
  mserr=(sstot - sstrt)/dferr;


* Define F;

         F=mstrt/mserr;

* Print results;

*     print L i F;


         if f>dummy then do;
           dummy=f;
           dummy2=mserr;
         end;
*          print 'dummy =' dummy;

  end;

* define Fmax;


fmax1=dummy;
    ferror=dummy2;
print  '        L=' L;
```

```
        print 'fmax1=' fmax1 'ferror=' ferror;
    dummy = 0;
    dummy2 = 0;


* Make the cuts for three subpops;

  do i=1 to (n-2);
    n1=i;
    y1=y[1:i];
      do j=1 to (n-i-1);
        n2=j; n3=n-n1-j;
        y2=y[(i+1):(i + j)];
        y3=y[(i + 1 + j):n];


* Define Sum of Squares (treatment) for three levels;

        sstrt=(sum(y1)**2)/n1 + (sum(y2)**2)/n2 +(sum(y3)**2)/n3;

* Define Degrees of Freedom for treatment and error;
  dftrt=2;
  dferr=dftot - dftrt;

* Define Mean Squares for Treatment and Error;
  mstrt=(sstrt - cfact)/dftrt;
  mserr=(sstot - sstrt)/dferr;


* Define F;

        F=mstrt/mserr;

* Print results;

*        print dftrt mstrt mserr F;

* Find Fmax2;

        if f>dummy then do;
          dummy=f;
          dummy2=mserr;
        end;
*        print 'dummy =' dummy;

    end;
  end;
    fmax2 = dummy;
    zzz = ferror;
    ferror=dummy2;
        print 'fmax2=' fmax2 'ferror=' ferror;
    dummy = 0;
    dummy2 = 0;
```

```
        gain = fmax2 - fmax1;
        ereduct = zzz/ferror;
        print 'Gain=' gain  'Error reduction ratio =' ereduct;
put @1 L 4.0 + 3 differ 3.1 + 3 n 5.0 + 3 zzz 8.7 +3 ferror 8.7;

