NON-CONGRUENCE OF STATISTICAL DISTRIBUTIONS:
HOW DIFFERENT IS DIFFERENT?

by

TERRY LEE APPLEGATE

B.S., Kansas State University, 1977

A MASTER'S THESIS

Submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1978

Approved by:

Doris Grosh

Major Professor

ACKNOWLEDGEMENTS

I would like to thank my family, the faculty and personnel of the Department of Industrial Engineering at Kansas State, and my fellow students. Each of these groups helped make this segment of my education meaningful. Special thanks is given to Dr. Doris Grosh, whose faith and guidance were instrumental in my completing my course of studies and to my wife, Debora, without whose encouragement and understanding nothing would have been worthwhile.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	1.0. Genesis	1
	1.1. The Problem	1
	1.2. Purpose and Objective	1
	1.3. Method	2
	1.3.1 The Index of Non-Congruity - δ	2
	1.3.2. The Procedure	3
	1.3.3. The McGill Random Number Generator	8
II.	THE EXPONENTIAL CASE	10
	2.0. Introduction	10
	2.1. Determination of the Point of Intersection	10
	2.2. The Index of Non-Congruity: Exponential Case	12
	2.3. Description of Exponential Program Features	13
	2.3.1. Program Listing	13
	2.3.2. Definition of Program Variables	13
	2.3.3. Inputs to the Program	14
	2.3.4. Outputs of the Program	15
	2.3.5. Special Programming Considerations	15
	2.4. Results of the Exponential Program	16
III.	THE NORMAL CASE	29
	3.0. Introduction	29
	3.1. Determination of the Point or Points of Intersection	29
	3.2. The Index of Non-Congruity: Normal Case	31
	3.3. Description of Normal Program Features	34
	3.3.1. Program Listing	34

CHAPTER	PAGE
3.3.2. Definition of Program Variables	34
3.3.3. Inputs to the Program	37
3.3.4. Outputs of the Program	37
3.3.5. Special Programming Considerations	37
3.4. Results of the Normal Program	40
IV. CONCLUSION	68
SELECTED BIBLIOGRAPHY	70
APPENDIX 1	71
APPENDIX 2	80

LIST OF FIGURES

Figure		Page
1-1.	Illustration of the Index of Non-Congruity	4
1-2.	Distribution divided into Equi-probability Regions	6
2-1.	The Index of Non-Congruity for Exponential Distributions	11
2-2.	δ vs. λ_2/λ_1	19
2-3.	χ^2_{PS} vs. λ_2/λ_1	20
2-4.	χ^2_{PS} vs. δ	21
2-5.	χ^2_R vs. χ^2_{PS}	22
2-6.	$\hat{\alpha}$ vs. λ_2/λ_1	23
2-7.	$\hat{\alpha}$ vs. δ	24
2-8.	Possible Model of the Relationship of $\hat{\alpha}$ and δ	26
3-1.	The Normal Case with Two Points of Intersection	32
3-2.	The Normal Case for Equal Standard Deviations	33
3-3.	Tree Diagram for Sorting Procedure	39
3-4.	χ^2_{PS} vs. δ - Single Variation	42
3-5.	χ^2_{PS} vs. δ - Dual Variation	43
3-6.	δ vs. n_1 - Single Variation	44
3-7.	δ vs. n_1 - Dual Variation	45
3-8.	χ^2_{PS} vs. n_1 - Single Variation	46
3-9.	χ^2_{PS} vs. n_1 - Dual Variation	47
3-10.	χ^2_R vs. χ^2_{PS} - Mean-Variate Data	48
3-11.	χ^2_R vs. χ^2_{PS} - Variance-Variate Data $\sigma_1<\sigma_2$	49
3-12.	χ^2_R vs. χ^2_{PS} - Variance-Variate Data $\sigma_1>\sigma_2$	50
3-13.	χ^2_R vs. χ^2_{PS} - Dual Variation	51
3-14.	$\hat{\alpha}$ vs. δ - Mean-Variate Data	52
3-15.	$\hat{\alpha}$ vs. δ - Variance-Variate Data $\sigma_1<\sigma_2$	53

Figure		Page
3-16.	$\hat{\alpha}$ vs. δ - Variance-Variate Data $\sigma_1>\sigma_2$	54
3-17.	$\hat{\alpha}$ vs. δ - Dual Variation	55
3-18.	$\hat{\alpha}$ vs. η_1 - Mean-Variate Data	56
3-19.	$\hat{\alpha}$ vs. η_1 - Variance-Variate Data $\sigma_1<\sigma_2$	57
3-20.	$\hat{\alpha}$ vs. η_1 - Variance-Variate Data $\sigma_1>\sigma_2$	58

LIST OF TABLES

TABLE		PAGE
2-1.	Results of the Exponential Program	17
3-1.	Results of the Normal Program	62

Chapter 1

Introduction

1.0 GENESIS

This study had its origin in an earlier study for the United States Nuclear Regulatory Commission concerned with diesel engine failure data obtained from several nuclear power plants throughout the United States (3). A result of this earlier study was a number of questions concerning differences between statistical distributions with the primary question being "How does one determine statistically significant differences between statistical distributions?". A subsequent study by K. Lakshminarayan (4) applied the methodology used in this study to Beta distributions.

1.1 THE PROBLEM

Statistical significance is usually concerned with comparing different sets of sample data or with making inferences about the populations being sampled. Various parametric and non-parametric techniques therefore exist for making these decisions. However, no such techniques exist for comparing the populations themselves. The problem to be investigated is: given a family of statistical distributions, how much may a pair of distributions from this family differ before they can be detected as being significantly different, or "How different is different?".

1.2 PURPOSE AND OBJECTIVE

There are primarily two reasons for studying differences between similar distributions of the same family. The first reason deals with the theoretical insights which can result from studying the effects that perturbations of distribution parameters have on the "sameness" of family members.. With better understanding of the role of distribution parameters and their relative importance in determining the characteristics of a particular distribution, hopefully more powerful estimating and comparative statistical techniques can be developed. The second and probably more important reason is the practical applications which could result from studying differences between statistical distributions. Applications could include new methods for establishing when sample data from similar sources could be pooled, parametric "goodness of fit" tests, and sample-free hypothesis testing.

With these two broad underlying reasons for studying differences in statistical distributions from the same family, the expressed objective of this study is: to develop a method of comparing differences in statistical distributions from the same family (normal distributions and exponential distributions are the families of statistical distributions studied), to use this technique to examine the effects of varying the parameters of the distributions on their "sameness", and to attempt to draw some conclusions pertaining to the usefulness of this technique in answering the question of "How different is different?".

1.3 METHOD

1.3.1. The Index of Non-Congruity - δ

The following discussion is an adaptation of material presented by Lakshminarayan (3).

Theoretically, two continuous distributions are the same only if their probability density functions are identical and for their probability density functions to be identical the two distributions must have exactly the same parameters. In practical situations however, two distributions whose probability density functions (and therefore parameters) are nearly the same, may produce random samples which are indistinguishable from one another. It is this type of situation which indicates that merely examining the probability density functions (or the parameters) of two distributions to see if they are identical does not provide enough information to judge if the two distributions are similar enough to consider them practically as being the same, or if they are different enough that they must be considered as different.

One measure that determines differences between continuous distributions is the difference in the areas bounded by each probability density function in the region of interest or the amount of non-overlapping area bounded by the curves. If $f_1(x)$ and $f_2(x)$ are the probability density functions (Figure 1-1) of the two distributions being compared, then the amount of non-overlapping area or what we have termed "the index of non-congruity" is given by:

$$\delta = \int_{-\infty}^{\infty} |f_1(x) - f_2(x)| dx \qquad (1)$$

The non-overlapping area is shown by the shaded portion of Figure 1-1 and the total amount of this shaded area equals δ. To qualify as probability density functions, $f_1(x)$ and $f_2(x)$ each must satisfy the criterion:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \qquad (2)$$

Therefore the values that can be assumed by the index of non-congruity are $0 \leq \delta \leq 2$. If two distributions are approximately the same then δ will be close to zero and if two distributions are radically different the value of δ will approach 2.

1.3.2. The Procedure

The general procedure used in this study is to choose a particular

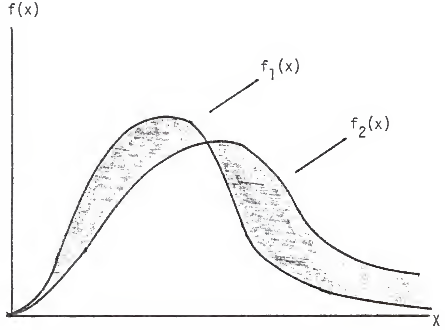

FIGURE 1-1 Illustration of the Index of Non-Congruity

distribution from a family of distributions and then compare it to a similar distribution from the same family. The first distribution will be termed the "model distribution" (Distribution 1) and the similar distribution will be termed the "alternative distribution" (Distribution 2).

The procedure by which the alternative distribution is compared with the model distribution consists of a number of steps. The first step is to calculate the index of non-congruity between the two distributions as explained in Section 1.3.1.

Secondly, the model distribution is divided into ten equi-probability regions. A set of values $\{x(i)\}$ of the independent variable is calculated such that:

$$\int_{-\infty}^{x(i)} f_1(x)dx = i/10 \qquad i = 1,\ldots,10 \tag{3}$$

The values $\{x(i)\}$ are such that the sample space of the independent variable is divided into regions which have the same area under the curve of the probability density function, as shown in Figure 1-2. After the equi-probability regions for the model distribution have been determined, a "perfect" sample is drawn from the alternative distribution by using the $\{x(i)\}$ from the model distribution as interval boundaries of the alternative distribution. A "pseudo" - χ^2 statistic is then calculated from this "perfect" sample. This pseudo-χ^2 statistic, χ^2_{PS}, is :

$$\chi^2_{PS} = \sum_{i=1}^{10} \frac{\{[F_2(i)-F_2(i-1)]M - .1(M)\}^2}{.1(M)} \tag{4}$$

where

$$F_2(i) = \int_{-\infty}^{x(i)} f_2(x)\, dx, \tag{5}$$

M is the sample size and $F_2(0) = 0$. We are interested in small sample sizes because small sample sizes are usually encountered in practical

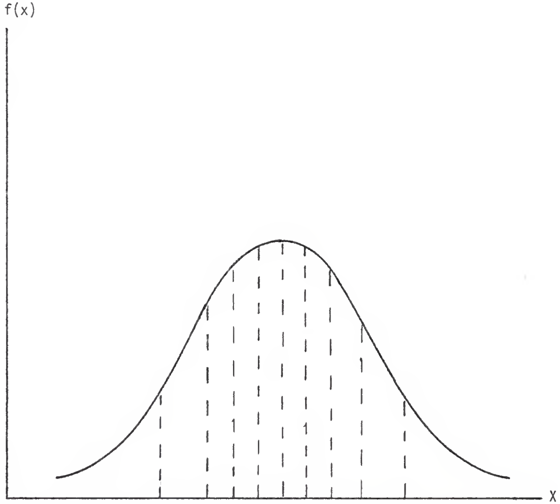

FIGURE 1-2 Distribution divided into Equi-probability Regions

situations, and with very large sample sizes even small differences are discernable. Therefore a value of 50 is used for M in this study. The "pseudo" χ^2 and "perfect" sample are used to reduce the effects introduced by random fluctuations and to better ascertain the basic relationship between the index of non-congruity and differences between the model and alternative distributions.

The third step in the procedure comparing the alternative distribution to the model distribution is that a genuine random sample of size M is drawn from the alternative distribution and the sample is compared to the model distribution. The comparison is made by an ordinary χ^2 goodness of fit test. The random χ^2 statistic is calculated to provide a check on the χ^2_{PS} results and to provide additional insight into the question of differences between statistical distributions. Once the random χ^2, denoted by χ^2_R , has been determined, the level of significance $\hat{\alpha}$ is calculated. Usually when a χ^2 statistic is calculated it is then compared to a value obtained from an appropriately chosen χ^2 distribution to determine significance at a prescribed confidence level. In this situation this technique is not totally satisfactory since we are not only interested in whether a particular χ^2_R value is significant but also in how significant it is. The level of significance $\hat{\alpha}$ is the area to the right of the computed (observed) χ^2_R of a χ^2 distribution with 9 degrees of freedom. There are 9 degrees of freedom since the model distribution is divided into 10 equi-probability regions and the random observations are sorted into these regions for ease of computation. The level of significance gives a more intuitively comprehensible measure of difference than the χ^2_{PS} and χ^2_R values.

The final step in comparing the alternative distribution with the model distribution is the calculation of parametric indicators which attempt to quantify the differences between the model distribution and the alternative distribution. It is hoped that a relationship can be discovered between a parametric indicator and the index of non-congruity. Using this relationship combined with knowledge about the relationship between the index of non-congruity and statistical significance it may be possible to find a measure of statistical difference between distributions based solely on the parameters of the distributions. Such a parametric indicator would be of considerable practical importance because it would be easy to calculate.

In summary, the comparison procedure given a model distribution and an alternative distribution is:

1) Calculate δ, the index of non-congruity
2) Calculate χ^2_{PS}, the "pseudo" chi-square statistic from a "perfect" sample
3) Calculate χ^2_R from a random sample from the alternative distribution
4) Calculate $\hat{\alpha}$, the level of significance for χ^2_R
5) Calculate various parametric indicators

1.3.3 The McGill Random Number Generator

This study is primarily based on the use of a computer to perform the comparison procedure outlined in Section 1.3.2. One of the major problems in the development of a program to perform this procedure is the generation of a random sample from the alternative distribution to be used in calculating χ^2_R. The McGill Random Number Generator developed by members of the School of Computer Science of McGill University seemed particularly well suited for the requirements of this study. The McGill RNG has several features which led to its selection. First, the use of the McGill RNG is FORTRAN compatible and the rest of the program will be written in FORTRAN. Second, the McGill RNG is called as a FORTRAN function rather than as a subroutine, which is advantageous in terms of computation time. Third, the previous value returned is maintained internally by the McGill RNG which leads to easier programming. Fourth, the McGill RNG has special procedures for generating random samples from normal and exponential distributions, thereby eliminating the need to program a transformation for converting a uniform distribution to either of these distributions. Finally, the McGill RNG is included in the subroutine library of the Kansas State University computing system, eliminating the need to include an additional subprogram for the random number generator in the index of non-congruity program.

Chapter 2

THE EXPONENTIAL CASE

2.0 INTRODUCTION

This chapter describes the specific procedure for calculating the index of non-congruity for exponential distributions, the program developed to perform the comparison procedure outlined in Section 1.3.2. for exponential distributions, and the results of using this procedure to compare several pairs of exponential distributions.

2.1 DETERMINATION OF THE POINT OF INTERSECTION

Let $f_1(x)$ be the probability density function of an exponential distribution with parameter λ_1 and let $f_2(x)$ be the probability density function of an exponential distribution with parameter λ_2. Assume that $\lambda_2 > \lambda_1$. Consider Figure 2-1 which shows two exponential distributions fulfilling these requirements. The shaded area represents the index of non-congruity for this pair of distributions. This area difference is given by Equation (1) which is repeated again for clarity.

$$\delta = \int_{-\infty}^{\infty} |f_1(x) - f_2(x)|\, dx \qquad (1)$$

To facilitate calculation of δ, this integral can be divided into 2 components.

$$\delta = \int_{0}^{\infty} |f_1(x)-f_2(x)|dx = \int_{0}^{X_A} [f_2(x)-f_1(x)]dx + \int_{X_A}^{\infty} [f_1(x)-f_2(x)]dx \qquad (6)$$

X_A is the point of intersection of the two probability density functions. Finally note that this expression applies for the case where $\lambda_2 > \lambda_1$ and for the case where $\lambda_1 > \lambda_2$ the limits of integration for the component integrals would have to be exchanged to insure the proper sign for δ.

