ESTIMATING STATISTICALLY SIGNIFICANT DIFFERENCES

BETWEEN A PAIR OF BETA DISTRIBUTIONS

by

KRISHNAN LAKSHMINARAYAN

B. Tech (Mechanical Engineering)

Indian Institute of Technology, Madras, 1974

.

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements of the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1980

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

1.1 Objective

The beta distribution has been used frequently to fit sample data (1). It has been used, for example, as a prior distribution for a binomial proportion ((1) Chapter 8). For a given set of data, different methods can be used to estimate the beta parameters, usually yielding different estimates. In order to distinguish between a pair of beta distributions, it may be desirable to invent an index, or indices, that reveal the extent of these differences.

The objective of this investigation was to develop a measure (or measures) of separation between a pair of beta distributions, and to correlate these measures with some standard test of significance, such as the χ^2 goodness-of-fit test. An important sub-objective of this research was to develop a computer code that can be used to obtain these correlations for a wide variety of beta distributions.

1.2 Literature Search

Several coefficients have been suggested in statistical literature to reflect the fact that some probability distributions are "closer together" than others, and consequently that it may be "easier to distinguish" between the distributions of one pair than between those of another (2). Such coefficients have been variously called measures of distance between two distributions (3), measures

of separation (4), measures of discriminatory information (5,6) and measures of variation distance (7). These coefficients have the common property of increasing as the two distributions involved "move apart".

Rao (4) states the need for a distance function between two populations, and describes mathematical concepts involving the construction of such a function. Mahalanobis (9) defined a generalized distance to be used in comparing normal populations. Karl Pearson proposed a measure of racial likeness (10) which has been used by anthropologists for analyzing skeletal remains. He defined a coefficient of racial likeness (C.R.L) which is a measure of the distance between two populations. Adhikari and Joshi (3) give a complete summary of the various coefficients of distance developed by several investigators.

1.3 The Problem

The concept of significance has been developed for analyzing sample data and for making inferences about the populations being sampled. Various tests have been developed for comparing groups of sample data with each other or for comparing a sample with a population (or model). The question to be asked when comparing two populations is: How different is different?

The type of distribution considered here is the beta family described by the density function

$$f(x|a,b) = \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} \quad (0 \le x \le 1) \tag{1.1}$$

where

$$B(a,b) = \frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)} \tag{1.2}$$

and the gamma function $\Gamma(a)$ is defined as

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} \, dx \quad [0 < a < \infty]. \tag{1.3}$$

For convenience of terminology, such a density function (or a random variable thus distributed) is described as beta (a,b).

In general, by varying the parameter a and b of the beta distribution, a variety of curve shapes are obtained (Figure 1.1). For this reason, the beta distribution is often suitable in modeling work for random variables over the interval $(0,1)$.

Density A can be called significantly different from density B if a random sample from A does not pass a χ^2 goodness-of-fit test, when tested against B (or vice versa). The significance level of an observed random sample (or perhaps the average of many) can be noted and compared with some index which depends only on the parameters of the curves. However, the construction of random samples presents some difficulties:

(i) Beta random samples are expensive to generate, when concentrated over a small range in $(0,1)$

(ii) True random samples would require many points to ascertain the relationship between any two indices of non-equality, because there would be the random fluctuation effect which would obscure the basic relationship.

For this reason, emphasis is placed in this study on "pseudo-samples" where the "observed" frequency for a class is taken to be the same as the "expected" frequency for that class for the particular distribution under consideration. The chi-squared value obtained from testing density A against density B under these conditions is referred to as a "pseudo-chi-squared" statistic, and has the following properties:

(a) It is zero when density A and density B are the same

(b) It has no random component and is purely a measure of lack of fit, since the "sample" is a perfect replica of its generating density function.

The objective, as explained in more detail below, is to

i) Compute a pseudo-chi-square test statistic for a sample from density A tested against density B. Here B is defined as the "model" and A is defined as the "alternative".

ii) Compute an "index of non-congruity", denoted by δ, and relate this to the pseudo chi-square value.

iii) Compute certain other indices that are easily determined from the parameters of both distributions.

iv) Generate a random sample from A and test against B using the chi-square goodness-of-fit test.

v) Lastly, study the variation of the chi-square value obtained in iv) with the indices obtained in iii).

1.4 Analysis

Theoretically, two distributions are considered equal only if their density functions are identical. But in practice minor variations in the parameters may be of no consequence because two "nearby" densities may yield samples that are indistinguishable from each other. For example, it may be impossible to detect any difference in a random sample from a beta (3,3) density and one from a beta (3.2, 3.5) density. If it were possible to find the sampling distribution of the statistics $\hat{a}$ and $\hat{b}$ (the estimates of a and b), then the two distributions could be considered indistinguishable if it were possible, using the computed $\hat{a}_i$, $\hat{b}_i$ (i = 1,2), to accept the null hypothesis:

$$H_0: \quad a_1 = a_2$$

$$b_1 = b_2$$

against the alternative:

$$H_1: \quad a_1 \neq a_2$$

$$b_1 \neq b_2$$

In the absence of knowledge of the distribution of $\hat{a}$ and $\hat{b}$, other techniques must be employed.

An alternative approach would be to first construct "expected" frequencies for each class interval for each distribution. These are called "perfect" samples. One distribution is then referred to as the "model", and can be tested against the "alternative", using a chi-squared goodness of fit. The chi-square value thus obtained will be a "pseudo" chi-square value whenever we construct "perfect" samples from each distribution. If a genuine random sample were drawn from the "alternative" and compared with the "model" using the chi-squared test statistic, the chi-square value thus obtained would be the usual chi-squared test statistic. Discussion of the computation of the chi-squared value is given in Chapter II.

1.4.1 <u>Measures of Separation</u>

Let beta (a_1, b_1) and beta (a_2, b_2) be the beta distributions under consideration (Figures 3.1 (i) and (ii)). The "index of non-congruity", is defined as

$$\delta = \int_0^1 |f(x|a_1, b_1) - f(x|a_2, b_2)| \, dx \qquad (1.4)$$

where

$$0 \leq \delta \leq 2.$$

If the two curves are identical, then $\delta = 0$. However, if the two curves are very different, δ may be near the maximum possible value of 2.

Various other indices of non-equality are also computed in this study. These are defined as

$$\eta_1 = |\mu_1 - \mu_2| + |\sigma_1 - \sigma_2| \tag{1.5}$$

$$\eta_2 = \left| \frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2} \right| \tag{1.6}$$

where μ_i and σ_i are, respectively, the mean and standard deviation of the ith distribution, $i = 1,2$.