end;
closefile uuuu;
finish; run main;
```

APPENDIX B

SIMULATION RESULTS

Results of the Proposed Procedure on
Samples Drawn from a Standard Normal Distribution

Size of Sample	F-Max for Two Groups	Error Ratio for Two Groups	F-Max for Three Groups	Error Ratio for Three Groups
24	54.4045	.3010296	59.8438	.1634828
48	144.7457	.2464014	150.5583	.1357924
72	146.9927	.3272000	171.6055	.1722419
96	170.7372	.3588465	207.8456	.1867537
120	172.3062	.4099120	258.3688	.1877749
144	279.1056	.3395823	338.6563	.1747498
24	27.5798	.4638983	34.3485	.2564188
48	96.7521	.3292422	103.2943	.1868129
72	147.8129	.3259678	205.5104	.1479103
96	189.3872	.3352303	213.0099	.1830373
120	203.0548	.3706532	228.8509	.2070639
144	279.6146	.3391723	386.6032	.1564198
24	28.8573	.4522457	78.2253	.1296135
48	103.1684	.3150801	108.3480	.1795977
72	101.0849	.4149986	138.3337	.2053998
96	147.7562	.3929578	196.8437	.1951972
120	186.6637	.3905946	279.2472	.1761672
144	252.3670	.3626064	314.0165	.1859478
24	29.0254	.4507556	31.4516	.2741253
48	115.2825	.2914141	126.4767	.1577428
72	111.9497	.3902178	145.7072	.1969955
96	169.7658	.3601680	252.5824	.1588191
120	219.1001	.3530109	290.6244	.1704264
144	308.9251	.3171258	355.7165	.1677551
24	65.3808	.2632157	118.8747	.0888891
48	74.3037	.3906778	116.6061	.1689358
72	110.4093	.3935496	166.0841	.1769831
96	148.2153	.3922130	236.2982	.1679643
120	244.2312	.3285194	347.0215	.1467247
144	249.4085	.3653472	308.7544	.1885278
24	49.1032	.3234733	85.2615	.1200900
48	109.9774	.3013257	150.5189	.1358233
72	118.8191	.3760213	125.3671	.2220595

96	179.7062	.3470875	196.5936	.1953980
120	217.6948	.3544887	229.5311	.2065749
144	253.7921	.3613008	338.8865	.1746516
24	71.9810	.2447303	123.0050	.0861391
48	125.8310	.2735245	138.9515	.1455546
72	140.7557	.3368829	144.2975	.1985486
96	186.5051	.3386747	229.1935	.1722928
120	196.3014	.3786174	281.0570	.1752283
144	213.4510	.4023058	312.0213	.1869177
24	55.9001	.2952500	84.8049	.1206653
48	76.3700	.3840811	112.8436	.1736321
72	134.3579	.3474297	155.3914	.1869489
96	185.0857	.3403973	221.5525	.1772041
120	175.4081	.4055784	221.4959	.2125031
144	268.0173	.3487657	299.8579	.1930565
24	50.1243	.3188940	84.5712	.1209619
48	89.8287	.3460242	156.1453	.1315456
72	160.3420	.3082374	156.7396	.1856310
96	187.9246	.3369695	265.8333	.1520811
120	192.2747	.3835311	260.6623	.1864255
144	277.4083	.3409565	333.9482	.1767841
24	50.2405	.3183808	63.5550	.1552900
48	70.4795	.4035046	92.1774	.2049226
72	127.8592	.3588410	176.2155	.1684736
96	157.3099	.3780193	197.4191	.1947367
120	214.9873	.3573710	289.5491	.1709529
144	217.5539	.3977150	311.9323	.1869612
24	43.4459	.3514353	55.1660	.1751287
48	93.1168	.3378455	127.1028	.1570826
72	105.9000	.4036384	151.9980	.1903506
96	131.8848	.4205683	210.4380	.1848695
120	233.3057	.3387363	291.9635	.1697752
144	214.5283	.4010901	280.3347	.2037997
24	33.6533	.4132731	36.6256	.2440287
48	117.6154	.2872590	132.1484	.1519576
72	103.0120	.4103761	129.5246	.2164310
96	177.9081	.3493827	187.7939	.2027368
120	189.2550	.3873005	279.0949	.1762467
144	226.4414	.3881214	272.0914	.2087035
24	44.6900	.3448791	53.4702	.1797711
48	109.2578	.3027223	143.1717	.1418468
72	119.6548	.3743643	152.0186	.1903296
96	161.5258	.3717824	215.6819	.1811719
120	247.7186	.3253868	286.4974	.1724651
144	233.3120	.3810163	374.4124	.1607058

24	36.9750	.3899954	82.0518	.1242547
48	70.8950	.4020703	107.9035	.1802099
72	152.9484	.3184593	184.3155	.1622372
96	156.4645	.3792952	279.3714	.1457630
120	164.2618	.4215944	234.7178	.2029208
144	260.2698	.3554828	392.4611	.1544406
24	39.7748	.3723200	72.7114	.1382022
48	48.1767	.4990620	61.0947	.2811182
72	116.9431	.3797948	137.9273	.2058838
96	151.3020	.3872777	188.3625	.2022460
120	261.6168	.3134740	294.1334	.1687305
144	257.4895	.3579568	330.3290	.1783803
24	48.1948	.3276595	91.7322	.1124891
48	78.5668	.3773076	87.9677	.2127318
72	158.2135	.3111122	171.9111	.1719869
96	188.7841	.3359454	245.1117	.1628878
120	190.9318	.3851983	251.1569	.1921481
144	243.0894	.3713424	365.0738	.1641513
24	42.4391	.3569264	57.3266	.1695501
48	75.4592	.3869611	101.2832	.1898481
72	143.8478	.3320118	200.1150	.1513117
96	152.0078	.3861667	189.2653	.2014716
120	203.2095	.3704747	300.6261	.1656800
144	259.9968	.3557242	318.4685	.1838195
24	45.8408	.3390291	71.0818	.1409629
48	89.1907	.3476570	120.5076	.1643270
72	111.8002	.3905387	171.8278	.1720563
96	172.3489	.3566751	268.1004	.1509852
120	259.1273	.3155433	291.1948	.1701484
144	243.4682	.3709775	304.5487	.1906419
24	49.9809	.3195293	56.4061	.1718826
48	99.4678	.3230954	102.8035	.1875446
72	157.9304	.3114985	177.6241	.1673549
96	158.1799	.3767152	172.7098	.2166874
120	191.1971	.3848677	328.2284	.1538547
144	225.9822	.3886057	355.3973	.1678808
24	37.9223	.3838306	62.1849	.1582172
48	114.2955	.2932085	123.1463	.1613498
72	123.3468	.3672157	141.0953	.2021694

VITA

Mark W. Palko

Candidate for the Degree of

Master of Science

Thesis: A TEST FOR ESTIMATING THE NUMBER OF COMPONENT DISTRIBUTIONS IN A NORMAL MIXTURE DISTRIBUTION AND THE RESULTS OF A MONTE CARLO SIMULATION OF THE TEST

Major Field: Statistics

Biographical:

Personal Data: Born in Dallas, Texas, November 5, 1962. the son of Tom and Hilda Palko.

Education: Graduated from Russellville High School, Russellville, Arkansas, in June 1981; received Bachelor of Fine Arts in Creative Writing from Arkansas Tech University in May, 1985; completed requirements for the Master of Science degree at Oklahoma State University in May, 1993.

Professional Experience: Teacher, County Line High School, Branch Arkansas, August, 1987 to June, 1988; tutor, Chrysalis Associates, Little Rock, Arkansas, September, 1988 to May, 1990; Teaching Assistant, Department of Statistics, Oklahoma State University, August, 1990 to May, 1992; Assistant Director, Mathematics Resource Center, University of Arkansas, August, 1992 to present.