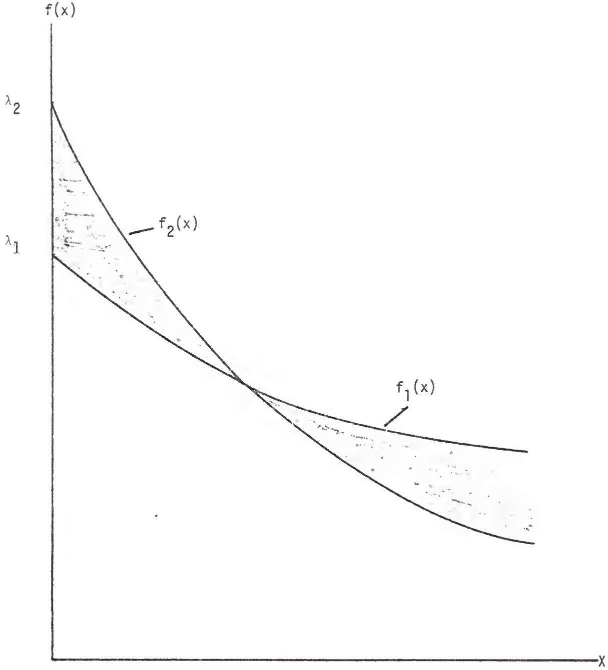

FIGURE 2-1 The Index of Non-Congruity for Exponential Distributions

The following calculation demonstrates the determination of the point of intersection X_A. At $X = X_A$

$$\lambda_1 e^{-\lambda_1 X_A} = \lambda_2 e^{-\lambda_2 X_A} \tag{7}$$

$$e^{(\lambda_2-\lambda_1)X_A} = \frac{\lambda_2}{\lambda_1} \tag{8}$$

$$(\lambda_2-\lambda_1)X_A = \ln \frac{\lambda_2}{\lambda_1} \tag{9}$$

$$X_A = [\frac{1}{\lambda_2-\lambda_1}] \ln \frac{\lambda_2}{\lambda_1} \tag{10}$$

2.2 THE INDEX OF NON-CONGRUITY: EXPONENTIAL CASE

Having determined the point of intersection X_A the expression for the index of non-congruity can be modified. Substituting into Equation (6) we obtain:

$$\delta = \int_0^{X_A} [\lambda_2 e^{-\lambda_2 X} - \lambda_1 e^{-\lambda_1 X}]dx + \int_{X_A}^{\infty} [\lambda_1 e^{-\lambda_1 X} - \lambda_2 e^{-\lambda_2 X}]dx \tag{11}$$

$$= \int_0^{X_A} \lambda_2 e^{-\lambda_2 X}\, dx - \int_0^{X_A} \lambda_1 e^{-\lambda_1 X}\, dx + \int_{X_A}^{\infty} \lambda_1 e^{-\lambda_1 X}\, dx - \int_{X_A}^{\infty} \lambda_2 e^{-\lambda_2 X}\, dx \tag{12}$$

$$= -e^{-\lambda_2 X}\Big|_0^{X_A} - -e^{-\lambda_1 X}\Big|_0^{X_A} + -e^{-\lambda_1 X}\Big|_{X_A}^{\infty} - -e^{-\lambda_1 X}\Big|_{X_A}^{\infty} \tag{13}$$

$$= [-e^{-\lambda_2 X_A} + 1] - [-e^{-\lambda_1 X_A} + 1] + [0 + e^{-\lambda_1 X_A}] - [0 + e^{-\lambda_2 X_A}] \tag{14}$$

$$= 2e^{-\lambda_1 X_A} - 2e^{-\lambda_2 X_A} \tag{15}$$

Therefore

$$\delta = 2[e^{-\lambda_1 X_A} - e^{-\lambda_2 X_A}] \tag{16}$$

This development is for the case where $\lambda_2 > \lambda_1$. For the case where

$\lambda_1 > \lambda_2$ the index of non-congruity is:

$$\delta = 2[e^{-\lambda_2 X_A} - e^{-\lambda_1 X_A}] \qquad (15a)$$

and:

$$X_A = \frac{1}{\lambda_1 - \lambda_2} \ln \frac{\lambda_1}{\lambda_2} \qquad (10a)$$

2.3 DESCRIPTION OF EXPONENTIAL PROGRAM FEATURES

2.3.1 Program Listing

A complete listing of the computer program to perform the comparison procedure for exponential distributions is given in Appendix 1.

2.3.2. Definition of Program Variables

AHAT: the level of significance, $\hat{\alpha}$

CADTR: function subroutine to determine $\hat{\alpha}$

DIFFL: absolute difference of λ_1 and λ_2, $|\lambda_1 - \lambda_2|$

DELTA: the index of non-congruity, δ

EI: the expected frequency in the equi-probability regions, M/10

FREQ(10): the array containing the frequency counts of the random sample sorted into the equi-probability regions

F2SUM: the sum of the components of the array FREQ squared

ISEED: one of the seeds for the McGill RNG

JSEED: the other seed for the McGill RNG

K: the index for the array FREQ. K can take on integer values from 1 to 10.

LAMDA1: the parameter of the model exponential distribution

LAMDA2: the parameter of the alternative exponential distribution

M: the sample size

NU: the degrees of freedom for the χ^2 distribution which χ^2_{PS} and χ^2_R are compared with

P2(10): the array containing the cumulative probability of the alternative distribution at the region boundries {x(i)}

RATIO: the ratio of λ_1 and λ_2, if $\lambda_2 > \lambda_1$

$$\text{RATIO} = \frac{\lambda_2}{\lambda_1} \text{ and if } \lambda_1 > \lambda_2 \quad \text{RATIO} = \frac{\lambda_1}{\lambda_2}$$

REXP: subroutine to generate a random deviate from an exponential distribution with mean $\lambda = 1$

RSTART: subroutine to initialize the McGill Random Number Generator

SAMPL(200): the array containing the random sample from the alternative distribution

T1: the point of intersection X_A

X1(10): the equi-probability region boundaries {x(i)}

X2ACT: the random χ^2 statistic, χ^2_R

X2PS: the pseudo χ^2 statistic, χ^2_{PS}

X2SUM: the sum of $[P2(10) - X1(10)]^2$ used in the calculation of X2PS

Z1: variable equal to the negative of the product of LAMDA1 and T1

Z2: variable equal to the negative of the product of LAMDA2 and T1

2.3.3. Inputs to the Program

The variables required as input to the program are:

LAMDA1, LAMDA2, M, ISEED, JSEED

The input is given on two separate cards with the indicated format.

LAMDA1, LAMDA2, M	1 card	(2E10.4, I5)
ISEED, JSEED	1 card	(2I5)

Multiple runs of the program can be made by supplying additional input cards (two per replication) containing the information described above. Program completion is indicated by a blank card.

2.3.4. Outputs of the Program

The program provides two types of output. The first type is an echo check of the input. The second type is information calculated by the program. The following information is produced as output of the second type: the array X1, the array P2, the first M components of the array SAMPL, the array FREQ, DELTA, X2PS, X2ACT, and AHAT. A sample output is shown in Appendix 1.

2.3.5. Special Programming Considerations

Using the McGill Random Number Generator The use of the McGill RNG is accomplished by the two subroutines RSTART and REXP. RSTART initializes the RNG. The arguments of RSTART are ISEED and JSEED. The transfer from the main program to the RSTART subroutine is made by the statement Call RSTART (ISEED, JSEED). The RNG provides default values if RSTART is not used. REXP generates a random exponential deviate from an exponenetial population with parameter $\lambda = 1$. This random deviate is transformed into a random deviate from an exponential population with parameter λ_A by dividing by λ_A e.g. $z = x/\lambda_A$ where z is the random deviate from the desired distribution and x is the generated random deviate. The argument of REXP is a dummy integer constant which is ignored by the program. In other words use of REXP(10) or REXP(98765) produces the same effect i.e. the generation of an exponential random deviate. Use of the function subroutine REXP is accomplished by using REXP(1) in an arithmetic function e.g. SAMPL(I) = REXP(1)/LAMDA2.

Sorting the Random Sample Observations In designing a sorting procedure the objective is to minimize the expected number of sorting trials for a sample set while maintaining a level of simplicity in the programming. The sorting procedure used in the program consists of a loop containing a set of test statements which compares the random deviate with the equal probability region boundaries sequentially until the deviate is less than the boundary value. The deviate is then placed in the frequency region which has the boundary value as its upper bound. This sorting procedure would minimize the expected number of trials for the case where $\lambda_2 > \lambda_1$. However for the case where $\lambda_1 >> \lambda_2$ this procedure would result in a high expected number of trials. This was not considered a significant problem since we are primarily concerned with cases where λ_1 and λ_2 are nearly equal.

Calculating $\hat{\alpha}$ The level of significance, $\hat{\alpha}$, is calculated by the function subroutine CADTR which is a slightly modified version of the CDTR subroutine contained in IBM's Scientific Subroutine Package. The modifications include changing the subprogram from a subroutine subprogram to a function subprogram and modifying the inputs and outputs of the subprogram.

2.4. RESULTS OF THE EXPONENTIAL PROGRAM

Four different sets of values of random number generator seeds were used in investigating the exponential case. Values of distributions compared varied from $\lambda_2/\lambda_1 = 1/3$ to $\lambda_2/\lambda_1 = 3$. For $\lambda_1 > \lambda_2$ the value of λ_2 was set to equal 10 and for $\lambda_2 > \lambda_1$ the value of λ_1 was set equal to 10. Therefore in every pair of distributions compared the smallest parameter was equal to 10. This was done for computation convenience since only the ratio is pertinent (as seen from Equations (10) and (16)), rather than the absolute size of λ_1 and λ_2. The sample size used in the comparison was set equal to 50 in all cases. The results of the various comparison runs are summarized in Table 2-1.

Various relationships between comparison indices are graphically presented in Figures 2-2 to 2-7. Examination of these figures indicates that there is good reason to believe that there is a strong relationship

TABLE 2-1

Results of the Exponential Program

λ_1	λ_2	RATIO	RNG Seed	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
30	10	1/3	A	.7698	76.80	86.00	.0000
			B			84.40	.0000
			C			120.40	.0000
			D			75.20	.0000
20	10	1/2	A	.5	28.67	24.40	.0037
			B			39.20	.0000
			C			48.80	.0000
			D			41.60	.0000
17.5	10	4/7	A	.4064	17.93	20.00	.0179
			B			46.80	.0000
			C			37.20	.0000
			D			36.00	.0000
15	10	2/3	A	.2963	8.88	9.20	.4190
			B			28.80	.0007
			C			21.60	.0102
			D			18.40	.0508
12.5	10	4/5	A	.1638	2.48	4.80	.8514
			B			14.40	.1088
			C			10.40	.3191
			D			14.00	.1223
10	12.5	5/4	A	.1638	2.00	10.40	.3191
			B			12.40	.1917
			C			10.00	.3505
			D			9.20	.4190
10	15	3/2	A	.2963	6.13	10.40	.3191
			B			13.60	.1373
			C			10.80	.2897
			D			12.80	.1719
10	17.5	7/4	A	.4064	11.12	17.60	.0401
			B			20.00	.0179
			C			18.80	.0269
			D			18.80	.0269
10	20	2/1	A	.5	16.50	19.60	.0205
			B			23.20	.0058
			C			18.80	.0269
			D			26.40	.0018

Table 2-1 continued

λ_1	λ_2	RATIO	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
10	30	3/1	A	.7698	39.50	48.40	.0000
			B			40.40	.0000
			C			43.60	.0000
			D			47.60	.0000

M = 50, RNG = A(51562, 62155) B(62155, 51562)
C(50020, 11292) D(11292, 50020)

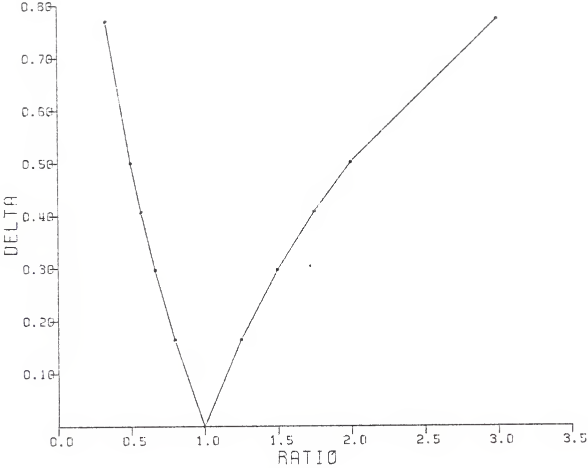

FIGURE 2-2 δ vs. λ_2/λ_1

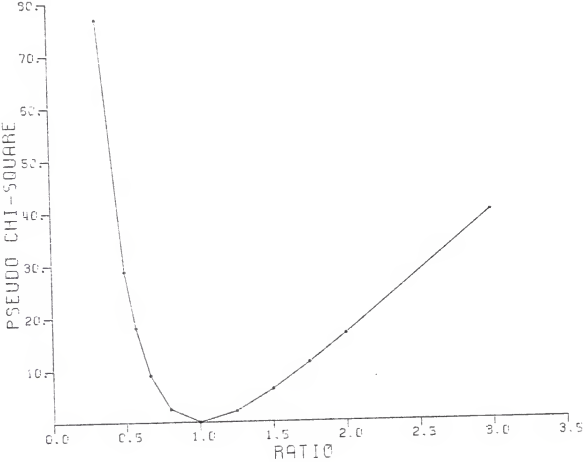

FIGURE 2-3 χ^2_{PS} vs. λ_2/λ_1

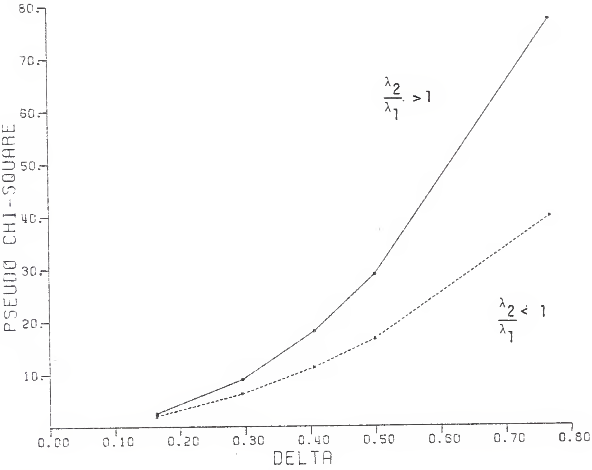

FIGURE 2-4 χ^2_{PS} vs. δ

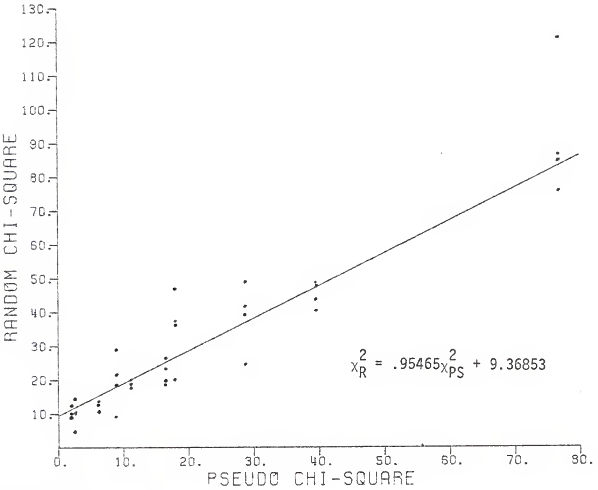

FIGURE 2-5 χ_R^2 vs. χ_{PS}^2

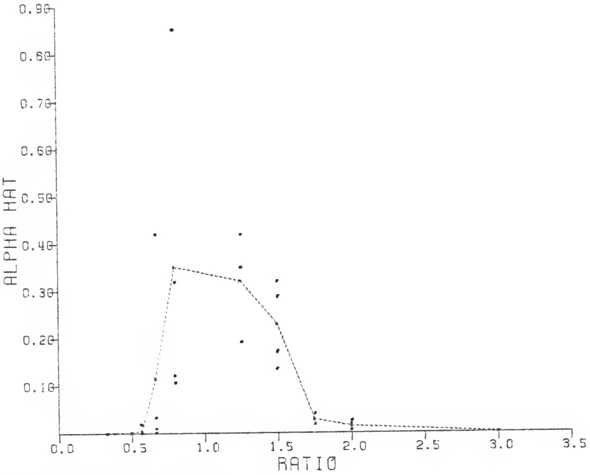

FIGURE 2-6 $\hat{\alpha}$ vs. λ_2/λ_1

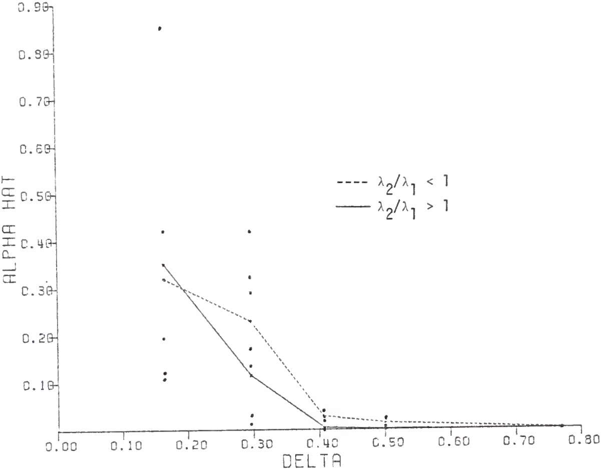

FIGURE 2-7 $\hat{\alpha}$ vs. δ

between all of the comparison indices plotted. It is only possible at present to make very tentative estimates of most of these relationships. The comparison variable pairs where a clear relationship exists are δ vs λ_2/λ_1, χ^2_{PS} vs δ, and χ^2_{PS} vs λ_2/λ_1. These relationships do not vary with different random samples. However, for the relationships involving variables which are affected by different random samples-namely χ^2_R and $\hat{\alpha}$ - it is only possible to make tentative estimates of the relationships. To obtain better insight into these relationships, it would be necessary to increase the number of replications (using different random number seed values each time). Increasing the number of replications would increase the knowledge of the distribution of the values assumed by χ^2_R and $\hat{\alpha}$ at fixed values of the non-effected variables (δ, λ_2/λ_1, and χ^2_{PS}) and would also increase the accuracy of the estimate of the mean value of the affected variables. With sufficient replications it would be possible to obtain very reliable figures of the type shown in Figure 2-8 which illustrates a possible model of the relationship of $\hat{\alpha}$ and δ. The knowledge of these tentative relationships is believed to be sufficient for the purposes of this study but it is recognized that further research should proceed in the effort to better quantify these relationships.

Examination of Figures 2-2 to 2-7 provides some valuable insights and also produces some interesting observations. First, as shown in Figure 2-2, δ changes at a faster rate for a given change in the lambda ratio for $\lambda_2/\lambda_1 < 1$ than for $\lambda_2/\lambda_1 > 1$. This indicates that if an alternative distribution is compared to a model distribution having a smaller parameter, of size $\lambda_2 - \varepsilon$ say, it is more likely that the two distributions can be considered as equivalent (because of a smaller δ value) than if the alternative distribution was compared to a model distribution having a correspondingly larger parameter of $\lambda_2 + \varepsilon$.

Second, as indicated in Figures 2-3 and 2-4, χ^2_{PS} (and therefore

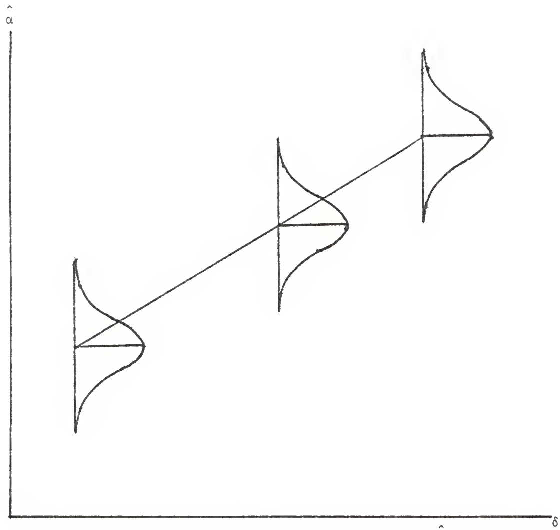

FIGURE 2-8 Possible model of the Relationship of $\hat{\alpha}$ and δ

presumably χ_R^2 and $\hat{\alpha}$) is more sensitive to differences in the distributions being compared for $\lambda_2/\lambda_1 < 1$ than for $\lambda_2/\lambda_1 > 1$. This sensitivity is in addition to the effect on δ of the relative magnitude difference of the parameters of the distributions. This indicates that in addition to the added likelihood that an alternative distribution and a model distribution can be considered equivalent for $\lambda_2/\lambda_1 > 1$ due to the magnitude effect, there is also an additional effect caused by the reduced sensitivity to difference for $\lambda_2/\lambda_1 > 1$.