Results have been obtained using beta $(4,4)$ as the model and beta $(3,b_2)$ as the alternative, where b_2 has been varied in the interval $[2,4]$ in steps of 0.1. The pseudo sample size used is 100. Relationships have been obtained between η_1 and δ, δ and χ^2, η_1 and χ^2, and η_2 and χ^2. Lastly, random samples have been drawn from the alternative beta and the χ^2 value thus obtained (χ_s^2) has been plotted against η_1 and η_2.
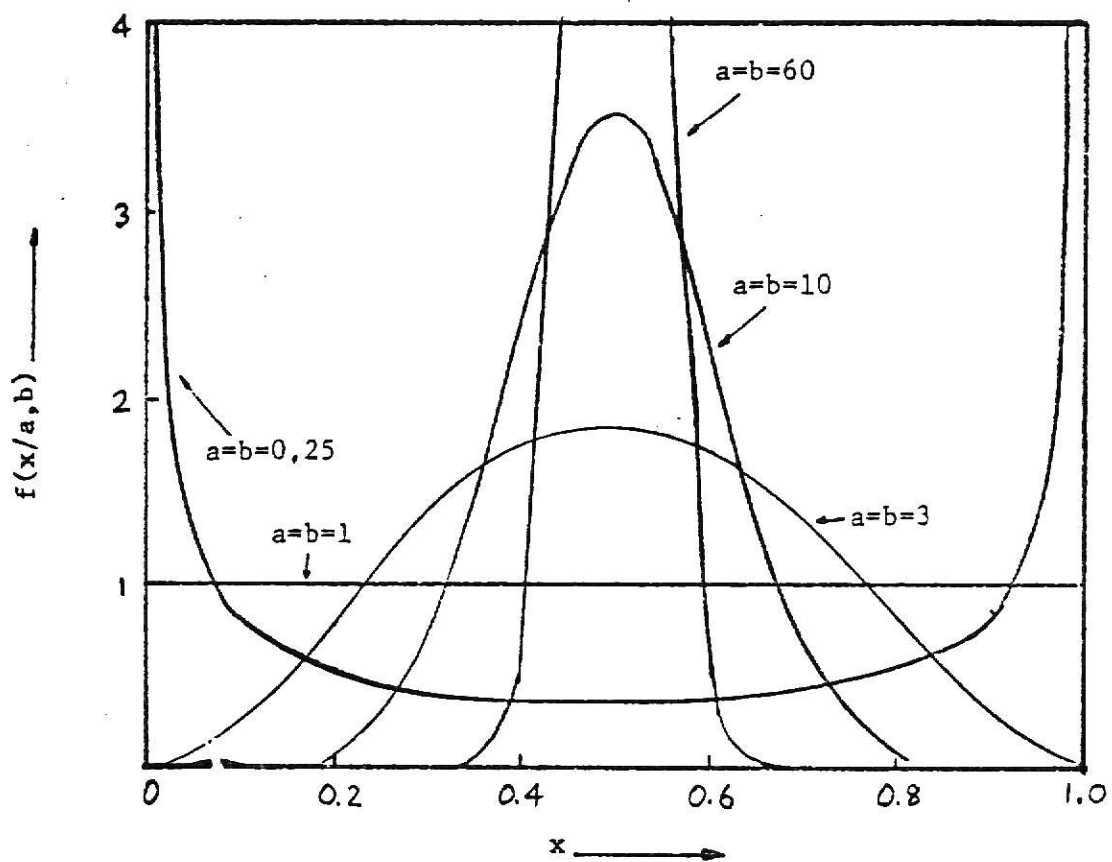
Fig. 1.1. Graphs of beta densities for various values of the parameters a and b. (From (11)).

2.0 THE χ^2 TEXT STATISTIC

2.1 Construction of the Test

Let $f(x|a_1,b_1)$ represent the density function for the model and $f(x|a_2,b_2)$ represent the density function for the alternative distribution. The first step in computing the chi-squared value is to obtain a frequency distribution for both the model and the alternative.

The ordinate of the density function at a point does not represent the probability at that point. For this reason, a finite number of classes are constructed over the interval $[0,1]$. The probability element for each class interval is normally computed using the equation

$$F = \int_{x_1}^{x_2} f(x|a,b) \, dx \qquad x_1,x_2 \varepsilon [0,1] \qquad (2.1)$$

where x_1 and x_2 are the lower and upper limits, respectively, of the class interval. Since it is not possible in general to analytically evaluate the integral in Eq. (2.1), an approximation is made by evaluating the ordinate at the midpoint of the class interval and multiplying by the class width. This probability element has the physical meaning of relative frequency, since it denotes the probability of an event occurring in that class interval. In the case of the alternative curve, these probability elements are "pseudo" relative frequencies, since they are obtained from an exact model and not from some random process occurring in nature.

In the computer program in Appendix I the probability element is designated as PRBEL. Using the subscript 1 for the model and 2 for the alternative, let N be the number of class intervals (cells) chosen. Then the pseudo chi-square statistic is defined for pseudo-sample size M as

$$\chi^2_{ps} = M \sum_{i=1}^{N} \frac{[(PRBEL)_{1i} - (PRBEL)_{2i}]^2}{(PRBEL)_{1i}} \tag{2.2}$$

with degrees of freedom DF = N-1.

The following example illustrates the construction of the chi-square test statistic.

EXAMPLE Let beta (4,4) be the model and beta (3,3) be the alternative. A sample size M = 50 is assumed, and N = 10. Thus

$$f(x|4,4) = \frac{1}{B(4,4)} x^3 (1-x)^3 \quad [0 \le x \le 1] \tag{2.3}$$

$$f(x|3,3) = \frac{1}{B(3,3)} x^2 (1-x)^2 \quad [0 \le x \le 1] \tag{2.4}$$

The "observed" frequencies, designated OF, and the "expected" frequencies, designated EF, were obtained using the program developed for this investigation (Appendix 1). These results are illustrated in Table 2.1, which also illustrates the probability elements for each class interval. If a sample of M observations is assumed, then each probability element must be multiplied by M to obtain the "frequency count" for that particular cell. In the computer program, OF and EF were designated as PRBEL(L,I,M), (for the Lth curve, I^{th} class, sample size M). Thus

$$PRBE(L,I,M) = PRBEL(L,I) *M \tag{2.5}$$

where

L = 1, (model)   I = 1, ..., N

= 2 (alternative)

The OF and EF values in Table 2.2 correspond to the PRBE values computed above.

In order to compute the pseudo-chi square test statistic a classical (though perhaps overly strict) requirement is that the expected frequency count in each interval be at least 5. This is ensured by grouping adjacent intervals whenever necessary. This reduces the effective number of intervals. In the above problem classes 1, 2 and 3 are grouped together to form one interval and intervals 8, 9, and 10 are grouped together to form another, the others remaining unaltered. The effective number of intervals is then reduced to 10-6+2 = 6. The pseudo chi-square statistic then becomes

$$\chi^2_{ps} = M \sum_{I=1}^{6} \frac{(PRBEL(1,I) - PRBEL(2,I))^2}{PRBEL(1,I)} \qquad (2.6)$$

and carrying out the computations yields:

$$\chi^2_{ps} = 1.474 \text{ with df} = 6-1 = 5.$$

As explained below, the next step is to calculate the significance level $\hat{\alpha}$ of the test.

The usual χ^2 table gives the critical point for a given α. But it is desired to know what significance level, called $\hat{\alpha}$, is associated with an observed χ^2 value. In the present work a function subprogram CADTR (12) is employed for this purpose. The input parameters of the subroutine are the known chi-square value and the degrees of freedom, both of which are treated as real continuous parameters (abbreviated as χ^2 and DF, respectively, for convenience).

The resultant $\hat{\alpha}$ value calculated by subprogram CADTR is $\hat{\alpha} = 0.9161$, for $\chi^2_{ps} = 1.474$ with df = 5.

CLASS MARK	$(PRBEL)_{2i}$	$(PRBEL)_{1i}$
.05	.0068	.0015
.15	.0488	.0290
.25	.1055	.0923
.35	.1553	.1648
.45	.1838	.2123
.55	.1838	.2123
.65	.1553	.1648
.75	.1055	.0923
.85	.0488	.0290
.95	.0068	.0015
TOTAL	1.0	1.0

TABLE 2.1 Probability elements for model beta (4,4) and alternative beta (3,3).
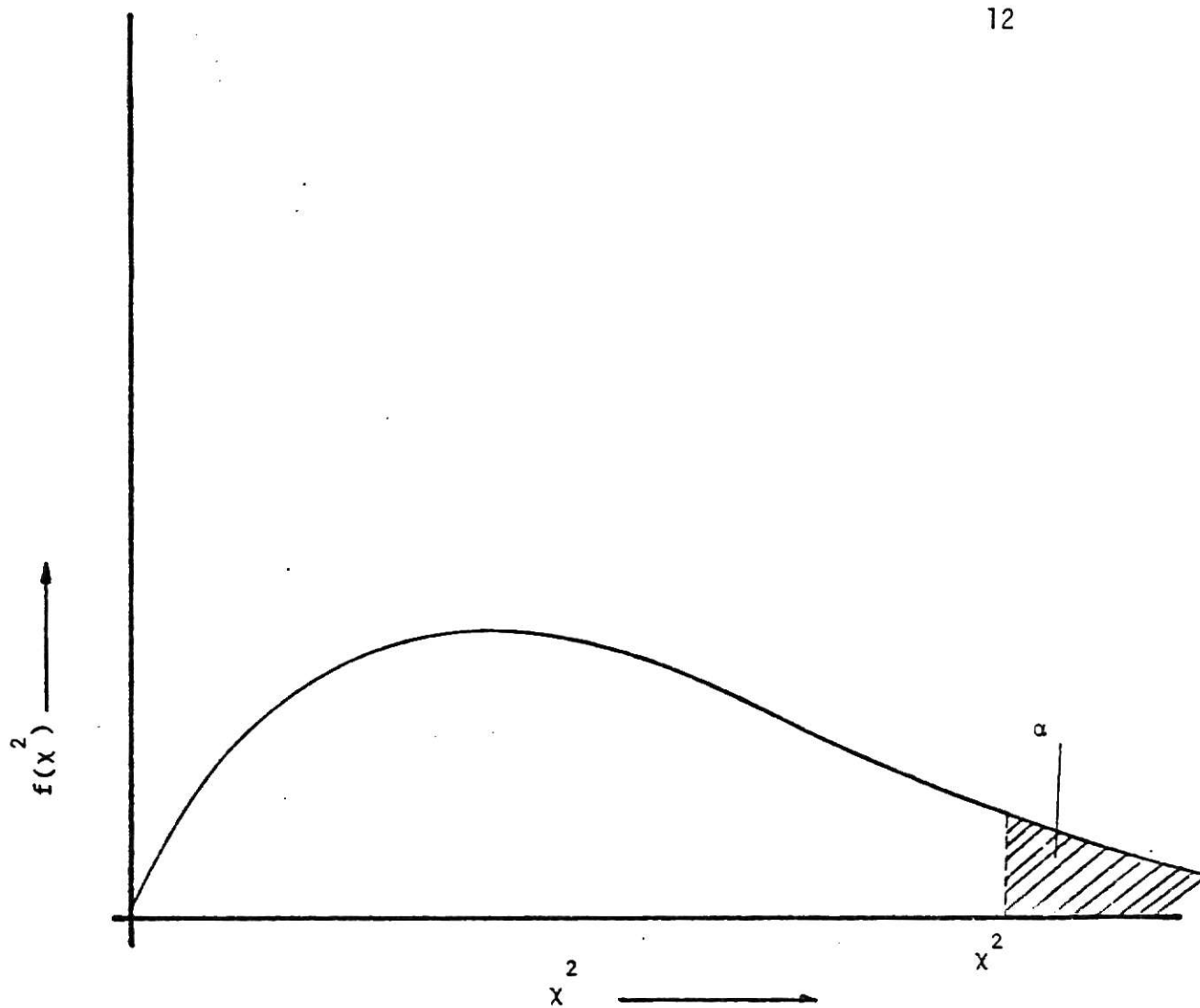
Fig. 2.1. χ^2 distribution showing right hand tail area for a given χ^2 value.

3.0 <u>DETERMINATION OF THE INDEX OF NON-CONGRUITY</u>

3.1 <u>Introduction</u>

The index of non-congruity, δ, is calculated in the following manner. As before, let $f(x|a_1,b_1)$ and $f(x|a_2,b_2)$ be two beta distributions described in the manner of Figure 3.1.

It can be proved that there will be at least one point of intersection for the two distributions but never more than two. Let the points of intersection be denoted by

$$x = x_i \qquad i = 1, 2 \qquad\qquad (3.1)$$

At each such x

$$f(x|a_1,b_1) = f(x|a_2,b_2) \qquad\qquad (3.2)$$

Thus,

$$\frac{1}{B(a_1,b_1)} x^{a_1-1}(1-x)^{b_1-1} = \frac{1}{B(a_2,b_2)} x^{a_2-1}(1-x)^{b_2-1}$$

hence

$$x^{a_1-a_2}(1-x)^{b_1-b_2} = \frac{B(a_1,b_1)}{B(a_2,b_2)}$$

$$= \frac{\Gamma(a_1)\,\Gamma(b_1)}{\Gamma(a_1+b_1)} \cdot \frac{\Gamma(a_2+b_2)}{\Gamma(a_2)\,\Gamma(b_2)} = C \qquad\qquad (3.3)$$

The right hand side of equation (3.3) is easily evaluated by subroutine GMMA, documented in [12], which evaluates the gamma function for real arguments.

In order to determine the points of intersection, it is necessary to determine the zeros of the equation