Figure 2-5 indicates that χ_{PS}^2 in general tends to be smaller than χ_R^2. It also appears that there is a linear trend between χ_{PS}^2 and χ_R^2 if the point $\chi_{PS}^2 = 76.8$, $\chi_R^2 = 120.40$ is not considered. A least-squares line was calculated for this relationship. The equation of this line is $\chi_R^2 = .95465\ \chi_{PS}^2 + 9.36853$.

Figure 2-5 and Table 2-1 also produce two interesting observations. First, there is a tighter grouping of χ_R^2 from cases where $\lambda_2/\lambda_1 > 1$. Second, RNG seed A seems to produce peculiar results for cases where $\lambda_2/\lambda_1 < 1$. The underlying causes of these two phenomena are not fully understood.

From Figure 2-7 it appears that δ tends to be significant at the .05 level above values of .4. At values between $.3 \leq \delta \leq .4$ the result is ambiguous if the judgment is to be based on $\hat{\alpha}$ (χ_R^2). If χ_{PS}^2 is used to judge the two distributions for equivalence the range of uncertainty for δ can be determined from Table 2-1. Significance occurs at χ^2 values around 17 at the .05 level. Therefore χ_{PS}^2 would indicate significance at the .05 level above 8 since .95465(8) + 9.36853 = 17. It appears that if χ_{PS}^2 is used as the judgment criterion, then for δ to indicate significance its value must be greater than .3 for $\lambda_2/\lambda_1 < 1$ and greater than .35 for $\lambda_2/\lambda_1 > 1$. Using χ_{PS}^2 as the judgment criterion the uncertainty range for

δ appears to be

$$.2 \leq \delta \leq .3 \qquad \lambda_2/\lambda_1 < 1$$

$$.25 \leq \delta \leq .35 \qquad \lambda_2/\lambda_1 > 1$$

These δ values imply that if λ_2/λ_1 is to be used to test for significance as opposed to δ then the uncertainty regions are

$$\left\{ \begin{array}{ll} .57 \leq \lambda_2/\lambda_1 \leq .67 & \lambda_2/\lambda_1 < 1 \\ 1.5 \leq \lambda_2/\lambda_1 \leq 1.75 & \lambda_2/\lambda_1 > 1 \end{array} \right\} \hat{\alpha} \text{ basis}$$

$$\left\{ \begin{array}{ll} .65 \leq \lambda_2/\lambda_1 \leq .75 & \lambda_2/\lambda_1 < 1 \\ 1.45 \leq \lambda_2/\lambda_1 \leq 1.6 & \lambda_2/\lambda_1 > 1 \end{array} \right\} \chi^2_{PS} \text{ basis}$$

and significance is indicated for values of

$$\lambda_2/\lambda_1 \leq .57 \quad \text{or} \quad \lambda_2/\lambda_1 \geq 1.75 \quad \hat{\alpha} \text{ basis}$$

$$\lambda_2/\lambda_1 \leq .65 \quad \text{or} \quad \lambda_2/\lambda_1 \geq 1.6 \quad \chi^2_{PS} \text{ basis}$$

Chapter 3

THE NORMAL CASE

3.0 INTRODUCTION

This chapter describes the specific procedure for calculating the index of non-congruity for normal distributions, the program developed to perform the comparison procedure outlined in Section 1.3.2. for normal distributions, and the results of using the procedure to compare several pairs of normal distributions.

3.1 DETERMINATION OF THE POINT OR POINTS OF INTERSECTION

Let $f_1(x)$ be the probability density function of a normal distribution with mean u_1 and standard deviation σ_1. Let $f_2(x)$ be the probability density function of a normal distribution with mean u_2 and standard deviation σ_2. At a point of intersection we have,

$$\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-1/2(\frac{x-u_1}{\sigma_1})^2} - \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-1/2(\frac{x-u_2}{\sigma_2})^2} = 0 , \qquad (17)$$

so that $$e^{-1/2[(\frac{x-u_1}{\sigma_1})^2 - (\frac{x-u_2}{\sigma_2})^2]} = \frac{\sigma_1}{\sigma_2} \qquad (18)$$

and $$[(\frac{x-u_1}{\sigma_1})^2 - (\frac{x-u_2}{\sigma_2})^2] = -2\ln \frac{\sigma_1}{\sigma_2} . \qquad (19)$$

Then $$(\sigma_2^2 - \sigma_1^2)x^2 - 2(\sigma_2^2 u_1 - \sigma_1^2 u_2) + [\sigma_2^2 u_1^2 - \sigma_1^2 u_2^2 + 2\sigma_1^2\sigma_2^2 \ln \frac{\sigma_1}{\sigma_2}] = 0 \qquad (20)$$

Equation (20) is of the form

$$A x^2 + Bx + C = 0 \qquad (21)$$

The number of roots of equations of this form can be determined from the discriminant. If $(B^2 - 4AC)$ is negative there are no real roots

zero one real root $X_A = \frac{-B}{2A}$

positive two real roots, $X_{A1} = \frac{-B + \sqrt{B^2-4AC}}{2A}$

$$X_{A2} = \frac{-B - \sqrt{B^2-4AC}}{2A}$$

Denoting the discriminant by R, we find that

$$R = 4(\sigma_2{}^2u_1 - \sigma_1{}^2u_2)^2 - 4(\sigma_2{}^2 - \sigma_1{}^2)[\sigma_2{}^2u_1{}^2 - \sigma_1{}^2u_2{}^2 + 2\sigma_1{}^2\sigma_2{}^2 \ln \frac{\sigma_1}{\sigma_2}] \qquad (22)$$

which reduce to the conditions

R negative yields no real roots

R zero yields one root $X_A = \frac{\sigma_2{}^2u_1 - \sigma_1{}^2u_2}{\sigma_2{}^2 - \sigma_1{}^2}$

R positive yields two roots $X_{A1} = X_A + \frac{\sqrt{R}}{2(\sigma_2{}^2 - \sigma_1{}^2)}$

$$X_{A2} = X_A - \frac{\sqrt{R}}{2(\sigma_2{}^2 - \sigma_1{}^2)}$$

Now in the special case where $\sigma_2 = \sigma_1 = \sigma$, then Equation (17) reduces to

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-1/2(\frac{x-u_1}{\sigma})^2} - \frac{1}{\sqrt{2\pi}\,\sigma} e^{-1/2(\frac{x-u_2}{\sigma})^2} = 0 \qquad (17a)$$

which yields $(x-u_1)^2 = (x-u_2)^2$ (24)

$$x^2 - 2xu_1 + u_1{}^2 = x^2 - 2xu_2 + u_2{}^2 \qquad (25)$$

$$2xu_2 - 2xu_1 = u_2{}^2 - u_1{}^2 \qquad (26)$$

$$X = \frac{u_2{}^2 - u_1{}^2}{2(u_2 - u_1)} \qquad (27)$$

$$X_A = \frac{u_2 + u_1}{2} \qquad (28)$$

3.2 THE INDEX OF NON-CONGRUITY: NORMAL CASE

The two most common situations with normal distributions, where there are two points of intersection and $\sigma_1 = \sigma_2$, are shown in Figures 3-1 and 3-2. The shaded area in these figures equals the index of non-congruity.

The index of non-congruity is given by Equation (6).

$$\delta = \int_{-\infty}^{\infty} |f_1(x) - f_2(x)| \, dx \tag{6}$$

This expression can be modified according to the number of intersection points. If there is no point of intersection then obviously $\delta = 2$. If there is one point of intersection, X_A, the integral can be decomposed into two terms. Assume that $f_1(x) > f_2(x)$ for $X < X_A$ then

$$\delta = \int_{-\infty}^{X_A} (f_1(x) - f_2(x))\, dx + \int_{X_A}^{\infty} (f_2(x) - f_1(x))\, dx \tag{29}$$

$$\delta = \int_{-\infty}^{X_A} f_1(x)\, dx - \int_{-\infty}^{X_A} f_2(x)\, dx + \int_{X_A}^{\infty} f_2(x)\, dx - \int_{X_A}^{\infty} f_1(x)\, dx \tag{30}$$

$$\delta = F_1(X_A) - F_2(X_A) + [1 - F_2(X_A)] - [1 - F_1(X_A)] \tag{31}$$

$$\delta = 2[F_1(X_A) - F_2(X_A)] \tag{32}$$

where $F_i(x)$ is the cumulative probability function of $f_i(x)$. In general, for the case where there is one point of intersection

$$\delta = 2|F_1(X_A) - F_2(X_A)| \tag{33}$$

If there are two points of intersection (see Figure 3-2) then the index of non-congruity integral can be separated into three parts. Let C_1 and C_2 $(C_1 < C_2)$ be the points of intersection.

Assume that $f_1(x) > f_2(x)$ for $X < C_1$ then

$$\delta = \int_{-\infty}^{C_1} (f_1(x)-f_2(x))\, dx + \int_{C_1}^{C_2} (f_2(x) - f_1(x))\, dx + \int_{C_2}^{\infty} (f_1(x)-f_2(x))\, dx \tag{34}$$

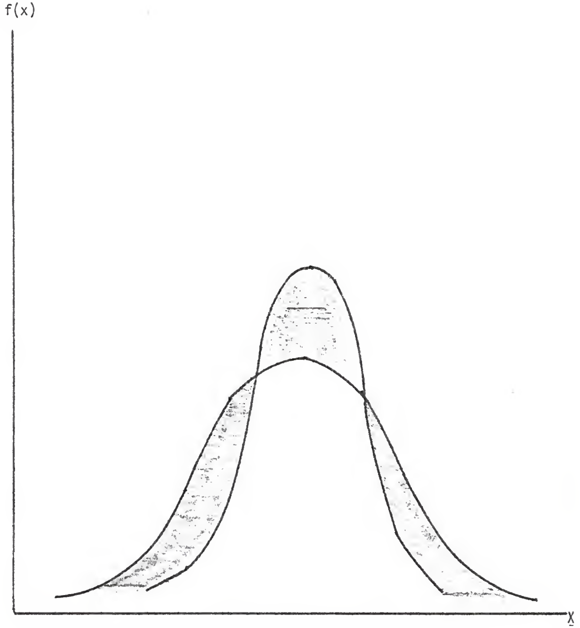

FIGURE 3-1 The Normal Case with Two Points of Intersection

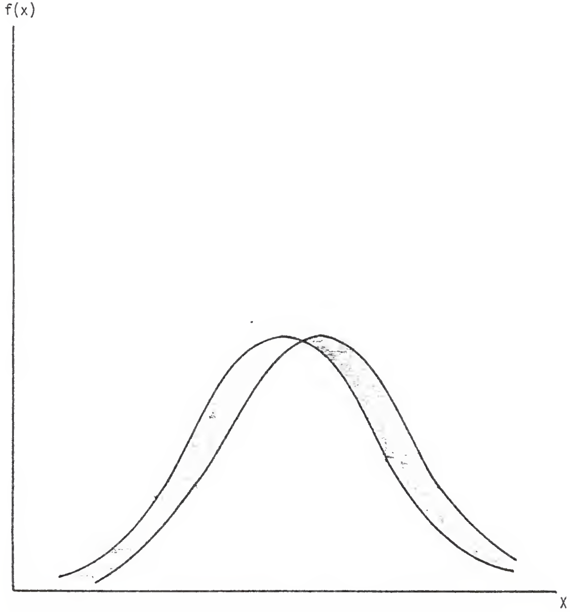

FIGURE 3-2 The Normal Case for Equal Standard Deviations

$$\delta = F_1(C_1) - F_2(C_1) + [F_2(C_2) - F_2(C_1) - F_1(C_2) + F_1(C_1)] + [1 - F_1(C_2) - 1 + F_2(C_2)] \qquad (35)$$

$$\delta = 2[(F_2(C_2) - F_2(C_1)) - (F_1(C_2) - F_1(C_1))] \qquad (36)$$

In general for the case where there are two points of intersection

$$\delta = |2[(F_2(C_2) - F_2(C_1)) - (F_1(C_2) - F_1(C_1))]| \qquad (37)$$

3.3 DESCRIPTION OF NORMAL PROGRAM FEATURES

3.3.1. Program Listing

This section describes the essential features of the program for comparing normal distributions. A complete listing of the program is given in Appendix 2.

3.3.2. Definition of Program Variables

AHAT: the level of significance, $\hat{\alpha}$

C1: the lower point of intersection, C_1

C2: the upper point of intersection , C_2

CADTR: function subroutine to determine $\hat{\alpha}$

D: an output parameter of subroutine NDTR which is not used in the main program

DELTA: the index of non-congruity, δ

EI: the expected frequency in the equi-probability regions, M/10

F1: the cumulative probability $F_1(X)$ at X_A

F1C1: the cumulative probabiltiy $F_1(X)$ at C_1

F1C2: the cumulative probability $F_1(X)$ at C_2

F1DIF: F1C2 - F1C1

F2: the cumulative probability $F_2(X)$ at X_A

F2C1: the cumulative probability $F_2(X)$ at C_1

F2C2: the cumulative probability $F_2(X)$ at C_2

F2DIF: F2C2 - F2C1

F2SUM: the sum of the components of the array FREQ squared

FREQ(10): The array containing the frequency counts of the random sample sorted into the equi-probability regions

FRSFAC: $4(\sigma_2{}^2u_1 - \sigma_2{}^2u_2)^2$

IER: error indicator used in subroutine NDTRI

IND: indicator showing the number of intersection points

ISEED: one of the seeds for the McGill RNG

JSEED: the other seed for the McGill RNG

K: the index for the array FREQ

MU1: population mean μ_1

MU2: population mean μ_2

M: the sample size

NDTR: subroutine to calculate $F_i(z)$

NDTRI: subroutine to calculate the standard normal deviate z, given $F_i(z)$

P: input parameter to subroutine NDTRI containing $F_i(z)$

P2(10): the array containing the cumulative probability of the alternative distribution at the region boundaries {x(i)}

RAD: R of Equation (22)

RADPRT: $\sqrt{R}/2(\sigma_2{}^2 - \sigma_1{}^2)$

RATIO: the ratio of population standard deviations, σ_1/σ_2

RNOR: McGill RNG function to generate standard normal deviate

RSTART: subroutine to initialize the McGill RNG

SAMPL(200): the array containing the random sample from the alternative distribution

SIGMA1: population standard deviation, σ_1

SIGMA2: population standard deviation, σ_2

SNDFAC: $4(\sigma_2^2 - \sigma_1^2)[\sigma_2^2 u_1^2 - \sigma_1^2 u_2^2 + 2\sigma_1^2\sigma_2^2 \ln \frac{\sigma_1}{\sigma_2}]$

SQMU1: u_1^2

SQMU2: u_2^2

VAR1: σ_1^2

VAR2: σ_2^2

VARDIF: VAR2 - VAR1

VXM12: $\sigma_1^2 u_2$

VXM21: $\sigma_2^2 u_1$

VXMD2: $(\sigma_2^2 u_1 - \sigma_1^2 u_2)^2$

VXMDIF: $(\sigma_2^2 u_1 - \sigma_1^2 u_2)$

VXS12: $\sigma_1^2 u_2^2$

VXS21: $\sigma_2^2 u_1^2$

X1(10): the equi-probability region boundaries {X(i)}

X2ACT: the random χ^2 statistic, χ_R^2

X2PS: the pseudo χ^2 statistic, χ_{PS}^2

X2SUM: the sum of [P2(10) - X1(10)] used in the calculation of X2P

XLNFAC: $2\ \sigma_2^2\sigma_1^2 \ln \frac{\sigma_1}{\sigma_2}$

XX: the single point of intersection, X_A

XX1: one of two points of intersection, X_{A1}

XX2: the other of two points of intersection, X_{A2}

Z: a standard normal deviate used in calculating X1(10)

Z1: standard normal equivalent of X_A for distribution 1

Z1C1: standard normal equivalent of C_1 for distribution 1

Z1C2: standard normal equivalent of C_2 for distribution 1

Z2: standard normal equivalent of X_A for distribution 2

Z2C1: standard normal equivalent of C_1 for distribution 2

Z2C2: standard normal equivalent of C_2 for distribution 2

Z2PS(10): array containing standard normal equivalent of X1(10) for distribution 2

3.3.3. Inputs to the Program

The inputs to the program are the distribution parameters - u_1, σ_1, u_2, σ_2; the sample size M; and the RNG seed values - ISEED, JSEED. The input is given on two separate cards with the indicated format.

MU1, SIGMA1, MU2, SIGMA2, M	1 Card	(4E10.4, I5)
ISEED, JSEED	1 Card	(2I5)

Multiple runs are possible by supplying additional input cards (two per replication). Program completion is indicated by a blank card.

3.3.4. Outputs of the Program

The output of the normal program is almost identical with the output of the exponential program. There are two types of output of the normal program. The first type is an echo check of the inputs to the program and the second type is the set of values calculated by the program. As in the exponential program, the second type of output for the normal program consists of the values of δ, χ^2_{PS}, χ^2_R, $\hat{\alpha}$, X1(10), P2(10), the first M elements of SAMPL(200), and FREQ(10). In addition to these second type outputs, the normal program also prints the number of intersection points and their values. A sample output is contained in Appendix 2.

3.3.5. Special Programming Considerations

Definition of Non-Overlapping Distributions In actuality any two normal distributions will intersect in at least one point, since both span the interval from $-\infty$ to $+\infty$ and both have unit area. However, for widely separated distributions the amount of overlapping area is small and at some point could be considered zero. In the program, if $u_S + 5\sigma_S$ is less than $u_L - 5\sigma_L$ (where u_S refers to the smaller mean, u_L refers to the larger mean, and σ_S and σ_L refer to the corresponding standard deviations) then the distributions are defined as non-overlapping and δ is set equal to 2.

Using the McGill RNG   Use of the McGill RNG is accomplished in this program through the use of two subprograms - RSTART and RNOR. The use of RSTART is described in section 2.3.4. RNOR is used to generate a sample from a standard normal distribution. Its use is similar to the use of REXP described in Section 2.3.4. The sample from a standard normal distribution is transformed to a sample from a normal distribution with mean u and standard deviation σ by the equation

$$X = u + \sigma z \tag{38}$$

where z is the sample from a standard normal distribution and X is the observation from the desired distribution.

Sorting the Random Sample Observations   The normal index of noncongruity program uses a different sorting scheme than the exponential program did because of the different shapes of the two distributions. The sorting scheme is shown in the tree diagram shown in Figure 3-3. This scheme is designed to search the middle equi-probability regions first. The efficiency of this sorting scheme is dependent on the nature of the alternative distribution and, therefore, the scheme does not minimize the expected number of tests in all situations. However, the scheme does reduce the maximum number of tests to 5, as compared with 9 in the scheme used in the exponential program.

Calculating the Normal Cumulative Probability of X   The cumulative probability for an argument X is calculated by first converting the number to standard form by the transformation,

$$z = (x - u)/\sigma \ . \tag{39}$$

The cumulative probability is then calculated by the IBM Scientific Subroutine Package subroutine NDTR which uses the following approximation taken from Hastings

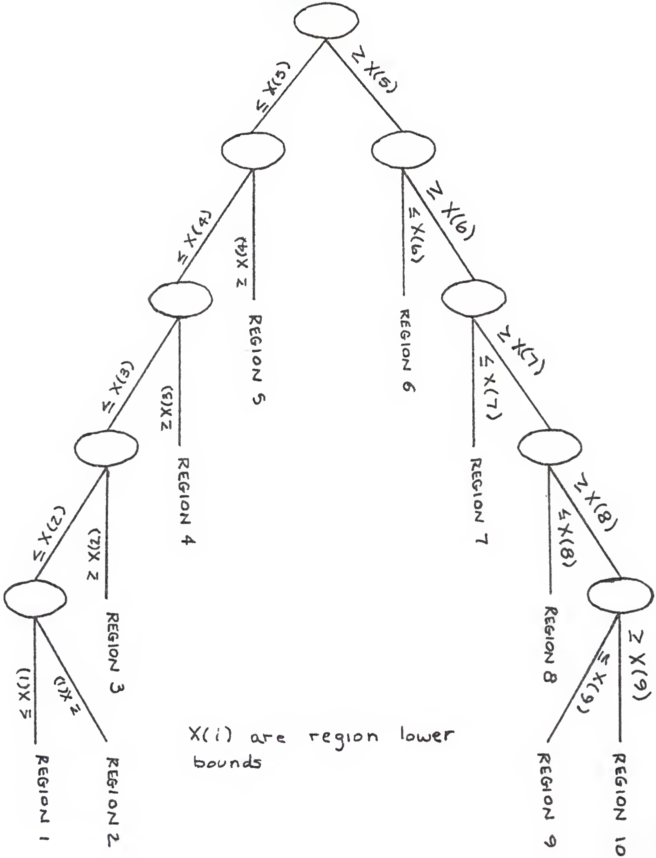

FIGURE 3-3 Tree Diagram for Sorting Procedure

$$F(z) = 1 - f(z)(b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) \tag{40}$$

where

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^{2/2}} , \quad t = \frac{1}{1 + rz} , \quad r = .2316419 ,$$

$$b_1 = .31938153, \quad b_2 = -.356563782, \quad b_3 = 1.7181477937,$$

$$b_4 = -1.821255978, \text{ and } b_5 = 1.33027449$$

This approximation has a maximum error of 7.5×10^{-8} and is valid only for $z \geq 0$. For $z < 0$ the complement of $F(-z)$ gives the desired value.

Calculating X given a Normal Cumulative Probability A value X from a $N(\mu,\sigma^2)$ population can be calculated from a given normal cumulative probability P by first calculating the value z from a standard normal distribution with cumulative probability P and then applying the transformation given in Equation (38). The value z is calculated by the IBM Scientific Subroutine Package subroutine NDTRI which uses the following approximation taken from Hastings,

$$z = w - \sum_{i=o}^{2} a_i w^i / \sum_{i=o}^{3} b_i w^i \tag{41}$$

where $w = \sqrt{\ln(1/p^2)}$, $a_0 = 2.515517$, $a_1 = .802853$, $a_2 = .010328$,

$b_o = 1$, $b_1 = 1.432788$, $b_2 = .189269$, $b_3 = .001308$

This approximation has a maximum error of 4.5×10^{-4} and is valid only for $P \leq .5$. For $P > .5$, z of 1-P is calculated and then the sign of z is changed.

3.4 RESULTS OF THE NORMAL PROGRAM

Four different sets of values for random number generator seeds were used in investigating the normal case. The standard normal distribution was chosen as the model distribution. Three different types of alternative distributions were considered. In the first type the mean of the alternative distribution was different from zero and the standard deviation was equal to one. This set of alternative distributions is referred to as the mean-variate set. In the second type the mean of the alternative distribution was equal to zero and the standard deviation was different from one. This set of alternative distributions is referred to as the variance-variate set. In the third type of alternative distribution considered, the mean of the alternative distribution was different from zero and the standard deviation was different from one. This set is referred to as the mean-variance-variate set. The sample size used in the comparisons was 50 in all cases.

One of the steps in the comparison procedure outlined in Section 1.3.2 is the calculation of various parametric indicators. The parametric indicator which was chosen in the normal case was,

$$\eta_1 = |u_1 - u_2| + |\sigma_1 - \sigma_2| \qquad (42)$$

Various relationships between comparison indices are graphically presented in Figures 3-4 to 3-20. There appears to be a strong relationship between all of the comparison indices plotted. Examination of these figures provides some valuable information. Figure 3-4 shows the relationship between χ^2_{PS} and δ. This figure indicates that for the cases of the mean-variate and variance-variate sets of alternative distributions χ^2_{PS} and δ are strongly related. The figure also indicates that two distributions with a particular δ value are more easily detected as being significantly different if their means are different than if their standard deviations are different. The figure also shows a greater difference in χ^2_{PS} for a given δ value for $\sigma_1 < \sigma_2$ than for $\sigma_1 > \sigma_2$. (Recall that in the exponential case this type of a relationship existed; χ^2_{PS} for $\lambda_2/\lambda_1 < 1$ was greater than χ^2_{PS} for $\lambda_2/\lambda_1 > 1$. This is a similar result since the standard deviation of an exponential distribution is $1/\lambda$ and therefore