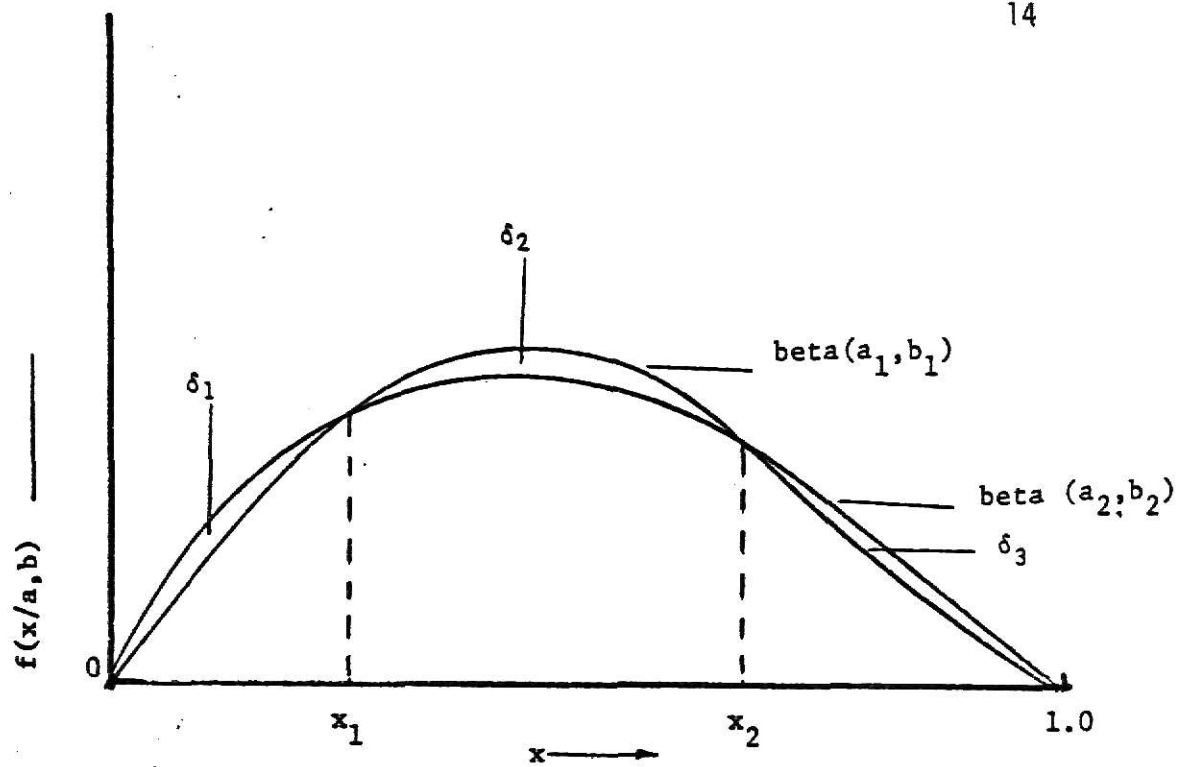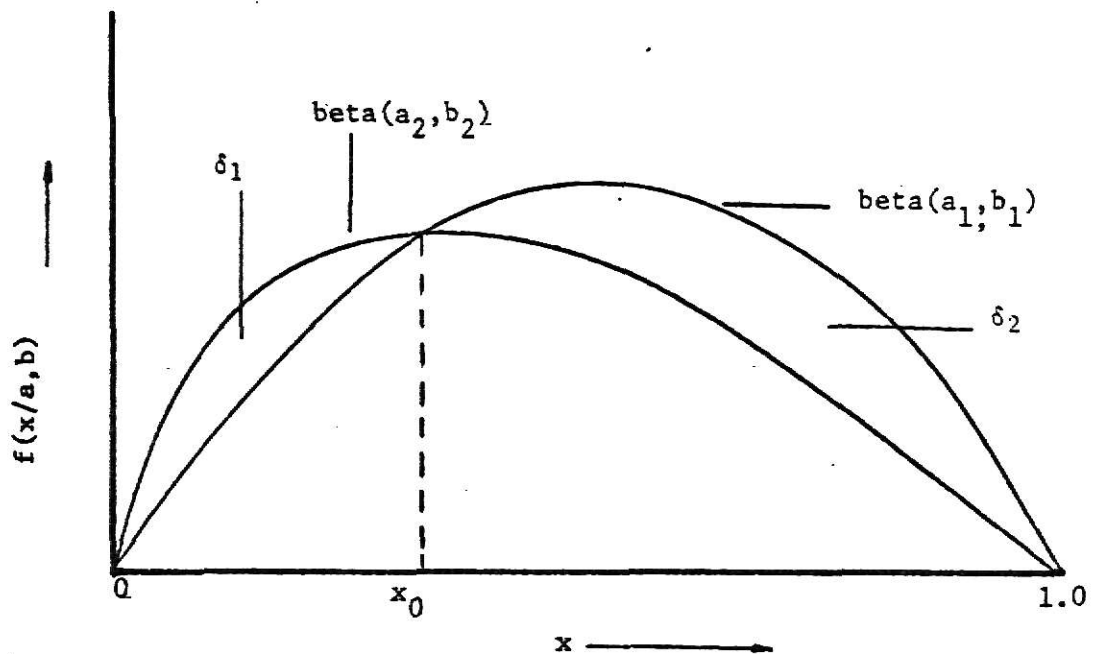Fig. 3.1(i). Graphs of beta (a_1, b_1) and beta (A_2, b_2) showing δ (two intersection points).

Fig. 3.1(ii). Graphs of beta (a_1, b_1) and beta (a_2, b_2) showing δ (one intersection point).
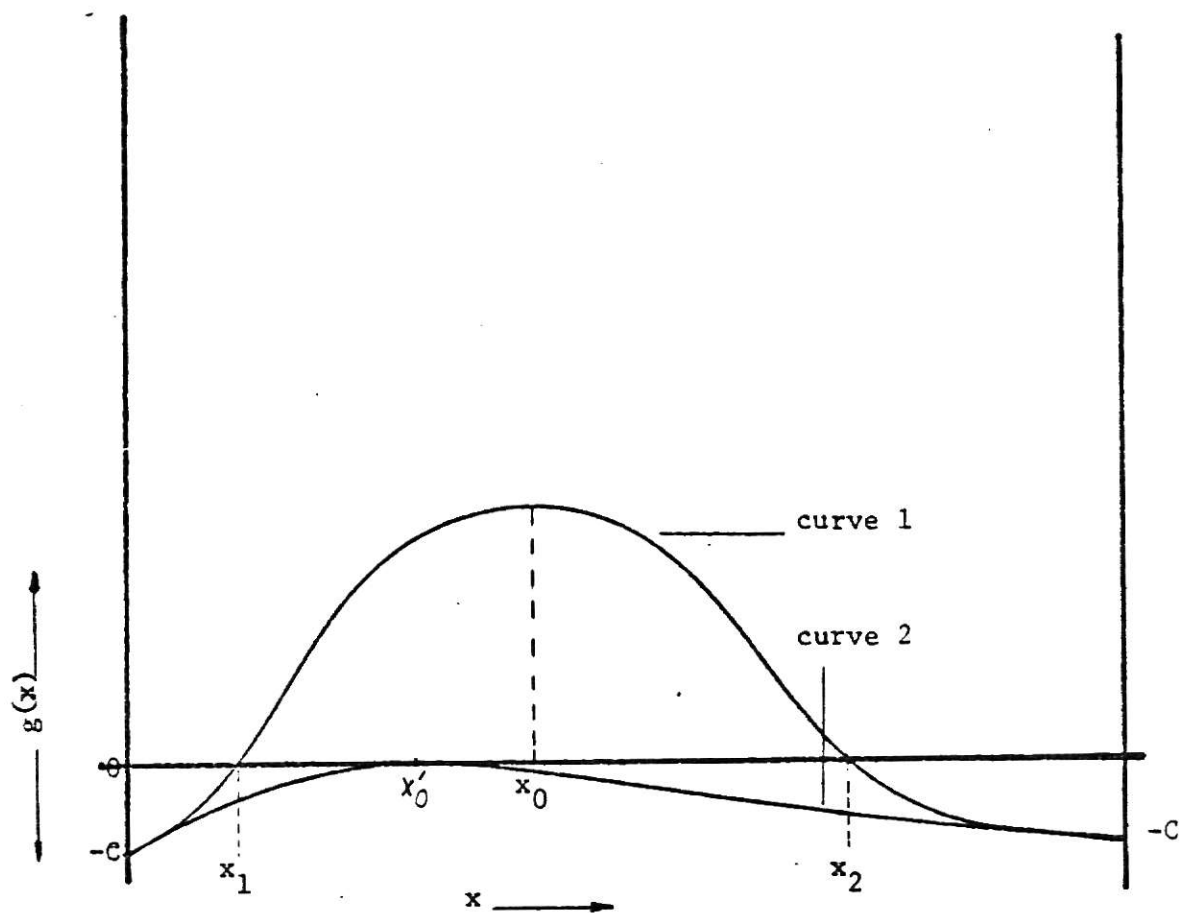
Fig. 3.2. Graph of g(x) against x.

$$g(x) = x^{a_1-a_2}(1-x)^{b_1-b_2} - C = 0 \qquad (3.4)$$

The critical point is determined by differentiating Eq. (3.4) and setting the result to 0 at $x=x_0$, the critical point. Thus,

$$\left.\frac{dg(x)}{dx}\right|_{x=x_0} = (a_1-a_2)x_0^{a_1-a_2-1}(1-x_0)^{b_1-b_2}$$

$$+ x_0^{(a_1-a_2)}(b_1-b_2)(1-x)^{b_1-b_2-1} = 0$$

hence,

$$x_0^{a_1-a_2-1}(1-x_0)^{b_1-b_2-1}[(a_1-a_2)(1-x_0) - (b_1-b_2)x_0] = 0$$

$$(a_1-a_2)(1-x_0) - (b_1-b_2)x_0 = 0$$

$$x_0 = \frac{a_1-a_2}{(a_1-a_2) + (b_1-b_2)} \qquad (3.5)$$

Equation (3.6) reveals that there is only one critical point over [0,1], given by x_0. Furthermore

$$\left.\begin{array}{l} g(x)_{x=0} = -C \\[2mm] g(x)_{x=1} = -C \end{array}\right\} \qquad (3.6)$$

Using (3.6) and (3.7) one can describe the general shape of $g(x)$ over x as shown by curve 1 and curve 2 in Figure 3.2.