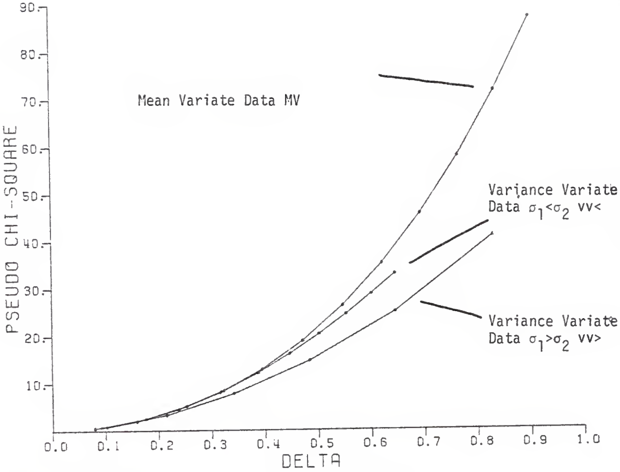

FIGURE 3-4 χ^2_{PS} vs. δ - Single Variation

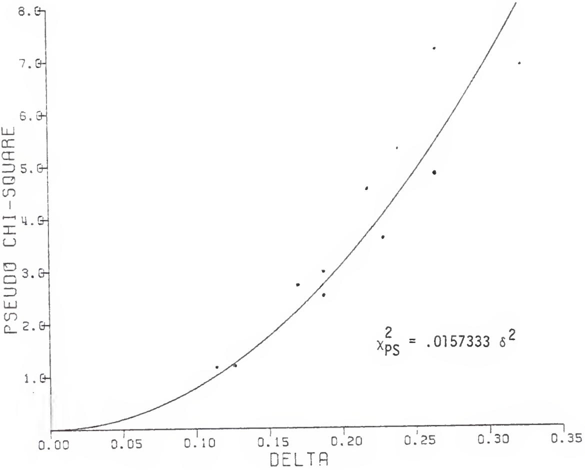

FIGURE 3-5 χ^2_{PS} vs. δ - Dual Variation

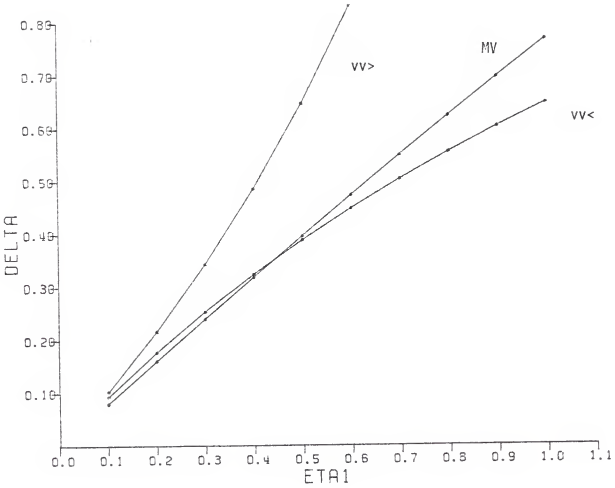

FIGURE 3-6 δ vs. η_1 - Single Variation

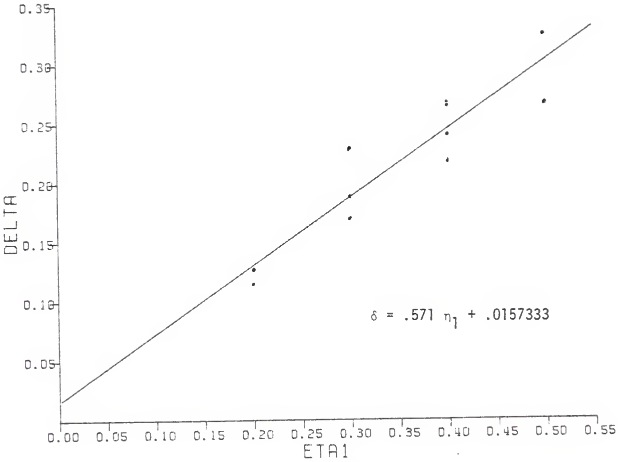

FIGURE 3-7 δ vs. η_1 - Dual Variation

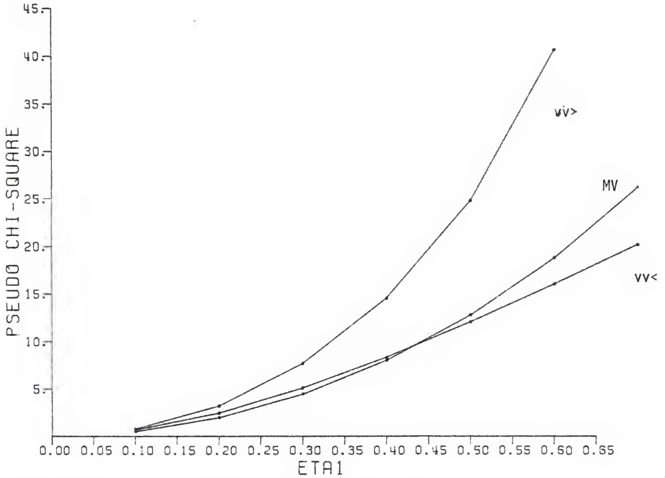

FIGURE 3-8 χ^2_{PS} vs. η_1 - Single Variation

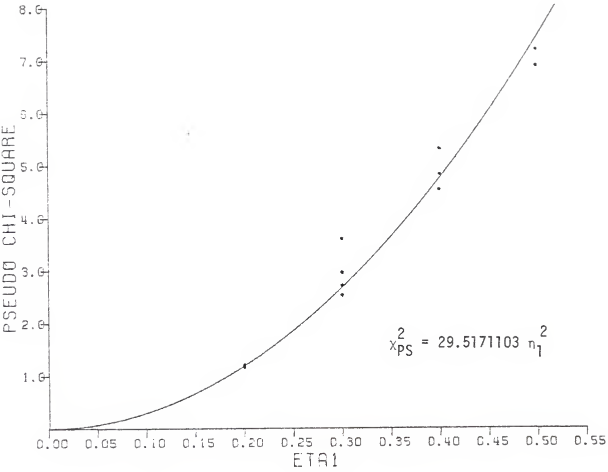

FIGURE 3-9 χ^2_{PS} vs. η_1 - Dual Variation

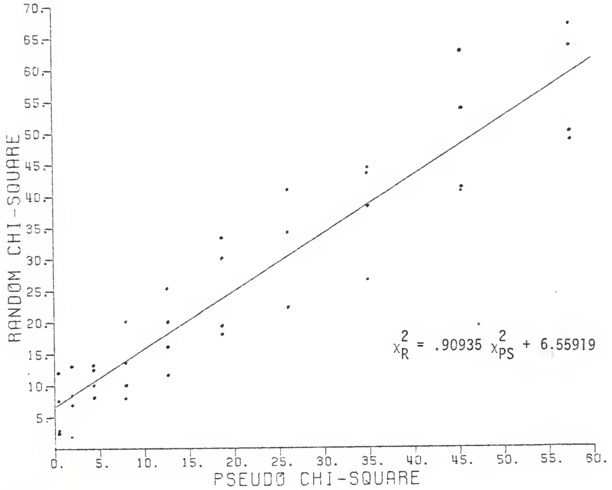

FIGURE 3-10 χ_R^2 vs. χ_{PS}^2 - Mean-Variate Data

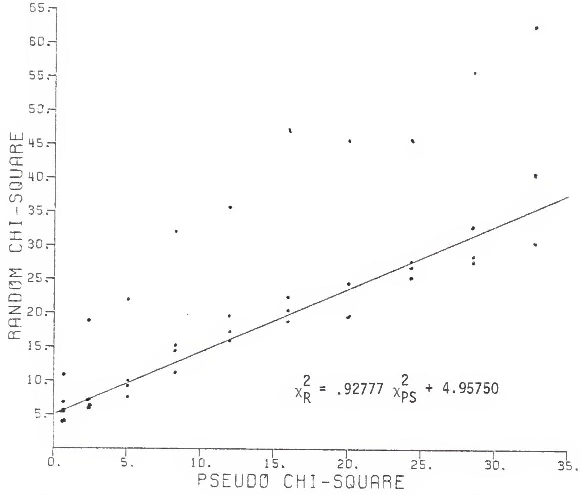

FIGURE 3-11 χ_R^2 vs. χ_{PS}^2 - Variance-Variate Data $\sigma_1<\sigma_2$

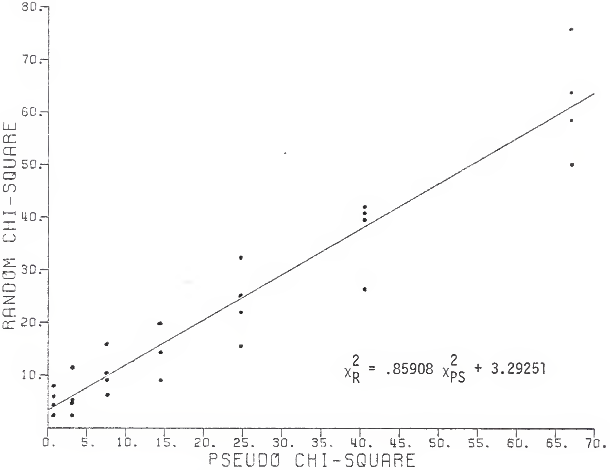

FIGURE 3-12 χ_R^2 vs. χ_{PS}^2 - Variance-Variate Data $\sigma_1 > \sigma_2$

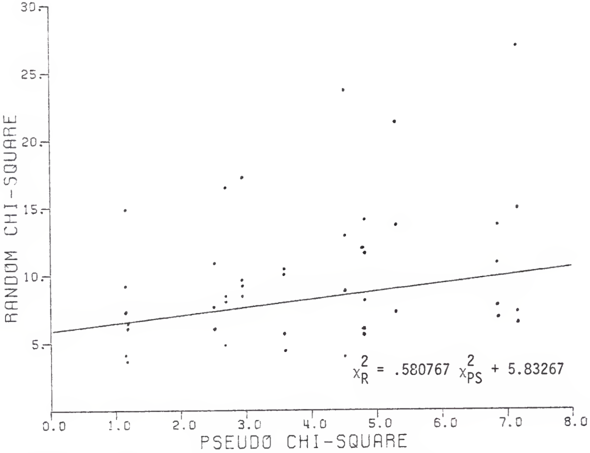

FIGURE 3-13 χ_R^2 vs. χ_{PS}^2 - Dual Variation

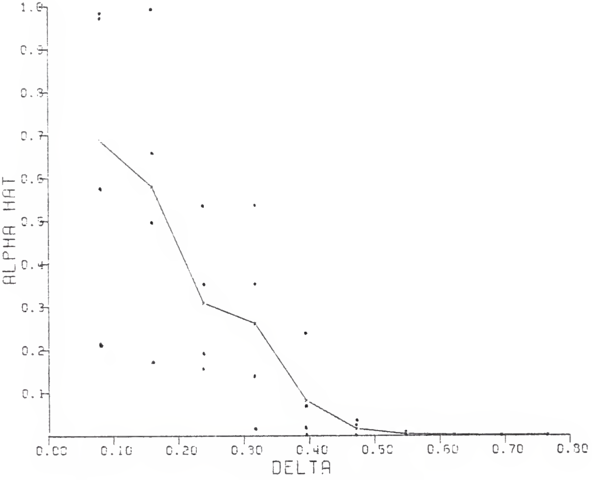

FIGURE 3-14 $\hat{\alpha}$ vs. δ - Mean-Variate Data

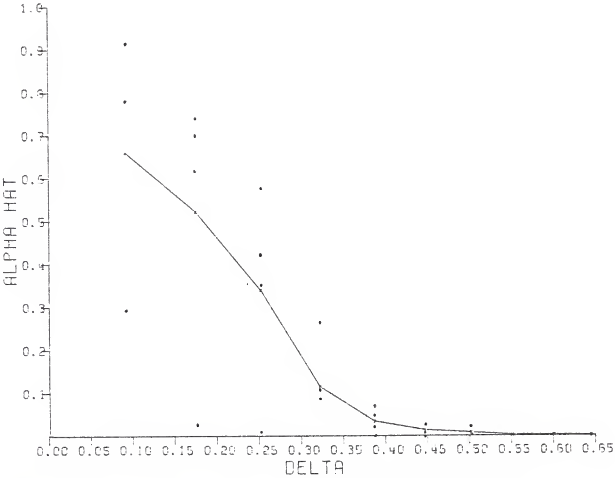

FIGURE 3-15 $\hat{\alpha}$ vs. δ - Variance-Variate Data $\sigma_1 < \sigma_2$

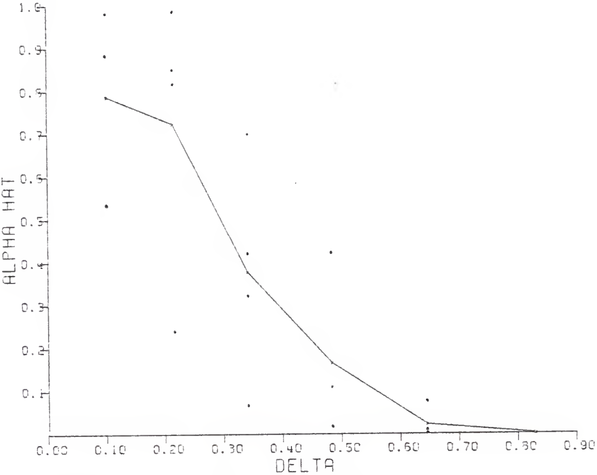

FIGURE 3-16 $\hat{\alpha}$ vs. δ - Variance-Variate Data $\sigma_1 > \sigma_2$

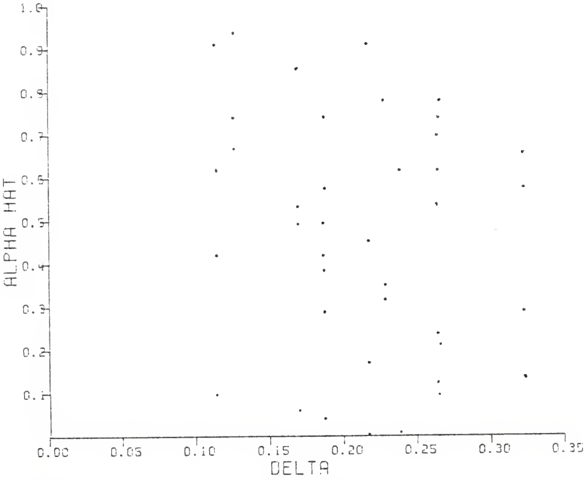

FIGURE 3-17 $\hat{\alpha}$ vs. δ - Dual Variation

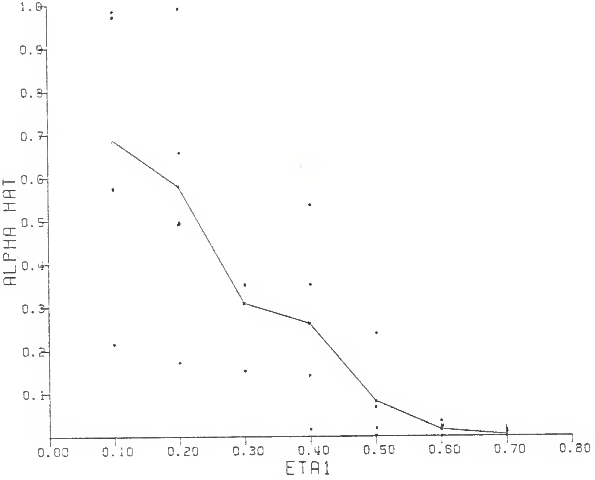

FIGURE 3-18 $\hat{\alpha}$ vs. η_1.- Mean-Variate Data

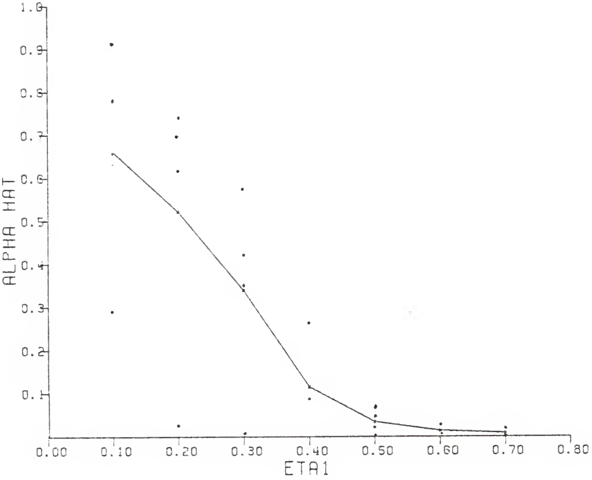

FIGURE 3-19 $\hat{\alpha}$ vs. η_1 - Variance-Variate Data $\sigma_1 < \sigma_2$

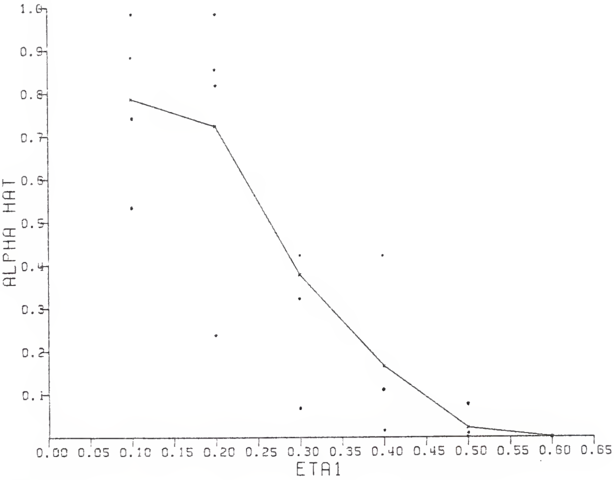

FIGURE 3-20 $\hat{\alpha}$ vs. η_1 - Variance-Variate Data $\sigma_1 \neq \sigma_2$

the ratio of the standard deviations is $1/\lambda_2/1/\lambda_1 = \lambda_1/\lambda_2$. Hence if $\lambda_2/\lambda_1 < 1$ then $\sigma_{exp1} < \sigma_{exp2}$.)

Figure 3-6 indicates the relationship between δ and η_1 for the mean-variate and variance-variate data. The most sensitive case is for the variance-variate case with $\sigma_1 > \sigma_2$. This sensitivity is partially offset by the relationship between χ^2_{PS} and δ for this case (which is not as sensitive as the other two cases shown) as shown in Figure 3-4. However it is not completely offset as seen in Figure 3-8 which shows that the case of variance-variate data with $\sigma_1 > \sigma_2$ produces a much higher χ^2_{PS} value for a given η_1 than the other two cases (which produce comparable values).

As in the exponential case χ^2_{PS} tends to be smaller than χ^2_R, as seen in Figures 3-10 to 3-13. There also appears to be a linear trend between χ^2_R and χ^2_{PS} in each of these figures. A least-squares line was calculated for each of these cases. The obviously outlying points of Figures 3-11 and 3-13 were omitted from the calculations. The derived lines are

Figure 3-10 Mean-Variate Data

$$\chi^2_R = .90935 \ \chi^2_{PS} + 6.55919$$

Figure 3-11 Variance-Variate Data $\sigma_1 < \sigma_2$

$$\chi^2_R = .92777 \ \chi^2_{PS} + 4.95750$$

Figure 3-12 Variance-Variate Data $\sigma_1 > \sigma_2$

$$\chi^2_R = .85908 \ \chi^2_{PS} + 3.29251$$

Figure 3-13 Mean-Variance-Variate Data

$$\chi^2_R = .580767 \ \chi^2_{PS} + 5.83267$$

From Figures 3-14 to 3-16 it appears that (based on $\hat{\alpha}$ as the criterion) δ tends to be significant at the .05 level as indicated below:

Figure	Type	Significance Range
3-14	Mean-Variate Data	$\delta \geq .47$

Figure	Type	Significance Range
3-15	Variance-Variate Data $\sigma_1 < \sigma_2$	$\delta \geq .39$
3-16	Variance-Variate Data $\sigma_1 > \sigma_2$	$\delta \geq .6$

If χ^2_{PS} is used as the criterion, χ^2_{PS} would indicate significance at the following values calculated from the regression equations to correspond to a χ^2_R value of 17 ($\hat{\alpha} = .05$).

Type		Significance Range
Mean-Variate Data		$\chi^2_{PS} \geq 11.5$
Variance-Variate Data	$\sigma_1 < \sigma_2$	$\chi^2_{PS} \geq 13.0$
Variance-Variate Data	$\sigma_1 > \sigma_2$	$\chi^2_{PS} \geq 16.0$

Based on these values and consulting Figure 3-4, δ would indicate significance at the .05 level as given below.

Type		Significance Range
Mean-Variate Data		$\delta \geq .36$
Variance-Variate Data	$\sigma_1 < \sigma_2$	$\delta \geq .4$
Variance-Variate Data	$\sigma_1 > \sigma_2$	$\delta \geq .55$

The uncertainty regions can be estimated as

Mean-Variate Data		$.38 \leq \delta \leq .47$	
Variance-Variate Data	$\sigma_1 < \sigma_2$	$.34 \leq \delta \leq .39$	$\hat{\alpha}$ basis
Variance-Variate Data	$\sigma_1 > \sigma_2$	$.48 \leq \delta \leq .6$	
Mean-Variate Data		$.32 \leq \delta \leq .36$	
Variance-Variate Data	$\sigma_1 < \sigma_2$	$.34 \leq \delta \leq .4$	χ^2_{PS} basis
Variance-Variate Data	$\sigma_1 > \sigma_2$	$.45 \leq \delta \leq .55$	

These values for δ imply that if η_1 is to be used as a test for significance instead of δ, then the uncertainty regions can be estimated from Figure 3-6 to be

Mean-Variate Data		$.48 \leq \eta_1 \leq .6$	
Variance-Variate Data	$\sigma_1 < \sigma_2$	$.43 \leq \eta_1 \leq .5$	$\hat{\alpha}$ basis
Variance-Variate Data	$\sigma_1 > \sigma_2$	$.4 \leq \eta_1 \leq .46$	
Mean-Variate Data		$.4 \leq \eta_1 \leq .44$	
Variance-Variate Data	$\sigma_1 < \sigma_2$	$.42 \leq \eta_1 \leq .52$	χ^2_{PS} basis
Variance-Variate Data	$\sigma_1 > \sigma_2$	$.38 \leq \eta_1 \leq .42$	

These regions compare favorably with results which can be taken from Figures 3-18 to 3-20 for the $\hat{\alpha}$ basis and Figure 3-8 for the χ^2_{PS} basis. Significance is therefore indicated for η_1 values of

Mean-Variate Data		$\eta_1 \geq .6$	
Variance-Variate Data	$\sigma_1 < \sigma_2$	$\eta_1 \geq .5$	$\hat{\alpha}$ basis
Variance-Variate Data	$\sigma_1 > \sigma_2$	$\eta_1 \geq .46$	
Mean-Variate Data		$\eta_1 \geq .44$	
Variance-Variate Data	$\sigma_1 < \sigma_2$	$\eta_1 \geq .52$	χ^2_{PS} basis
Variance-Variate Data	$\sigma_1 > \sigma_2$	$\nu_1 \geq .42$	

Figures 3-5, 3-7, 3-9, 3-13 and 3-17 show the various relationships between indices for the mean-variance-variate data. Indications of strong relationships between the various indices are shown by these figures. However, it is believed that the knowledge of these relationships is too limited to draw any satisfactory results. Further research, in which the range of the comparison indices is expanded, is needed to better quantify these relationships.

TABLE 3-1

Results of the Normal Program

DISTRIBUTION 2 (Alternative)	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
(.1, 1)	A	.0798	.48	7.60	.5749
	B			2.80	.9717
	C			2.40	.9835
	D			12.00	.2133
(.2, 1)	A	.1593	1.94	8.40	.4944
	B			2.00	.9915
	C			6.80	.6579
	D			12.80	.1719
(.3, 1)	A	.2385	4.41	13.20	.1538
	B			10.00	.3505
	C			8.00	.5341
	D			12.40	.1917
(.4, 1)	A	.3170	7.97	13.60	.1373
	B			8.00	.5341
	C			10.00	.3505
	D			20.00	.0179
(.5, 1)	A	.3948	12.70	20.00	.0179
	B			11.60	.2368
	C			16.00	.0669
	D			25.20	.0028
(.6, 1)	A	.4716	18.70	30.00	.0004
	B			19.20	.0235
	C			18.00	.0352
	D			33.20	.0001
(.7, 1)	A	.5473	26.08	34.00	.0001
	B			22.00	.0089
	C			30.40	.0004
	D			40.80	.0000
(.8, 1)	A	.6217	34.94	43.20	.0000
	B			26.40	.0018
	C			38.00	.0000
	D			44.00	.0000
(.9, 1)	A	.6946	45.38	62.40	.0000
	B			40.40	.0000
	C			40.80	.0000
	D			53.20	.0000

Table 3-1 continued

DISTRIBUTION 2 (Alternative)	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
(1.0, 1)	A	.7659	57.45	66.40	.0000
	B			48.40	.0000
	C			49.60	.0000
	D			63.20	.0000
(1.1, 1)	A	.8354	71.19	83.60	.0000
	B			67.20	.0000
	C			66.80	.0000
	D			77.60	.0000
(1.2, 1)	A	.9030	86.58	97.20	.0000
	B			76.00	.0000
	C			78.80	.0000
	D			92.80	.0000
(1.3, 1)	A	.9686	103.55	137.60	.0000
	B			85.60	.0000
	C			89.20	.0000
	D			106.00	.0000
(1.4, 1)	A	1.0321	121.99	143.20	.0000
	B			95.60	.0000
	C			111.60	.0000
	D			123.60	.0000
(1.5, 1)	A	1.0935	141.70	172.40	.0000
	B			118.80	.0000
	C			126.00	.0000
	D			126.00	.0000
(1.6, 1)	A	1.1526	162.46	184.00	.0000
	B			147.20	.0000
	C			141.20	.0000
	D			156.40	.0000
(1.7, 1)	A	1.2093	184.00	206.40	.0000
	B			163.20	.0000
	C			155.20	.0000
	D			180.00	.0000
(1.8, 1)	A	1.2638	206.00	246.40	.0000
	B			173.60	.0000
	C			175.60	.0000
	D			180.80	.0000
(1.9, 1)	A	1.3158	228.15	274.80	.0000
	B			197.20	.0000
	C			235.60	.0000
	D			206.80	.0000

Table 3-1 continued

DISTRIBUTION 2 (Alternative)	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
(2.0, 1)	A	1.3654	250.11	290.80	.0000
	B			226.80	.0000
	C			261.20	.0000
	D			220.80	.0000
(0, .1)	A	1.5964	194.35	182.80	.0000
	B			174.40	.0000
	C			192.80	.0000
	D			181.20	.0000
(0, .2)	A	1.2942	117.35	95.60	.0000
	B			125.60	.0000
	C			113.60	.0000
	D			88.80	.0000
(0, .3)	A	1.0435	67.04	63.60	.0000
	B			58.40	.0000
	C			75.60	.0000
	D			50.00	.0000
(0, .4)	A	.8300	40.56	39.60	.0000
	B			40.80	.0000
	C			42.00	.0000
	D			26.40	.0018
(0, .5)	A	.6453	24.72	25.20	.0028
	B			22.00	.0089
	C			32.40	.0002
	D			15.60	.0757
(0, .6)	A	.4840	14.48	14.40	.1088
	B			14.40	.1088
	C			20.00	.0179
	D			9.20	.4190
(0, .7)	A	.3416	7.63	9.20	.4190
	B			16.00	.0669
	C			10.40	.3191
	D			6.40	.6993
(0, .8)	A	.2151	3.19	4.80	.8514
	B			5.20	.8165
	C			11.60	.2368
	D			2.40	.9835
(0, .9)	A	.1019	.75	6.00	.7399
	B			2.40	.9835
	C			8.00	.5341
	D			4.40	.8832