Curve 1, Figure (3.2), illustrates the case where two intersection points at x_1 and x_2 were obtained. Curve 2 in the same figure illustrates the case where one intersection point was obtained. Referring to curve 1, x_0 is the maximum point and $g(x_0)$, from equation (3.5), is positive. Referring to curve 2, x_0' is the maximum point and $g(x_0')$ is zero. In

the latter case, the point of maximum is also the point of intersection of the two curves.

3.2 Procedure for Determining the Points of Intersection

A first step in determining the intersection points x_1 and x_2 is to establish whether the function $g(x)$ is of the type represented by curve 1 or by curve 2, referring to Figure 3.1. To do this, first evaluate

$$x_0 = \frac{a_1 - a_2}{(a_1 - a_2) + (b_1 - b_2)} \tag{3.7}$$

and then evaluate $g(x_0)$ from equation (3.5). The value of $g(x_0)$ determines the number of intersection points. Two cases are possible:

(i) $g(x_0) = 0$ => one root at x_0

(ii) $g(x_0) > 0$ => two roots, x_1 and x_2

3.2.1 Evaluation of δ for case (i) (one intersection point)

The intersection point x_0 is determined by equation (3.6). The "index of non-congruity" δ then becomes

$$= \int_0^{x_0} |f(x|a_1,b_1) - f(x|a_2,b_2)|\ dx + \int_{x_0}^1 |f(x|a_1,b_1) - f(x|a_2,b_2)|\ dx$$

$$= \delta_1 + \delta_2 \tag{3.8}$$

where

$$\delta_1 = \int_0^{x_0} |f(x|a_1,b_1) - f(x|a_2,b_2)|\ dx \tag{3.9}$$

$$\delta_2 = \int_{x_0}^1 |f(x|a_1,b_1) - f(x|a_2,b_2)|\ dx \tag{3.10}$$

3.2.1.1 Evaluation of δ_1

Since $f(x|a,b)$ is a positive function, Eq. (3.9) can also be written as:

$$\delta_1 = \left| \int_0^{x_0} f(x|a_1,b_1)dx - \int_0^{x_0} f(x|a_2,b_2)\,dx \right|$$

Each of the integrals is evaluated separately. However, neither of them can be evaluated analytically. Some form of series expansion method must be used, and subroutine BDTR [12] does this. The subroutine, documented in [12], computes $I_x(a,b)$ defined by:

$$I_x(a,b) = \int_0^x f(y|a,b)\,dy. \tag{3.11}$$

Then

$$\delta_1 = \left| I_{x_0}(a_1,b_1) - I_{x_0}(a_2,b_2) \right| \tag{3.12}$$

3.2.1.2 Evaluation of δ_2

Recall the formula for δ_2 (Eq. (3.10)) as

$$\delta_2 = \int_{x_0}^1 \left| f(x|a_1,b_1) - f(x|a_2,b_2) \right| dx$$

$$= \left| \int_{x_0}^1 f(x|a_1,b_1)dx - \int_{x_0}^1 f(x|a_2,b_2)dx \right| \tag{3.13}$$

where x_0 is the intersection point. Each of the above two integrals can be evaluated separately. Using the relationship:

$$I_x(a,b) = \int_0^x f(y|a,b)dy = 1 - \int_x^1 f(y|a,b)dy \tag{3.14}$$

each of the integrals in equation (3.12) is expressed as:

$$\int_{x_0}^{1} f(x \mid a_i, b_i) dx = 1 - \int_{0}^{x_0} f(x \mid a_i, b_i) dx$$

$$= 1 - I_{x_0}(a_i, b_i) \quad i = 1, 2 \quad (3.15)$$

hence

$$\delta_2 = \left| 1 - I_{x_0}(a_1, b_1) - (1 - I_{x_0}(a_2, b_2)) \right|$$

$$= \left| I_{x_0}(a_2, b_2) - I_{x_0}(a_1, b_1) \right| \quad (3.16)$$

Equation (3.16) for δ_2 is identical to equation (3.12) for δ_1. Thus

$$\delta_1 = \delta_2 \quad (3.17)$$

3.2.2 Evaluation of δ for two intersection points (case (ii))

Let x_1 and x_2 be the intersection points shown in Figure 1.2(i).
The index of non-congruity δ is given by

$$\delta = \delta_1 + \delta_2 + \delta_3$$

where

$$\delta_1 = \int_{0}^{x_1} \left| f(x \mid a_1, b_1) - f(x \mid a_2, b_2) \right| dx \quad (3.18)$$

$$\delta_2 = \int_{x_1}^{x_2} \left| f(x \mid a_1, b_1) - f(x \mid a_2, b_2) \right| dx \quad (3.19)$$

$$\delta_3 = \int_{x_2}^{1} \left| f(x \mid a_1, b_1) - f(x \mid a_2, b_2) \right| dx \quad (3.20)$$

Before computing each of the above integrals, it is necessary
to evaluate the points of intersection, x_1 and x_2, using equation
(3.5) which is reproduced below:

$$g(x) = x^{a_1 - a_2} (1-x)^{b_1 - b_2} C = 0 \quad (3.5)$$

Knowing that $g(x_0) > 0$ (recall Figure 3.1, curve 1) the roots of the nonlinear equation $g(x)$ must be obtained using a search technique. Mueller's iterative scheme, documented in [12], pages 217-218, is a convenient scheme to use, and has been incorporated into the computer program in Appendix 1.

3.3 Random Sampling from a Beta Distribution

Several methods can be used for generating random beta variates [13]. The method used here is a table look-up procedure with linear interpolation. Briefly, this method involves setting up a table of values of the cumulative distribution of beta (a,b) in the range. A random number is then interpolated directly into the cumulative tabled value to yield the desired beta variate. A discussion of the procedure used follows:

Let $f(x|a,b)$ defined by equations (1.1) and (1.2) be the given beta density function. The first step in generating a random sample is to set up a table of cumulative distribution function (c.d.f.) values. Subroutine BDTR, documented in [1], pages 78-80, can be employed for this purpose. The shape of the cumulative distribution function

$$f(x) = \int_0^x f(t|a,b)dt \qquad (3.25)$$

has the form depicted in Figure (3.3).

The next step is to generate a random number in the interval [0,1] and set this value equal to $F(x)$. Using linear interpolation (if necessary) the corresponding beta variate x can then be obtained.