Table 3-1 contined

DISTRIBUTION 2 (Alternative)	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
(0, 1.1)	A	.0922	.65	5.60	.7792
	B			6.80	.6579
	C			4.00	.9114
	D			10.80	.2897
(0, 1.2)	A	.1760	2.42	6.40	.6993
	B			7.20	.6163
	C			6.00	.7399
	D			18.80	.0269
(0, 1.3)	A	.2525	5.05	9.20	.4190
	B			10.00	.3505
	C			7.60	.5749
	D			22.00	.0089
(0, 1.4)	A	.3226	8.29	15.20	.0856
	B			11.20	.2622
	C			14.40	.1088
	D			32.00	.0002
(0, 1.5)	A	.3872	11.97	17.20	.0457
	B			19.60	.0205
	C			16.00	.0669
	D			35.60	.0000
(0, 1.6)	A	.4467	15.93	18.80	.0269
	B			20.40	.0156
	C			22.40	.0077
	D			47.20	.0000
(0, 1.7)	A	.5019	20.07	19.60	.0205
	B			24.40	.0037
	C			24.40	.0037
	D			45.60	.0000
(0, 1.8)	A	.5531	24.29	27.60	.0011
	B			26.80	.0015
	C			25.20	.0028
	D			45.60	.0000
(0, 1.9)	A	.6008	28.52	27.60	.0011
	B			32.80	.0001
	C			28.40	.0008
	D			55.60	.0000
(0, 2.0)	A	.6453	32.73	30.40	.0004
	B			40.40	.0000
	C			40.40	.0000
	D			62.40	.0000

Table 3-1 continued

DISTRIBUTION 2 (Alternative)	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
(.1, .8)	A	.2282	3.59	5.60	.7792
	B			10.00	.3505
	C			10.40	.3191
	D			4.40	.8832
(.1, .9)	A	.1262	1.19	6.40	.6693
	B			3.60	.9357
	C			6.00	.7399
	D			6.00	.7399
(.1, 1.1)	A	.1142	1.16	9.20	.4190
	B			7.20	.6163
	C			4.00	.9114
	D			14.80	.0966
(.1, 1.2)	A	.1868	2.95	9.60	.3838
	B			8.40	.4944
	C			9.20	.4190
	D			17.20	.0457
(.2, .8)	A	.2659	4.79	12.00	.2133
	B			5.60	.7792
	C			6.00	.7399
	D			6.00	.7399
(.2, .9)	A	.1876	2.52	7.60	.5749
	B			10.80	.2897
	C			6.00	.7399
	D			6.00	.7399
(.2, 1.1)	A	.1698	2.70	8.40	.4944
	B			4.80	.8514
	C			8.00	.5341
	D			16.40	.0590
(.2, 1.2)	A	.2177	4.52	12.80	.1719
	B			4.00	.9114
	C			8.80	.4559
	D			23.60	.0050
(.3, .8)	A	.3230	6.86	13.60	.1373
	B			6.80	.6579
	C			7.60	.5749
	D			10.80	.2897
(.3, .9)	A	.2639	4.81	11.60	.2368
	B			6.00	.7399
	C			8.00	.5341
	D			14.00	.1223

Table 3-1 continued

DISTRIBUTION 2 (Alternative)	RNG SEED	δ	χ^2_{PS}	χ^2_R	$\hat{\alpha}$
(.3, 1.1)	A	.2389	5.29	13.60	.1373
	B			7.20	.6163
	C			7.20	.6163
	D			21.20	.0118
(.3, 1.2)	A	.2648	7.16	14.80	.0966
	B			6.40	.6993
	C			7.20	.6163
	D			26.80	.0015

M = 50, RNG: A(51562, 62155) B(62155, 51562) C(50020, 11292) D(11292, 50020)

Chapter 4

CONCLUSION

In concluding this study, two questions need to be answered. The first question is "How well did this study accomplish its objective?". The second question is "What direction should future research in this area take?".

The first question can be answered by reconsidering the objective of this study which was to develop a comparison method, use this method to investigate the effects of varying distribution parameters, and to evaluate the usefulness of this technique. The first two parts of this objective have already been accomplished and the third part can be completed by a brief review of the results of the study. The technique used in this study to compare statistical distributions appears to have considerable usefulness because of the consistency of results (i.e. δ in all cases indicated significance in the range $.3 \leq \delta \leq .6$), the ability to compare distributions for statistical difference in terms of only their parameters (λ_2/λ_1 for exponential distributions, n_1 for most normal distributions), and the relative simplicity of the method.

The reader should recognize that in the normal case, a technique already exists to answer the question of "How different is different?" based on the classical z-test. This technique is statistically sufficient and is therefore more powerful than the method presented in this study. The use of the z-test is demonstrated below for M = 50, $\alpha = .05$, and $\sigma_1 = \sigma_2 = \sigma$.

$$1.96 \leq (u_1 - u_2)/\sqrt{\sigma^2/50} \qquad (43)$$

$$u_1 - u_2 \geq .277\ \sigma \qquad (44)$$

This indicates that for our investigation ($\sigma = 1$) significance would be indicated for a η_1 value greater than or equal to .277, as compared to a η_1 value of .60 for the index of non-congruity method.

The second question concerning the direction of future research is easily answered. There are four readily apparent directions for future research. They are:

1. Extension of this research in terms of additional replications and inclusion of more mean-variance-variate comparisons for the normal case as mentioned in Sections 2.4 and 3.4.
2. Application of the methodology used in this study to other continuous distributions such as the Weibull, Log-Normal, or Gamma distribution.
3. Development of a similar methodology which can be applied to the evaluation of statistically significant differences in discrete distributions.
4. Application of the methodology used in this study to study statistical differences of similar distributions which are from different families.

SELECTED BIBLIOGRAPHY

(1) Hasting, Cecil, Jr., Approximations for Digital Computers, Princeton: Princeton University Press, 1955.

(2) IBM System/360 Scientific Subroutine Package Version III Programmer's Manual, White Plains, New York: IBM Corporation, Technical Publications Department, 1968.

(3) Lakshminarayan, K., "Progress Report to the U.S. Nuclear Regulatory Commission under NRC Contract No. AT(49-24)-0339 on the Analysis of Diesel Engine Failure Probability Data", (Kansas State University Document KSU-2662-1, CES-24, October, 1976).

(4) _________, "Estimating Statistically Significant Differences Between a Pair of Beta Distributions", (Unpublished Masters Thesis, Kansas State University, 1978).

APPENDIX 1

FORTRAN IV G LEVEL 21 MAIN DATE = 78094 10/03/40

```
      C     PROGRAM TO EVALUATE "THE INDEX OF NON-CONGRUITY" FOR EXPONENTIAL
      C     DISTRIBUTIONS.  THIS INDEX IS BASED ON THE AMOUNT OF NON-OVERLAPPING
      C     AREA BETWEEN TWO DISTRIBUTIONS.
      C     VALUES TO BE SUPPLIED TO THE PROGRAM ARE DISTRIBUTION PARAMETERS,
      C      SAMPLE SIZE, AND INITIAL VALUES FOR THE RNG
      C     FORMAT FOR DATA CARDS:  LAMDA1,LAMDA2,M  1 CARD  (2E10.4,I5)
      C     ISEED,JSEED  1 CARD (2I5)
      C     MULTIPLE RUNS ARE POSSIBLE BY SUPPLYING ADDITIONAL INPUT CARDS
      C     (TWO PER REPLICATION).  PROGRAM COMPLETION IS INDICATED BY A
      C     BLANK CARD.
            DIMENSION XI(10),P2(10),SAMPL(200),FREQ(10)
            REAL LAMDA1,LAMDA2,NU
      C     INITIALIZE PROGRAM PARAMETERS
        200 X2SUM=0
            F2SUM=0
            DO 11 I=1,10
         11 FREQ(I)=0
      C     INPUT VALUES FOR THE PARAMETERS OF THE EXPONENTIAL DISTRIBUTIONS
      C     AND THE SAMPLE SIZE.
            READ(5,99)LAMDA1,LAMDA2,M
         99 FORMAT(2E10.4,I5)
      C     CHECK FOR PROGRAM COMPLETION
            IF(LAMDA1.EQ.0) GO TO 201
      C     CHECK TO DETERMINE WHICH DISTRIBUTION PARAMETER IS LARGER
            IF(LAMDA2.LT.LAMDA1) GO TO 120
      C     LAMDA2 IS GREATER THAN OR EQUAL TO LAMDA1.  EXPLICITLY DETERMINE
      C     DELTA, THE INDEX OF NON-CONGRUITY
            RATIO=LAMDA2/LAMDA1
            DIFFL=LAMDA2-LAMDA1
            T1=ALOG(RATIO)/DIFFL
            Z1=-LAMDA1*T1
            Z2=-LAMDA2*T1
            DELTA=2*(EXP(Z1)-EXP(Z2))
            GO TO 121
      C     LAMDA1 IS GREATER THAN LAMDA2.  EXPLICITLY DETERMINE DELTA, THE INDEX
      C     OF NON-CONGRUITY
        120 RATIO=LAMDA1/LAMDA2
            DIFFL=LAMDA1-LAMDA2
            T1=ALOG(RATIO)/DIFFL
            Z1=-LAMDA1*T1
            Z2=-LAMDA2*T1
            DELTA=2*(EXP(Z2)-EXP(Z1))
      C     OUTPUT VALUES OF M, LAMDA1, LAMDA2
        121 WRITE(6,98)M,LAMDA1,LAMDA2
         98 FORMAT(1H1,///,11X,'THE SAMPLE SIZE EQUALS',I6,//,11X,'LAMDA1 EQUAL
           1LS',E12.4,//,11X,'LAMDA2 EQUALS',E12.4)
      C     CALCULATE THE EXPECTED VALUE FOR CELL FREQUENCIES
            EI=M/10.
      C     DETERMINE EQUAL PROBABILITY REGIONS FOR DISTRIBUTION 1 AND DETERMINE
      C     THE CUMULATIVE PROBABILITY ASSOCIATED WITH THESE REGIONS FOR
      C     DISTRIBUTION 2
            XI(1)=-ALOG(.9)/LAMDA1
            P2(1)=1-EXP(-LAMDA2*XI(1))
            X2SUM=X2SUM+(EI-M*(P2(1)-0))**2
            DO 1 I=2,9
            XI(I)=-ALOG(1-.1*I)/LAMDA1
            P2(I)=1-EXP(-LAMDA2*XI(I))
            X2SUM=X2SUM+(EI-M*(P2(I)-P2(I-1)))**2
```

```
0034          1 CONTINUE
0035            X2SUM=X2SUM+(EI-M*(1-P2(9)))**2
        C       CALCULATE THE PSEUDO CHI-SQUARE STATISTIC
0036            X2PS=X2SUM/EI
        C       GENERATE RANDOM SAMPLE FROM SECOND DISTRIBUTION
        C       READ RANDOM NUMBER GENERATOR SEED VALUES
0037            READ(5,97)ISEED,JSEED
0038         97 FORMAT(2I5)
        C       ECHO SEED VALUES
0039            WRITE(6,96)ISEED,JSEED
0040         96 FORMAT(1X,/////,1X,'THE SEED VALUES FOR THIS RUN ARE',///,11X,'ISE
               1ED EQUALS',I12,//,11X,'JSEED EQUALS',I12)
        C       INITIALIZE RANDOM NUMBER GENERATOR
0041            CALL RSTART(ISEED,JSEED)
        C       GENERATE SAMPLE
0042            DO 2 I=1,M
0043            SAMPL(I)=REXP(1)/LAMDA2
0044          2 CONTINUE
        C       SORT RANDOM OBSERVATIONS INTO FREQUENCY CLASSES
0045        100 DO 3 I=1,M
0046            IF(SAMPL(I).LE.XI(1)) GO TO 101
0047            IF(SAMPL(I).LE.XI(2)) GO TO 102
0048            IF(SAMPL(I).LE.XI(3)) GO TO 103
0049            IF(SAMPL(I).LE.XI(4)) GO TO 104
0050            IF(SAMPL(I).LE.XI(5)) GO TO 105
0051            IF(SAMPL(I).LE.XI(6)) GO TO 106
0052            IF(SAMPL(I).LE.XI(7)) GO TO 107
0053            IF(SAMPL(I).LE.XI(8)) GO TO 108
0054            IF(SAMPL(I).LE.XI(9)) GO TO 109
0055            K=10
0056            GO TO 110
0057        101 K=1
0058            GO TO 110
0059        102 K=2
0060            GO TO 110
0061        103 K=3
0062            GO TO 110
0063        104 K=4
0064            GO TO 110
0065        105 K=5
0066            GO TO 110
0067        106 K=6
0068            GO TO 110
0069        107 K=7
0070            GO TO 110
0071        108 K=8
0072            GO TO 110
0073        109 K=9
0074        110 FREQ(K)=FREQ(K)+1
0075          3 CONTINUE
        C       CALCULATE THE ACTUAL CHI-SQUARE STATISTIC FOR RANDOM SAMPLE
0076            DO 4 I=1,10
0077            F2SUM=F2SUM+FREQ(I)**2
0078          4 CONTINUE
0079            X2ACT=F2SUM/EI-M
0080            NU=9.
        C       CALL FUNCTION TO CALCULATE "ALPHA HAT" FOR THE COMPUTED CHI-SQUARE
        C       VALUE
```

```
FORTRAN IV G LEVEL  21                    MAIN               DATE = 78094          10/03/40

0081              AHAT=CADTR(X2ACT,NU)
            C     OUTPUT VALUES OF DELTA, PSEUDO CHI-SQUARE, CHI-SQUARE, AND ALPHA HAT
0082              WRITE(6,95) DELTA,X2PS,X2ACT,AHAT
0083           95 FORMAT(/////,11X,'THE VALUE OF THE INDEX OF NON-CONGRUITY(DELTA) E
                 1QUALS',F12.4,//,11X,'THE VALUE OF THE PSEUDO CHI-SQUARE STATISTIC
                 2EQUALS',F12.2,//,11X,'THE VALUE OF THE CHI-SQUARE STATISTIC EQUALS
                 3',F12.2,//,11X,'THE AREA OF THE CHI-SQUARE DISTRIBUTION TO THE RIG
                 4HT OF THE CHI-SQUARE STATISTIC (ALPHA HAT) EQUALS',F12.4)
0084              WRITE(6,94)
0085              WRITE(6,92)(X1(I),I=1,10)
0086              WRITE(6,91)(P2(I),I=1,10)
0087              WRITE(6,93)(SAMPL(I),I=1,M)
0088              WRITE(6,90)(FREQ(I),I=1,10)
0089           94 FORMAT(//////)
0090           93 FORMAT('0','RANDOM SAMPLE',10F11.4)
0091           92 FORMAT('0','REGION BOUNDRIES',10F11.4)
0092           91 FORMAT('0','CUMULATIVE PROB',10F11.4)
0093           90 FORMAT('0','CELL FREQUENCY',10F11.1)
0094              GO TO 200
0095          201 STOP
0096              END
```

```
0001          FUNCTION CAOTR(X,G)
        C
        C        PURPOSE
        C           COMPUTES P(X) = PROBABILITY THAT THE RANDOM VARIABLE U,
        C           DISTRIBUTED ACCORDING TO THE CHI-SQUARE DISTRIBUTION WITH G
        C           DEGREES OF FREEDOM, IS LESS THAN OR EQUAL TO X.  F(G,X), THE
        C        USAGE
        C          PROB=CDTR(X,G)
        C
        C        DESCRIPTION OF PARAMETERS
        C           X    - INPUT SACLE FOR WHICH P(X) IS COMPUTED.
        C           G    - NUMBER OF DEGREES OF FREEDOM OF THE CHI-SQUARE
        C                  DISTRIBUTION.  G IS A CONTINUOUS PARAMETER.
        C           IER - RESULTANT ERROR CODE WHERE
        C                IER= 0 --- NO ERROR
        C                IER=-1 --- AN INPUT PARAMETER IS INVALID.  X IS LESS
        C                           THAN 0.0, OR G IS LESS THEN 0.5 OR GREATER
        C                           THAN 2*10**(+5).  P AND D ARE SET TO -1.E75.
        C                IER=+1 --- INVALID OUTPUT.  P IS LESS THAN ZERO OR
        C                           GREATER THAN ONE, OR SERIES FOR T1 (SEE
        C                           MATHEMATICAL DESCRIPTION) HAS FAILED TO
        C                           CONVERGE.  P IS SET TO 1.E75.
        C
        C        SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED
        C           OLGAM
        C           NOTR
        C
0002          DOUBLE PRECISION XX,OLXX,X2,OLX2,GG,G2,OLT3,THETA,THP1,
             1TI1,SER,CC,XI,FAC,TLOG,TERM,GTH,A2,A,B,C,DT2,OT3,THPI
        C
        C        TEST FOR VALID INPUT DATA
        C
0003          IF(G-(.5-1.E-5)) 590,10,10
0004       10 IF(G-2.E+5) 20,20,590
0005       20 IF(X) 590,30,30
        C
        C        TEST FOR X NEAR 0.0
        C
0006       30 IF(X-1.E-8) 40,40,80
0007       40 P=0.0
0008          IF(G-2.) 50,60,70
0009       50 D=1.E75
0010          GO TO 610
0011       60 D=0.5
0012          GO TO 610
0013       70 D=0.0
0014          GO TO 610
        C
        C        TEST FOR X GREATER THAN 1.E+6
        C
0015       80 IF(X-1.E+6) 100,100,90
0016       90 D=0.0
0017          P=1.0
0018          GO TO 610
        C
        C        SET PROGRAM PARAMETERS
        C
0019      100 XX=DBLE(X)
```

```
FORTRAN IV G LEVEL  21                CADTR               DATE = 78094          10/03/40

0020             OLXX=OLDG(XX)                                                        CDTR0295
0021             X2=XX/2.D0                                                           CDTR0300
0022             DLX2=OLOG(X2)                                                        CDTR0305
0023             GG=OBLE(G)                                                           CDTR0310
0024             G2=GG/2.D0                                                           CDTR0315
           C                                                                          CDTR0320
           C        TEST FOR G GREATER THAN 1000.0                                    CDTR0325
           C        TEST FOR X GREATER THAN 2000.0                                    CDTR0330
           C                                                                          CDTR0335
0025             IF(G-1000.) 160,160,180                                              CDTR0340
0026         160 IF(X-2000.) 190,190,170                                              CDTR0345
0027         170 P=1.0                                                                CDTR0350
0028             GO TO 610                                                            CDTR0355
0029         180 A=OLDG(XX/GG)/3.D0                                                   CDTR0360
0030             A=DEXP(A)                                                            CDTR0365
0031             B=2.D0/(9.D0*GG)                                                     CDTR0370
0032             C=(A-1.D0+B)/DSQRT(B)                                                CDTR0375
0033             SC=SNGL(C)                                                           CDTR0380
0034             CALL NOTR(SC,P,DUMMY)                                                CDTR0385
0035             GO TO 490                                                            CDTR0390
           C                                                                          CDTR0395
           C        COMPUTE THETA                                                     CDTR0400
           C                                                                          CDTR0405
0036         190 K= IDINT(G2)                                                         CDTR0410
0037             THETA=G2-DFLOAT(K)                                                   CDTR0415
0038             IF(THETA-1.D-8) 200,200,210                                          CDTR0420
0039         200 THETA=0.D0                                                           CDTR0425
0040         210 THP1=THETA+1.D0                                                      CDTR0430
           C                                                                          CDTR0435
           C        SELECT METHOD OF COMPUTING T1                                     CDTR0440
           C                                                                          CDTR0445
0041             IF(THETA)230,230,220                                                 CDTR0450
0042         220 IF(XX-10.D0)260,260,320                                              CDTR0455
           COMPUTE T1 FOR THETA EQUALS 0.0                                            CDTR0460
0043         230 IF(X2-1.68D02) 250,240,240                                           CDTR0465
0044         240 T1=1.0                                                               CDTR0470
0045             GO TO 400                                                            CDTR0475
0046         250 T11=1.D0-DEXP(-X2)                                                   CDTR0480
0047             T1=SNGL(T11)                                                         CDTR0485
0048             GO TO 400                                                            CDTR0490
           C                                                                          CDTR0495
           C        COMPUTE T1 FOR THETA GREATER THAN 0.0 AND                         CDTR0500
           C        X LESS THAN OR EQUAL TO 10.0                                      CDTR0505
           C                                                                          CDTR0510
0049         260 SER=X2*(1.D0/THP1 -X2/(THP1+1.D0))                                   CDTR0515
0050             J=+1                                                                 CDTR0520
0051             CC=DFLOAT(J)                                                         CDTR0525
0052             DO 270 IT1=3,30                                                      CDTR0530
0053             XI=DFLOAT(IT1)                                                       CDTR0535
0054             CALL DLGAM(XI,FAC,IOK)                                               CDTR0540
0055             TLOG= XI*DLX2-FAC-DLOG(XI+THETA)                                     CDTR0545
0056             TERM=DEXP(TLOG)                                                      CDTR0550
0057             TERM=DSIGN(TERM,CC)                                                  CDTR0555
0058             SER=SER+TERM                                                         CDTR0560
0059             CC=-CC                                                               CDTR0565
0060             IF(DABS(TERM)-1.D-9) 280,270,270                                     CDTR0570
0061         270 CONTINUE                                                             CDTR0575
0062             GO TO 600                                                            CDTR0580
```

```
      280 IF(SER) 600,600,290                                           CDTR0585
      290 CALL DLGAM(THP1,GTH,IOK)                                      CDTR0590
          TLOG=THETA*DLX2+DLOG(SER)-GTH                                 CDTR0595
          IF(TLOG+1.68D02) 300,300,310                                  CDTR0600
      300 T1=0.0                                                        CDTR0605
          GO TO 400                                                     CDTR0610
      310 T11=DEXP(TLOG)                                                CDTR0615
          T1=SNGL(T11)                                                  CDTR0620
          GO TO 400                                                     CDTR0625
C                                                                       CDTR0630
C         COMPUTE T1 FOR THETA GREATER THAN 0.0 AND                     CDTR0635
C         X GREATER THAN 10.0 AND LESS THAN 2000.0                      CDTR0640
C                                                                       CDTR0645
      320 A2=0.D0                                                       CDTR0650
          DO 340 I=1,25                                                 CDTR0655
          XI=DFLOAT(I)                                                  CDTR0660
          CALL DLGAM(THP1,GTH,ICK)                                      CDTR0665
          T11=-(13.D0*XX)/XI +THP1*DLOG(13.D0*XX/XI) -GTH-DLOG(XI)      CDTR0670
          IF(T11+1.68D02) 340,340,330                                   CDTR0675
      330 T11=DEXP(T11)                                                 CDTR0680
          A2=A2+T11                                                     CDTR0685
      340 CONTINUE                                                      CDTR0690
          A=1.0(282051+THETA/150.D0-XX/312.D0                           CDTR0695
          B=DABS(A)                                                     CDTR0700
          C= -X2+THP1*DLX2+DLOG(B)-GTH-3.951243718581427                CDTR0705
          IF(C+1.68D02) 370,370,350                                     CDTR0710
      350 IF (A) 360,370,380                                            CDTR0715
      360 C=-DEXP(C)                                                    CDTR0720
          GO TO 390                                                     CDTR0725
      370 C=0.D0                                                        CDTR0730
          GO TO 390                                                     CDTR0735
      380 C=DEXP(C)                                                     CDTR0740
      390 C=A2+C                                                        CDTR0745
          T11=1.D0-C                                                    CDTR0750
          T1=SNGL(T11)                                                  CDTR0755
C                                                                       CDTR0760
C         SELECT PROPER EXPRESSION FOR P                                CDTR0765
C                                                                       CDTR0770
      400 IF(G-2.) 420,410,410                                          CDTR0775
      410 IF(G-4.) 450,460,460                                          CDTR0780
C                                                                       CDTR0785
C         COMPUTE P FOR G GREATER THAN ZERO AND LESS THAN 2.0           CDTR0790
C                                                                       CDTR0795
      420 CALL DLGAM(THP1,GTH,IOK)                                      CDTR0800
          DT2=THETA*DLXX-X2-THP1*.6931471805599453 -GTH                 CDTR0805
          IF(DT2+1.68D02) 430,430,440                                   CDTR0810
      430 P=T1                                                          CDTR0815
          GO TO 490                                                     CDTR0820
      440 DT2=DEXP(DT2)                                                 CDTR0825
          T2=SNGL(DT2)                                                  CDTR0830
          P=T1+T2+T2                                                    CDTR0835
          GO TO 490                                                     CDTR0840
C                                                                       CDTR0845
C         COMPUTE P FOR G GREATER THAN OR EQUAL TO 2.0                  CDTR0850
C         AND LESS THAN 4.0                                             CDTR0855
C                                                                       CDTR0860
      450 P=T1                                                          CDTR0865
          GO TO 490                                                     CDTR0870
```

```
FORTRAN IV G LEVEL  21                  CADTR                DATE = 78094            10/03/40

               C                                                                      CDTR0875
               C         COMPUTE P FOR G GREATER THAN OR EQUAL TO 4.0                 CDTR0880
               C         AND LESS THAN OR EQUAL TO 1000.0                             CDTR0885
               C                                                                      CDTR0890
 0107            460 DT3=0.D0                                                         CDTR0895
 0108                DO 480 I3=2,K                                                    CDTR0900
 0109                THPI=DFLOAT(I3)+THETA                                            CDTR0905
 0110                CALL DLGAM(THPI,GTH,IOK)                                         CDTR0910
 0111                DLT3=THPI*DLX2-DLXX-X2-GTH                                       CDTR0915
 0112                IF(DLT3+1.68D02) 480,480,470                                     CDTR0920
 0113            470 DT3=DT3+DEXP(DLT3)                                               CDTR0925
 0114            480 CONTINUE                                                         CDTR0930
 0115                T3=SNGL(DT3)                                                     CDTR0935
 0116                P=T1-T3-T3                                                       CDTR0940
               C                                                                      CDTR0945
               C         SET ERROR INDICATOR                                          CDTR0950
               C                                                                      CDTR0955
 0117            490 IF(P) 500,520,520                                                CDTR0960
 0118            500 IF(ABS(P)-1.E-7) 510,510,600                                     CDTR0965
 0119            510 P=0.0                                                            CDTR0970
 0120                GO TO 610                                                        CDTR0975
 0121            520 IF(1.-P) 530,550,550                                             CDTR0980
 0122            530 IF(ABS(1.-P)-1.E-7) 540,540,600                                  CDTR0985
 0123            540 P=1.0                                                            CDTR0990
 0124                GO TO 610                                                        CDTR0995
 0125            550 IF(P-1.E-8) 560,560,570                                          CDTR1000
 0126            560 P=0.0                                                            CDTR1005
 0127                GO TO 610                                                        CDTR1010
 0128            570 IF((1.0-P)-1.E-8) 580,580,610                                    CDTR1015
 0129            580 P=1.0                                                            CDTR1020
 0130                GO TO 610                                                        CDTR1025
 0131            590 IER=-1                                                           CDTR1030
 0132                D=-1.E75                                                         CDTR1035
 0133                P=-1.E75                                                         CDTR1040
 0134                GO TO 620                                                        CDTR1045
 0135            600 IER=+1                                                           CDTR1050
 0136                P= 1.E75                                                         CDTR1055
 0137                GO TO 620                                                        CDTR1060
 0138            610 IER=0                                                            CDTR1065
 0139            620 CADTR=1.0-P                                                      CDTR1070
 0140                IF(IER.EQ.1)PRINT 910                                            CDTR1075
 0141            910 FORMAT('0',10X,'FAILURE TO CONVERGE IN X-SQ FUNCTION')           CDTR1080
 0142                IF(IER.EQ.-1)PRINT 911                                           CDTR1085
 0143            911 FORMAT('0',10X,'INVALID INPUT TO X-SQ FUNCTION')                 CDTR1090
 0144                RETURN                                                           CDTR1095
 0145                END                                                              CDTR1100
```

```
FORTRAN IV G LEVEL  21                NDTR                DATE = 78094          10/03/40

 0001          SUBROUTINE NDTR(X,P,D)                                                    CDTR1105
 0002          AX=ABS(X)                                                                 CDTR1110
 0003          T=1.0/(1.0+0.2316419*AX)                                                  CDTR1115
 0004          D=0.3989423*EXP(-X*X/2.0)                                                 CDTR1120
 0005          P=1.0-D*T*((((1.330274*T-1.821256)*T+1.781478)*T-0.3565638)*T+            CDTR1125
              $0.3193815)                                                                CDTR1130
 0006          IF(X.LT.0.0) P=1.0-P                                                      CDTR1135
 0007          RETURN                                                                    CDTR1140
 0008          END                                                                       CDTR1145
```

THE SAMPLE SIZE EQUALS 50

LAMDA1 EQUALS 0.1000E 02

LAMDA2 EQUALS 0.1250E 02

THE SEED VALUES FOR THIS RUN ARE

ISEED EQUALS 62155

JSEED EQUALS 51562

THE VALUE OF THE INDEX OF NON-CONGRUITY(DELTA) EQUALS 0.1638

THE VALUE OF THE PSEUDO CHI-SQUARE STATISTIC EQUALS 2.00

THE VALUE OF THE CHI-SQUARE STATISTIC EQUALS 12.40

THE AREA OF THE CHI-SQUARE DISTRIBUTION TO THE RIGHT OF THE CHI-SQUARE STATISTIC (ALPHA HAT) EQUALS 0.1917

REGION BOUNORIES	0.0105	0.0223	0.0357	0.0511	0.0693	0.0916	0.1204	0.1609	0.2303	0.0
CUMULATIVE PROB	0.1234	0.2434	0.3597	0.4719	0.5796	0.6819	0.7780	0.8663	0.9438	0.0
RANDOM SAMPLE	0.5302	0.0023	0.4798	0.3014	0.0083	0.0028	0.1542	0.0115	0.0736	0.1119
RANDOM SAMPLE	0.0837	0.3412	0.0602	0.0210	0.0250	0.0664	0.0686	0.2834	0.1306	0.0428
RANDOM SAMPLE	0.1768	0.0614	0.0193	0.0702	0.0275	0.0137	0.0085	0.1358	0.4015	0.0403
RANDOM SAMPLE	0.0212	0.0008	0.0145	0.3709	0.2129	0.0620	0.0581	0.0016	0.0265	0.1270
RANDOM SAMPLE	0.1136	0.0519	0.1323	0.1319	0.0296	0.0045	0.0603	0.0596	0.0573	0.0428
CELL FREQUENCY	7.0	6.0	4.0	3.0	10.0	3.0	2.0	6.0	2.0	7.0