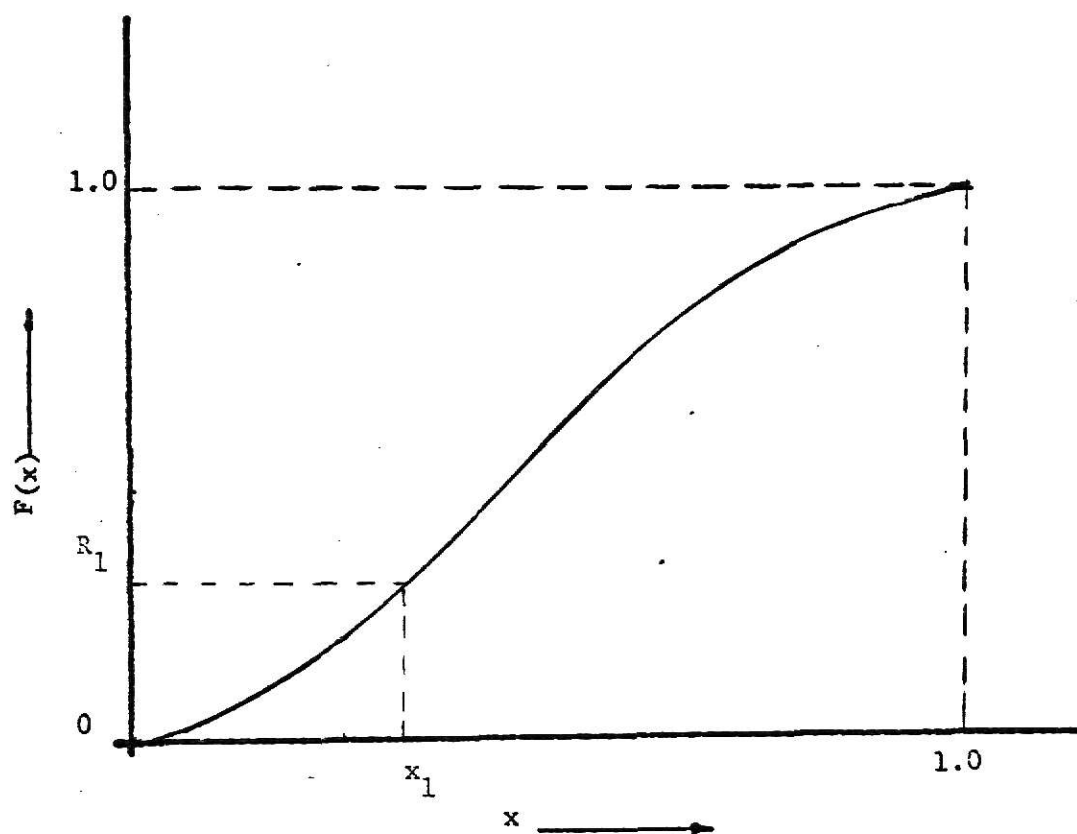Fig. 3.3.   Shape of cumulative distribution function for beta (a,b).

Referring to Figure 3.3, R_1 is a uniform random number and the corresponding beta variate is x_1.

The random number generator used in this investigation was obtained from the random number package "Super Duper" [14]. The function UNI(0) generates a uniform random variable in the half open interval [0,1].

The cumulative table is constructed with the beta variate x ranging from 0.0 (0.005) 1.0. For example, let RANUM(I) be the ith random number generated, and let PROB(K) and PROB(K-1) be the tabulated cumulative probability values with associated values of the beta variate X(K) and X(K-1), respectively, such that

$$PROB(K-1) < RANUM(I) < PROB(K) \tag{3.26}$$

and

$$X(K-1) < DEV(I) < X(K) \tag{3.27}$$

where DEV(I) is the desired variate.

$$DEV(I) = \frac{X(K-1)[PROB(K) - RANUM(I)] + X(K)[RANUM(I) - PROB(K-1)]}{[X(K) - X(K-1)]} \tag{3.28}$$

which is a linear interpolation method.

The next step is to group these random beta deviates into a frequency distribution.

This grouping is done as before for the case of the pseudo-random samples (Chapter 2). To compare these random frequencies (which may correspond to a real life situation) to the model (which is an assumed beta distribution) the χ^2 goodness-of-fit test is used (as was done for the case of the pseudo-random samples). The χ^2 value obtained in this manner is designated χ_s^2.

Comparisons were made between δ (discussed in this chapter), η_1 and η_2 (presented in Chapter 1) and the χ^2 values obtained between a pair of beta distributions. These comparisons will be discussed in the next chapter.

4.0 RESULTS

Results are presented in this chapter for a few special cases of beta distributions. A beta $(4,4)$ was used as the model, and beta $(3, b_2)$ were used as alternative distributions, where b_2 was varied in the range $(2,4)$ in steps of 0.1. Eight sets of relationships between the various measures of separation were studied, and each is briefly described below:

i) $\underline{\eta_1 \text{ versus } \delta}$ The results indicate that the index η_2 defines certain ranges over which η_1 and δ are related. Referring to Figure 4.1, we see that an increase in δ (associated with a pair of curves that are increasingly different) corresponds to an increase in η_1, as might be expected.

ii) $\underline{\delta \text{ versus } \chi^2_{ps}}$ Again here the index η_2 plays an auxiliary descriminating role in the functional relationship. Two distinct ranges of η_2 over which the relationship of δ against χ^2_{ps} is valid have been established, as shown in Figure 4.2. As might be expected, an increase in δ corresponds to an increase in χ^2_{ps}.

iii) $\underline{\eta_1 \text{ versus } \chi^2_{ps}}$ Referring to Figure 4.3, we see that over two ranges of η_2, η_1 increases as χ^2_{ps} increases.

iv) $\underline{\eta_2 \text{ versus } \chi^2_{ps}}$ Figure 4.4 depicts this relationship. A minimum value of χ^2_{ps} is obtained at $\eta_2 = 0.354$. This corresponds to a χ^2_{ps} value of 2.947, indicating that beta $(4,4)$ and beta $(3,3)$ are the least different from each other among those tested.

v) η_2 versus δ Figure 4.5 shows the relationship between η_2 and δ. A parabolic type relationship is again obtained, with a minimum at $\eta_2 = 0.379$. The value of δ at this point is 0.144. Comparing Figures 4.4 and 4.5, it appears that η_2 is a better index to work with, since the relationships obtained have been independent of η_1 or any other index.

vi) η_2 versus $(x^2_{ps}-x^2_s)$ Figure 4.6 shows the relationship. It was expected that in general $x^2_{ps} < x^2_s$, and a fixed quantity R (i.e., a random component) would have to be added to the x^2_{ps} value to obtain a more "realistic" approximation to a random sample value. Further, it was expected that R would be approximately constant from distribution to distribution and would always be postive. Figure 4.6, however, reveals something quite different. We see that x^2_s fluctuates around x^2_{ps}, indicating that in the limit, i.e., an x^2_2 increases indefinitely, x^2_s would approach the x^2_{ps} value.

vii) x^2_{ps} versus x^2_s This relationship is shown in Fig. (4.7). A least squares linear fit has been obtained for the data.

$$x^2_s = 1.518 \, x^2_{ps} - 2.949$$

The correlation coefficient for this case is

$$r^2 = 0.944.$$

viii) x^2_s versus δ Figure (4.8) depicts this relationship. A power curve of the form

$$y = ax^b, \quad x,y > 0$$

has been fitted to the data. This fit is given by

$$x^2_s = 298.35 \, \delta^{2.4}$$

and the correlation coefficient is

$$r^2 = 0.91$$

4.1 Conclusions

The results have been obtained for a few particular cases of beta distributions. Certain limited conclusions can be drawn from these cases, which may not apply for other model and alternative distributions. These conclusions are presented briefly below:

(i) For $\delta > .27$ (see Figures (4.2) and (4.8)), a χ_s^2 test would show significance at the $\hat{\alpha} = .05$ level of significance. This would apply for $a_2 = 3$ and $b_2 > 3.7$ or $b_2 < 2.5$. For $\hat{\alpha} = .01$, then for $\delta > .26$ a χ_s^2 test would show significance. This would apply for $a_2 = 3$ and $b_2 > 3.8$ or $b_2 < 2.4$.