APPENDIX 2

```
C     PROGRAM TO EVALUATE "THE INDEX OF NON-CONGRUITY" FOR NORMAL
C     DISTRIBUTIONS.  THIS INDEX IS BASED ON THE AMOUNT OF NON-OVERLAPPING
C     AREA BETWEEN TWO DISTRIBUTIONS.
C     VALUES TO BE SUPPLIED TO THE PROGRAM ARE DISTRIBUTION PARAMETERS,
C      SAMPLE SIZE, AND INITIAL VALUES FOR THE RNG
C     FORMAT FOR DATA CARDS:
C          MU1,SIGMA1,MU2,SIGMA2,M       1 CARD     (4E10.4,15)
C          ISEED,JSEED                   1 CARD     (2I5)
C     MULTIPLE RUNS ARE POSSIBLE BY SUPPLYING ADDITIONAL INPUT CARDS
C     (TWO PER REPLICATION).  PROGRAM COMPLETION IS INDICATED BY A
C     BLANK CARD.
      DIMENSION XI(10),P2(10),SAMPL(200),FREQ(10),Z2PS(10)
      REAL MU1,MU2,NU
C     INITIALIZE PROGRAM PARAMETERS
  300 X2SUM=0.
      F2SUM=0.
      DO 11 I=1,10
   11 FREQ(I)=0.
C     INPUT VALUES FOR THE PARAMETERS OF THE NORMAL DISTRIBUTIONS AND
C     THE SAMPLE SIZE
      READ 99,MU1,SIGMA1,MU2,SIGMA2,M
   99 FORMAT(4E10.4,15)
C     CHECK FOR PROGRAM COMPLETION
      IF(MU1.EQ.0..AND.SIGMA1.EQ.0.) GO TO 303
C     ECHO PARAMETER VALUES AND SAMPLE SIZE
      WRITE(6,98) MU1,SIGMA1,MU2,SIGMA2,M
   98 FORMAT(1H1,///,11X,'MEAN AND STANDARD DEVIATION OF DISTRIBUTION 1'
     1,2E12.4,//,11X,'MEAN AND STANDARD DEVIATION OF DISTRIBUTION 2',2E1
     22.4,//,11X,'THE SAMPLE SIZE EQUALS',I8)
C     DETERMINE IF DISTRIBUTIONS ARE OVERLAPPING
      IF(MU2.LT.MU1) GO TO 21
      IF(MU1+5.*SIGMA1.LT.MU2-5.*SIGMA2) GO TO 100
      GO TO 22
   21 IF(MU2+5.*SIGMA2.LT.MU1-5.*SIGMA1) GO TO 100
   22 CONTINUE
C     DETERMINE THE NUMBER OF INTERSECTION POINTS USING THE QUADRATIC
C     EQUATION
      IF(SIGMA1.EQ.SIGMA2) GO TO 1010
      VAR2=SIGMA2*SIGMA2
      VAR1=SIGMA1*SIGMA1
      RATIO=SIGMA1/SIGMA2
      SQMU1=MU1*MU1
      SQMU2=MU2*MU2
      VXM21=VAR2*MU1
      VXM12=VAR1*MU2
      VXMDIF=VXM21-VXM12
      VXMD2=VXMDIF*VXMDIF
      FRSFAC=4.*VXMD2
      VARDIF=VAR2-VAR1
      VXS21=VAR2*SQMU1
      VXS12=VAR1*SQMU2
      XLNFAC=2.*VAR1*VAR2*ALOG(RATIO)
      SNDFAC=4.*VARDIF*(VXS21-VXS12+XLNFAC)
      RAD=FRSFAC-SNDFAC
      IF(RAD) 100,101,102
C     THE DISTRIBUTIONS ARE NON-OVERLAPPING. THERE ARE NO INTERSECTION
C     POINTS
  100 WRITE(6,900)
```

```
      900 FORMAT(//,11X,'DISTRIBUTIONS ARE NON-OVERLAPPING. DELTA EQUALS 2')
          READ 97,ISEED,JSEED
          GO TO 300
C         THERE IS ONE INTERSECTION POINT
     1010 XX=(MU1+MU2)/2.
          GO TO 1011
      101 XX=(VXM21-VXM12)/VARDIF
     1011 Z1=(XX-MU1)/SIGMA1
          Z2=(XX-MU2)/SIGMA2
          CALL NDTR(Z1,F1,D)
          CALL NDTR(Z2,F2,D)
          DELTA=2.*ABS(F1-F2)
          IND=1
          GO TO 103
C         THERE ARE TWO INTERSECTION POINTS
      102 XX=(VXM21-VXM12)/VARDIF
          RADPRT=SQRT(RAD)/(2.*VARDIF)
          XX1=XX+RADPRT
          XX2=XX-RADPRT
          IF(XX1.LE.XX2) GO TO 1021
          C1=XX2
          C2=XX1
          GO TO 1022
     1021 C1=XX1
          C2=XX2
     1022 Z1C1=(C1-MU1)/SIGMA1
          Z2C1=(C1-MU2)/SIGMA2
          Z1C2=(C2-MU1)/SIGMA1
          Z2C2=(C2-MU2)/SIGMA2
          CALL NDTR(Z1C1,F1C1,D)
          CALL NDTR(Z1C2,F1C2,D)
          CALL NDTR(Z2C1,F2C1,D)
          CALL NDTR(Z2C2,F2C2,D)
          F2DIF=F2C2-F2C1
          F1DIF=F1C2-F1C1
          DELTA=2.*ABS(F2DIF-F1DIF)
          IND=2
C         DETERMINE VALUES FOR CLASS BOUNDRIES FOR EQUAL PROBABILITY REGIONS
C         OF MODEL DISTRIBUTION (DISTRIBUTION 1)
      103 DO 1 I=1,9
          P=.1*I
          CALL NDTRI(P,Z,D,IER)
        1 X1(I)=MU1+SIGMA1*Z
C         DETERMINE CUMULATIVE PROBABILITIES FOR CLASS BOUNDRIES FOR ALTERNATIVE
C         DISTRIBUTION (DISTRIBUTION 2)
          DO 2 I=1,9
          Z2PS(I)=(X1(I)-MU2)/SIGMA2
        2 CALL NDTR(Z2PS(I),P2(I),D)
C         CALCULATE THE EXPECTED VALUE FOR CELL FREQUENCIES
          EI=M/10.
C         CALCULATE THE PSEUDO CHI-SQUARE STATISTIC
          X2SUM=(EI-M*P2(1))**2
          DO 3 I=2,9
        3 X2SUM=X2SUM+(EI-M*(P2(I)-P2(I-1)))**2
          X2PS=X2SUM+(EI-M*(1.-P2(9)))**2
          X2PS=X2PS/EI
C         READ RANDOM NUMBER GENERATOR SEED VALUES
          READ(5,97)ISEED,JSEED
```

```
      97 FORMAT(2I5)
C        ECHO SEEO VALUES
         WRITE(6,96)ISEEO,JSEEO
      96 FORMAT(/////,11X,'THE SEEO VALUES FOR THIS RUN ARE',///,21X,'ISEED
        1 EQUALS',I12,//,21X,'JSEEO EQUALS',I12)
C        INITIALIZE RANOOM NUMBER GENERATOR
         CALL RSTART(ISEEO,JSEEO)
C        GENERATE RANOOM SAMPLE FROM SECONO OISTRIBUTION
         OO 4 I=1,M
       4 SAMPL(I)=MU2+SIGMA2*RNOR(1)
C        SORT RANOOM OBSERVATIONS INTO FREQUENCY CLASSES
         OO 5 I=1,M
         IF(SAMPL(I).LE.X1(5)) GO TO 201
         IF(SAMPL(I).LE.X1(6)) GO TO 206
         IF(SAMPL(I).LE.X1(7)) GO TO 207
         IF(SAMPL(I).LE.X1(8)) GO TO 208
         IF(SAMPL(I).LE.XI(9)) GO TO 209
         K=10
         GO TO 5
     201 IF(SAMPL(I).GT.X1(4)) GO TO 205
         IF(SAMPL(I).GT.X1(3)) GO TO 204
         IF(SAMPL(I).GT.X1(2)) GO TO 203
         IF(SAMPL(I).GT.XI(1)) GO TO 202
         K=1
         GO TO 5
     202 K=2
         GO TO 5
     203 K=3
         GO TO 5
     204 K=4
         GO TO 5
     205 K=5
         GO TO 5
     206 K=6
         GO TO 5
     207 K=7
         GO TO 5
     208 K=8
         GO TO 5
     209 K=9
       5 FREQ(K)=FREQ(K)+1.
C        CALCULATE THE ACTUAL CHI-SQUARE STATISTIC FOR RANDOM SAMPLE
         OO 6 I=1,10
       6 F2SUM=F2SUM+FREQ(I)**2
         X2ACT=F2SUM/EI-M
         NU=9.
C        CALCULATE "ALPHA HAT" FOR THE COMPUTEO CHI-SQUARE VALUE
         AHAT=CAOTR(X2ACT,NU)
C        OUTPUT VALUES OF OELTA, PSEUOO CHI-SQUARE, CHI-SQUARE, AND ALPHA HAT
         WRITE(6,95) OELTA,X2PS,X2ACT,AHAT
      95 FORMAT(/////,11X,'THE VALUE OF THE INOEX CF NON-CONGRUITY(OELTA) E
        1QUALS',F12.4,//,11X,'THE VALUE OF THE PSEUOO CHI-SQUARE STATISTIC
        2EQUALS',F12.2,//,11X,'THE VALUE OF THE CHI-SQUARE STATISTIC EQUALS
        3',F12.2,//,11X,'THE AREA OF THE CHI-SQUARE OISTRIBUTICN TO THE RIG
        4HT OF THE CHI-SQUARE STATISTIC (ALPHA HAT) EQUALS',F12.4)
C        OUTPUT VALUES OF INTERSECTICN POINT OR PCINTS
         WRITE(6,94)
         IF(INO.EQ.1) GO TO 301
```

```
      WRITE(6,920) C1,C2
  920 FORMAT(1X,'THERE ARE TWO POINTS OF INTERSECTION. THEY ARE',F10.2,5
     1X,'AND',F10.2)
      GO TO 302
  301 WRITE(6,910) XX
  910 FORMAT(1X,'THERE IS ONE POINT OF INTERSECTION. IT IS',F10.2)
C     OUTPUT VALUES OF X1, P2, SAMPL, AND FREQ
  302 WRITE(6,94)
      WRITE(6,92)(X1(I),I=1,9)
      WRITE(6,91)(P2(I),I=1,9)
      WRITE(6,93)(SAMPL(I),I=1,M)
      WRITE(6,90)(FREQ(I),I=1,10)
   94 FORMAT(/////)
   93 FORMAT('0','RANDOM SAMPLE',10F11.4)
   92 FORMAT('0','REGION BOUNDRIES',9F11.4)
   91 FORMAT('0','CUMULATIVE PROB',9F11.4)
   90 FORMAT('0','CELL FREQUENCY',10F11.1)
      GO TO 300
  303 STOP
      END
```

```
FORTRAN IV G LEVEL  21                NDTRI                DATE = 78094           10/00/33

      SUBROUTINE NDTRI(P,X,D,IE)
      IE=0
      X=.99999E+74
      D=X
      IF(P)1,4,2
    1 IE=-1
      GO TO 12
    2 IF(P-1.0) 7,5,1
    4 X=-.99999E+74
    5 D=0.0
      GO TO 12
    7 D=P
      IF(D-0.5) 9,9,8
    8 D=1.0-D
    9 T2=ALOG(1.0/(D*D))
      T=SQRT(T2)
      X=T-(2.515517+0.802853*T+0.010328*T2)/(1.0+1.432788*T+0.189269*T2+
     10.001308*T*T2)
      IF(P-0.5) 10,10,11
   10 X=-X
   11 D=0.3989423*EXP(-X*X/2.0)
   12 RETURN
      END
```

FORTRAN IV G LEVEL 21 CAOTR DATE = 78094 10/03/40

```
0001           FUNCTION CAOTR(X,G)
         C
         C        PURPOSE
         C           COMPUTES P(X) = PROBABILITY THAT THE RANCOM VARIABLE U,
         C           OISTRIBUTED ACCOROING TO THE CHI-SQUARE DISTRIBUTION WITH G
         C           DEGREES OF FREEDOM, IS LESS THAN OR ECUAL TO X.  F(G,X), THE
         C        USAGE
         C          PROB=COTR(X,G)
         C
         C        OESCRIPTION OF PARAMETERS
         C           X    - INPUT SACLE FOR WHICH P(X) IS COMPUTEO.
         C           G    - NUMBER OF DEGREES OF FREEDOM OF THE CHI-SQUARE
         C                  OISTRIBUTION.  G IS A CONTINUOUS PARAMETER.
         C           IER - RESULTANT ERRCR CODE WHERE
         C                IER= 0 --- NO ERRCR
         C                IER=-1 --- AN INPUT PARAMETER IS INVALIO.  X IS LESS
         C                           THAN 0.0, OR G IS LESS THEN 0.5 CR GREATER
         C                           THAN 2*10**(+5).  P ANO O ARE SET TO -1.E75.
         C                IER=+1 --- INVALIO OUTPUT.  P IS LESS THAN ZERO OR
         C                           GREATER THAN ONE, OR SERIES FOR T1 (SEE
         C                           MATHEMATICAL DESCRIPTION) HAS FAILEO TO
         C                           CONVERGE.  P IS SET TO 1.E75.
         C
         C        SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIREO
         C           OLGAM
         C           NOTR
         C
0002           DOUBLE PRECISION XX,OLXX,X2,OLX2,GG,G2,OLT3,THETA,THP1,
              1T11,SER,CC,XI,FAC,TLOG,TERM,GTH,A2,A,B,C,DT2,DT3,THPI
         C
         C        TEST FOR VALIO INPUT DATA
         C
0003           IF(G-(.5-1.E-5)) 590,10,10
0004        10 IF(G-2.E+5) 20,20,590
0005        20 IF(X) 590,30,30
         C
         C        TEST FOR X NEAR 0.0
         C
0006        30 IF(X-1.E-8) 40,40,80
0007        40 P=0.0
               IF(G-2.) 50,60,70
0008        50 D=1.E75
0009           GO TO 610
0010        60 D=0.5
0011           GO TO 610
0012        70 D=0.0
0013           GO TO 610
0014
         C
         C        TEST FOR X GREATER THAN 1.E+6
         C
0015        80 IF(X-1.E+6) 100,100,90
0016        90 D=0.0
0017           P=1.0
0018           GO TO 610
         C
         C        SET PROGRAM PARAMETERS
         C
0019       100 XX=DBLE(X)
```

```
      DLXX=DLOG(XX)
      X2=XX/2.D0
      DLX2=DLOG(X2)
      GG=DBLE(G)
      G2=GG/2.D0
C
C        TEST FOR G GREATER THAN 1000.0
C        TEST FOR X GREATER THAN 2000.0
C
      IF(G-1000.) 160,160,180
  160 IF(X-2000.) 190,190,170
  170 P=1.0
      GO TO 610
  180 A=DLOG(XX/GG)/3.D0
      A=DEXP(A)
      B=2.D0/(9.D0*GG)
      C=(A-1.D0+B)/DSQRT(B)
      SC=SNGL(C)
      CALL NDTR(SC,P,DUMMY)
      GO TO 490
C
C        COMPUTE THETA
C
  190 K= IDINT(G2)
      THETA=G2-DFLOAT(K)
      IF(THETA-1.0-8) 200,200,210
  200 THETA=0.D0
  210 THP1=THETA+1.D0
C
C        SELECT METHOD OF COMPUTING T1
C
      IF(THETA)230,230,220
  220 IF(XX-10.D0)260,260,320
COMPUTE T1 FOR THETA EQUALS 0.0
  230 IF(X2-1.68D02) 250,240,240
  240 T1=1.0
      GO TO 400
  250 T11=1.D0-DEXP(-X2)
      T1=SNGL(T11)
      GO TO 400
C
C        COMPUTE T1 FOR THETA GREATER THAN 0.0 AND
C        X LESS THAN OR EQUAL TO 10.0
C
  260 SER=X2*(1.D0/THP1 -X2/(THP1+1.D0))
      J=+1
      CC=DFLOAT(J)
      DO 270 IT1=3,30
      XI=DFLOAT(IT1)
      CALL DLGAM(XI,FAC,IOK)
      TLOG= XI*DLX2-FAC-DLOG(XI+THETA)
      TERM=DEXP(TLOG)
      TERM=DSIGN(TERM,CC)
      SER=SER+TERM
      CC=-CC
      IF(DABS(TERM)-1.D-9) 280,270,270
  270 CONTINUE
      GO TO 600
```

```
C
C        COMPUTE P FOR G GREATER THAN OR EQUAL TO 4.0
C        AND LESS THAN OR EQUAL TO 1000.0
C
  460 DT3=0.00
      DO 480 I3=2,K
      THPI=DFLOAT(I3)+THETA
      CALL DLGAM(THPI,GTH,IOK)
      DLT3=THPI*DLX2-DLXX-X2-GTH
      IF(DLT3+1.68D02) 480,480,470
  470 DT3=DT3+DEXP(DLT3)
  480 CONTINUE
      T3=SNGL(DT3)
      P=T1-T3-T3
C
C        SET ERROR INDICATOR
C
  490 IF(P) 500,520,520
  500 IF(ABS(P)-1.E-7) 510,510,600
  510 P=0.0
      GO TO 610
  520 IF(1.-P) 530,550,550
  530 IF(ABS(1.-P)-1.E-7) 540,540,600
  540 P=1.0
      GO TO 610
  550 IF(P-1.E-8) 560,560,570
  560 P=0.0
      GO TO 610
  570 IF((1.0-P)-1.E-8) 580,580,610
  580 P=1.0
      GO TO 610
  590 IER=-1
      D=-1.E75
      P=-1.E75
      GO TO 620
  600 IER=+1
      P= 1.E75
      GO TO 620
  610 IER=0
  620 CADTR=1.0-P
      IF(IER.EQ.1)PRINT 910
  910 FORMAT('0',10X,'FAILURE TO CONVERGE IN X-SQ FUNCTION')
      IF(IER.EQ.-1)PRINT 911
  911 FORMAT('0',10X,'INVALID INPUT TO X-SQ FUNCTION')
      RETURN
      END
```

```
FORTRAN IV G LEVEL  21              NDTR              DATE = 78094          10/03/40

 0001            SUBROUTINE NDTR(X,P,D)                                                 CDTR1105
 0002            AX=ABS(X)                                                              CDTR1110
 0003            T=1.0/(1.0+0.2316419*AX)                                               CDTR1115
 0004            D=0.3989423*EXP(-X*X/2.0)                                              CDTR1120
 0005            P=1.0-D*T*((((1.330274*T-1.821256)*T+1.781478)*T-0.3565638)*T+         CDTR1125
                $0.3193815)                                                             CDTR1130
 0006            IF(X.LT.0.0) P=1.0-P                                                   CDTR1135
 0007            RETURN                                                                 CDTR1140
 0008            END                                                                    CDTR1145
```

```
FORTRAN IV G LEVEL  21                 NDTR                 DATE = 78094          10/03/40

 0001          SUBROUTINE NDTR(X,P,D)                                                  CDTR1105
 0002          AX=ABS(X)                                                               CDTR1110
 0003          T=1.0/(1.0+0.2316419*AX)                                                CDTR1115
 0004          D=0.3989423*EXP(-X*X/2.0)                                               CDTR1120
 0005          P=1.0-D*T*((((1.330274*T-1.821256)*T+1.781478)*T-0.3565638)*T+          CDTR1125
              $0.3193815)                                                              CDTR1130
 0006          IF(X.LT.0.0) P=1.0-P                                                    CDTR1135
 0007          RETURN                                                                  CDTR1140
 0008          END                                                                     CDTR1145
```

MEAN AND STANDARD DEVIATION OF DISTRIBUTION 1 0.0 0.1000E 01

MEAN AND STANDARD DEVIATION OF DISTRIBUTION 2 0.3000E 00 0.1200E 01

THE SAMPLE SIZE EQUALS 50

THE SEED VALUES FOR THIS RUN ARE

ISEED EQUALS 62155

JSEED EQUALS 51562

THE VALUE OF THE INDEX OF NON-CONGRUITY(DELTA) EQUALS 0.2648

THE VALUE OF THE PSEUDO CHI-SQUARE STATISTIC EQUALS 7.16

THE VALUE CF THE CHI-SCUARE STATISTIC EQUALS 6.40

THE AREA OF THE CHI-SQUARE DISTRIBUTION TO THE RIGHT OF THE CHI-SQUARE STATISTIC (ALPHA HAT) EQUALS 0.6993

THERE ARE TWO POINTS OF INTERSECTION. THEY ARE -2.05 AND 0.68

REGION BOUNDRIES	-1.2817	-0.8415	-0.5240	-0.2529	0.0	0.2529	0.5240	0.8415	1.2817	
CUMULATIVE PROB	0.0937	0.1707	0.2461	0.3225	0.4013	0.4844	0.5740	0.6741	0.7934	
RANDOM SAMPLE	0.7770	4.3540	-1.4125	0.4722	-0.2033	-0.7779	-0.3549	1.0924	-0.5331	-0.3329
RANDOM SAMPLE	0.1506	-0.8829	-0.8850	-0.6316	1.1461	0.7400	-2.2633	0.8774	0.0849	-0.4979
RANDOM SAMPLE	0.2298	-0.0376	1.5552	0.3529	3.2375	3.7394	2.4928	0.6124	1.6426	0.5917
RANDOM SAMPLE	-1.0056	-0.7294	0.9289	-1.0840	-1.0789	0.0053	0.5922	0.9800	0.0062	0.1151
RANDOM SAMPLE	0.8669	0.7669	-0.7807	1.6423	-0.1953	1.1252	-2.8222	1.5932	-2.0086	-1.8013
CELL FREQUENCY	5.0	5.0	5.0	3.0	3.0	6.0	2.0	6.0	7.0	8.0

NON-CONGRUENCE OF STATISTICAL DISTRIBUTIONS:

HOW DIFFERENT IS DIFFERENT?

by

TERRY LEE APPLEGATE

B.S., Kansas State University, 1977

AN ABSTRACT OF A MASTER'S THESIS

Submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1978

This thesis studies differences between statistical distributions of the same family. In particular it studies members of the exponential and normal families of statistical distributions. In theory two distributions are the same only if their probability density functions are identical (which implies that their parameters are identical also). However, in practical situations, two distributions which have closely similar probability density functions may produce random samples of small size which are indistinguishable from one another. This thesis is concerned with studying this situation in an attempt to better understand the question of "How different is different?" in relation to differences in statistical distributions from the same family.

The methodology used to study the difference between a pair of statistical distributions from the same family consists of a number of steps. The first step in the comparison procedure consists of determining the amount of non-overlapping area bounded by the probability density functions of the two distributions being compared. The second step consists of drawing a "perfect" sample from one distribution and comparing it with the other distribution. The third step consists of drawing a random sample from one distribution and comparing it with the other distribution. The final step consists of calculating certain indices from the parameters of the distributions and relating these indices to the other comparison results.

Results of the comparison procedure for a sample size of 50 indicate that in both the exponential case and the normal case statistical significant differences at the .05 level would be indicated for amounts of non-overlapping area in excess of a threshhold value occurring somewhere in the region of .3 to .6. In addition to this there appears to be strong relationships between the indices derived from the parameters of the distributions being compared and the various other comparison indices.

These strong relationships would allow the comparison of statistical distributions solely on the basis of their parameters without requiring the use of sampling.

Special topics covered in the study which might be of interest to other researchers are the use of the McGill Random Number Generator developed by members of the School of Computer Science of McGill University and the suggestions for further research in this area.