(ii) The ranges of n_2 over which the χ_{ps}^2 test would show significance are illustrated in Figure 4.4. At $\hat{\alpha} = .05$ and $.01$, and an approximate range $n_2 > .5$ or $n_2 < .2$ a χ_{ps}^2 test would show significance.

The results obtained from the investigations for these special cases have been disappointing to the investigator in that the measures of separation developed are shown to be only moderately good, and are probably not promising enough to warrant further investigation.

| a_2 | b_2 | Pseudo Sample N = 100 | | | Random Sample N = 100 | | | $\chi^2_{ps} - \chi^2_s$ | η_1 | η_2 | δ |
		χ^2_{ps}	df	$\hat\alpha$	χ^2_s	df	$\hat\alpha$				
4	4	0	-	-	3.982	5	.5521	-3.982	-	-	-
3	3.9	17.11	5	.0043	16.91	5	.0047	-0.20	.075	.535	.308
3	3.8	14.27	5	.0140	13.82	5	.0168	0.45	.070	.518	.280
3	3.7	11.68	5	.0394	10.53	5	.0616	1.15	.065	.501	.252
3	3.6	9.802	5	.1330	7.99	5	.1568	1.81	.059	.483	.224
3	3.5	7.389	5	.1933	5.381	5	.3712	2.01	.054	.465	.197
3	3.4	5.725	5	.3339	4.264	5	.5121	1.46	.048	.445	.174
3	3.3	4.419	5	.4908	4.499	5	.4800	-0.08	.042	.424	.156
3	3.2	3.499	5	.6235	5.229	5	.3885	-1.73	.036	.402	.146
3	3.1	2.997	5	.7004	3.371	5	.6430	-0.37	.029	.379	.144
3	3.0	2.947	5	.7081	3.784	5	.5809	-0.84	.022	.354	.149
3	2.9	3.384	5	.6409	3.602	5	.6080	-0.22	.032	.328	.161
3	2.8	4.344	5	.5010	1.976	5	.8524	2.37	.042	.301	.179
3	2.7	5.862	5	.3199	4.270	5	.5112	1.59	.053	.272	.203
3	2.6	7.972	5	.1578	8.630	5	.1248	-0.658	.063	.240	.232
3	2.5	10.70	5	.0576	13.79	5	0.0170	-3.09	.074	.207	.266
3	2.4	14.08	5	.0151	17.36	5	0.0038	-3.28	.085	.172	.303
3	2.3	18.11	5	.0028	27.62	5	4.33×10^{-4}	-9.51	.097	.133	.343
3	2.2	22.80	5	$.369 \times 10^{-3}$	35.83	5	1.07×10^{-6}	-13.03	.109	.092	.386
3	2.1	28.03	5	$.34 \times 10^{-4}$	43.11	5	1.79×10^{-7}	-14.98	.121	.048	.431
3	2.0	34.05	5	$.24 \times 10^{-4}$	50.15	5	5.96×10^{-8}	-16.10	.133	0.000	.479

TABLE 4.1 Results using a beta $(4,4)$ model.
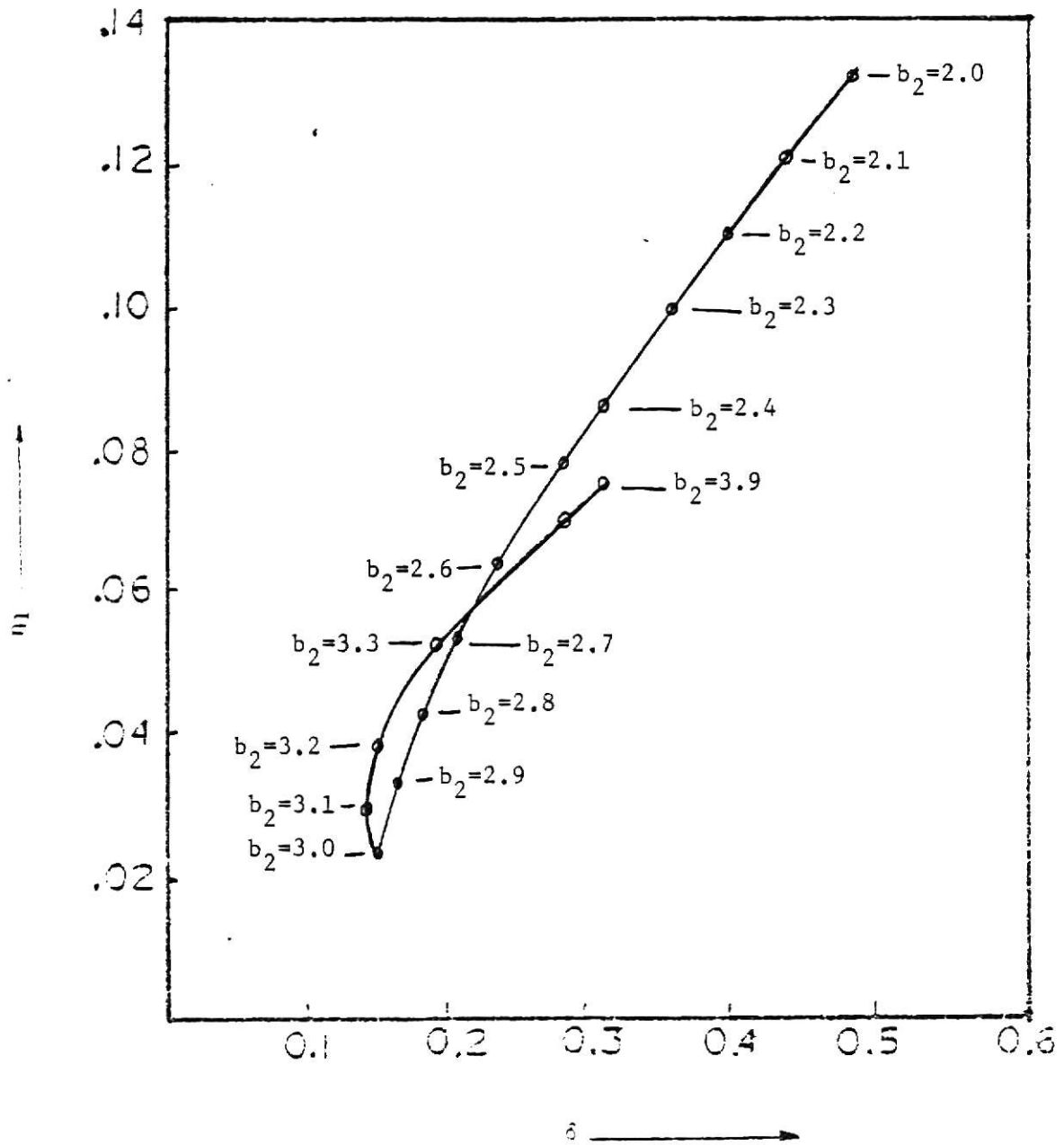
Fig. 4.1.   Relationship of η_1 and δ.
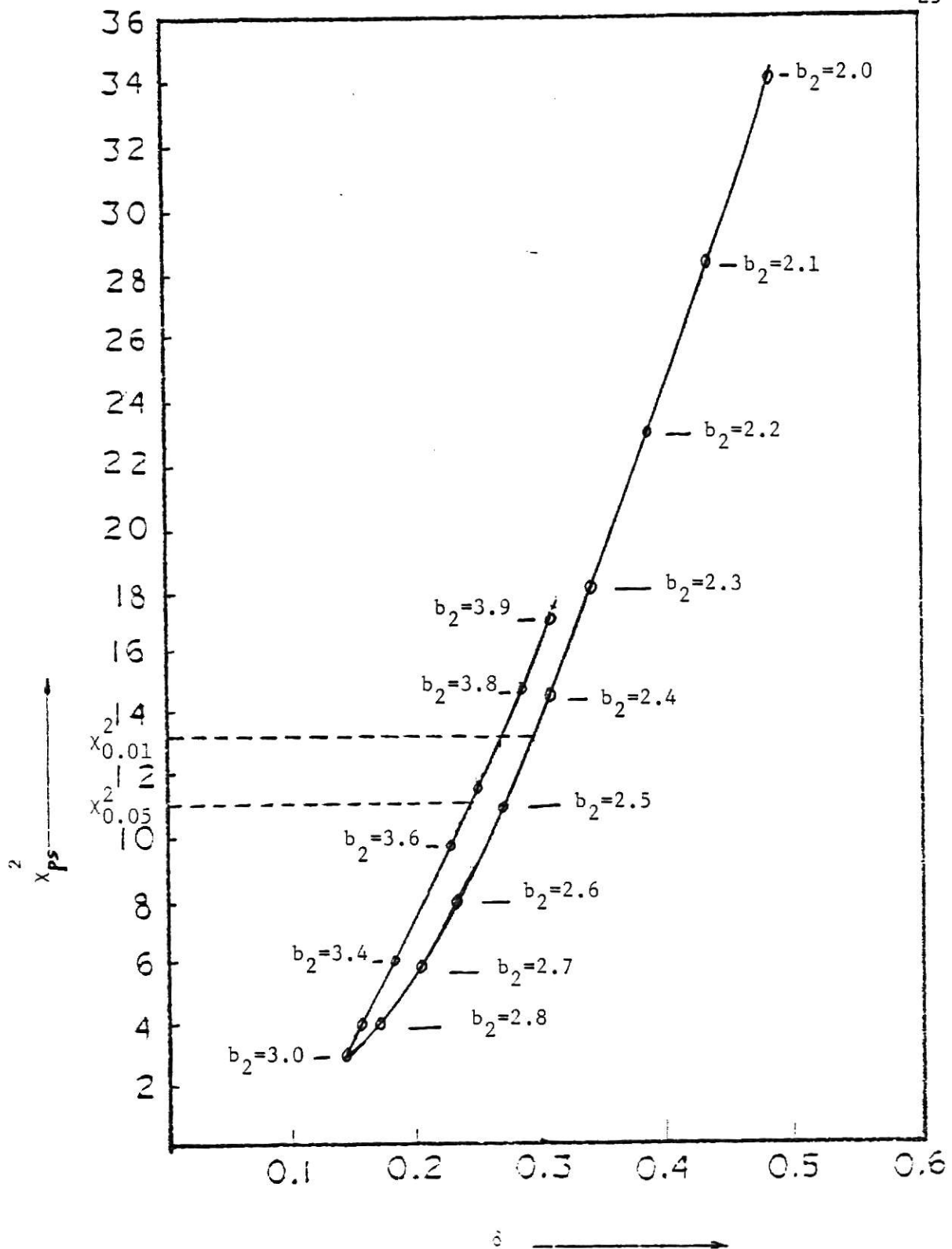
Fig. 4.2. Relationship of χ^2_{ps} and δ.
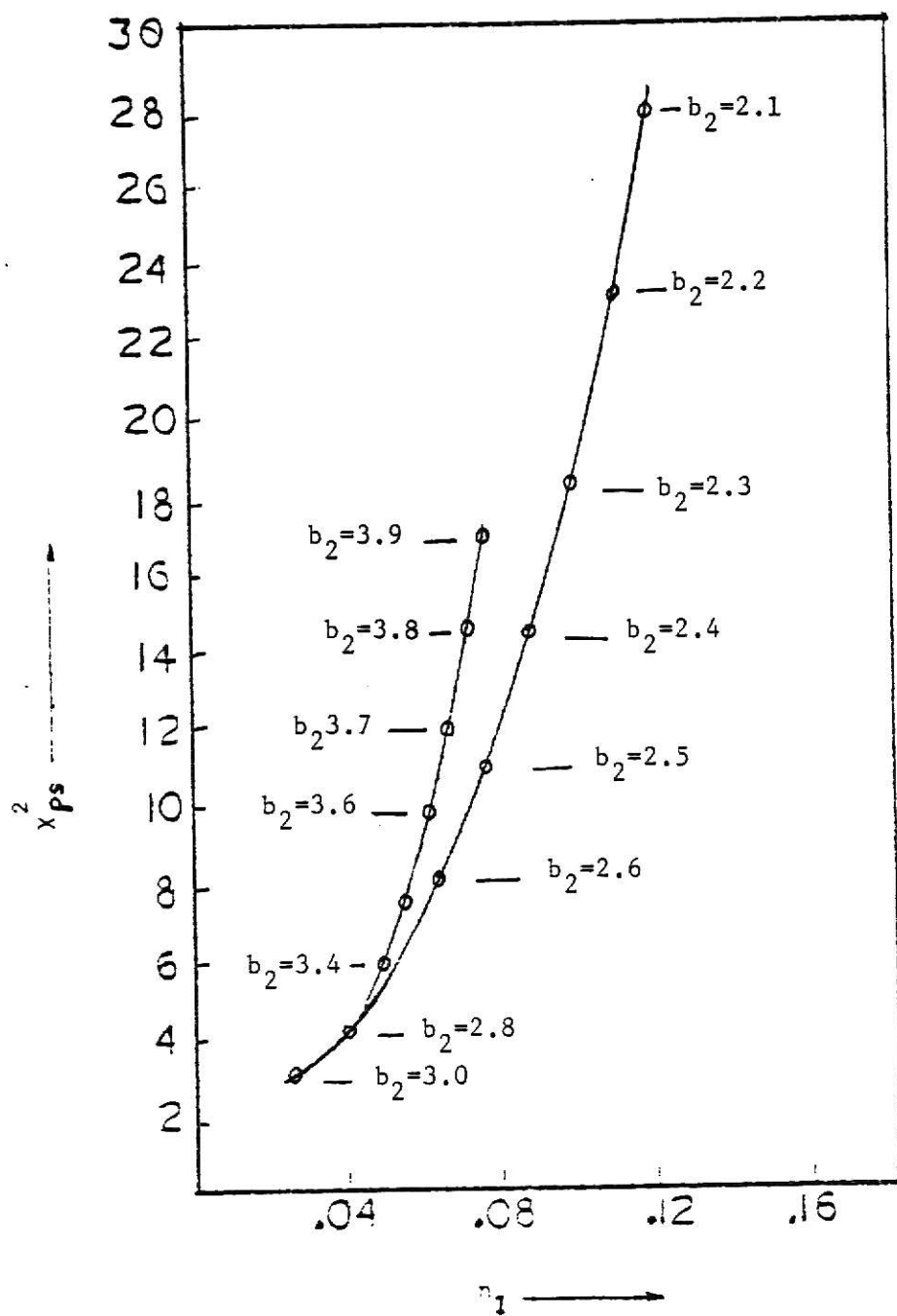
Fig. 4.3. Relationship of x^2_{ps} and η_1.
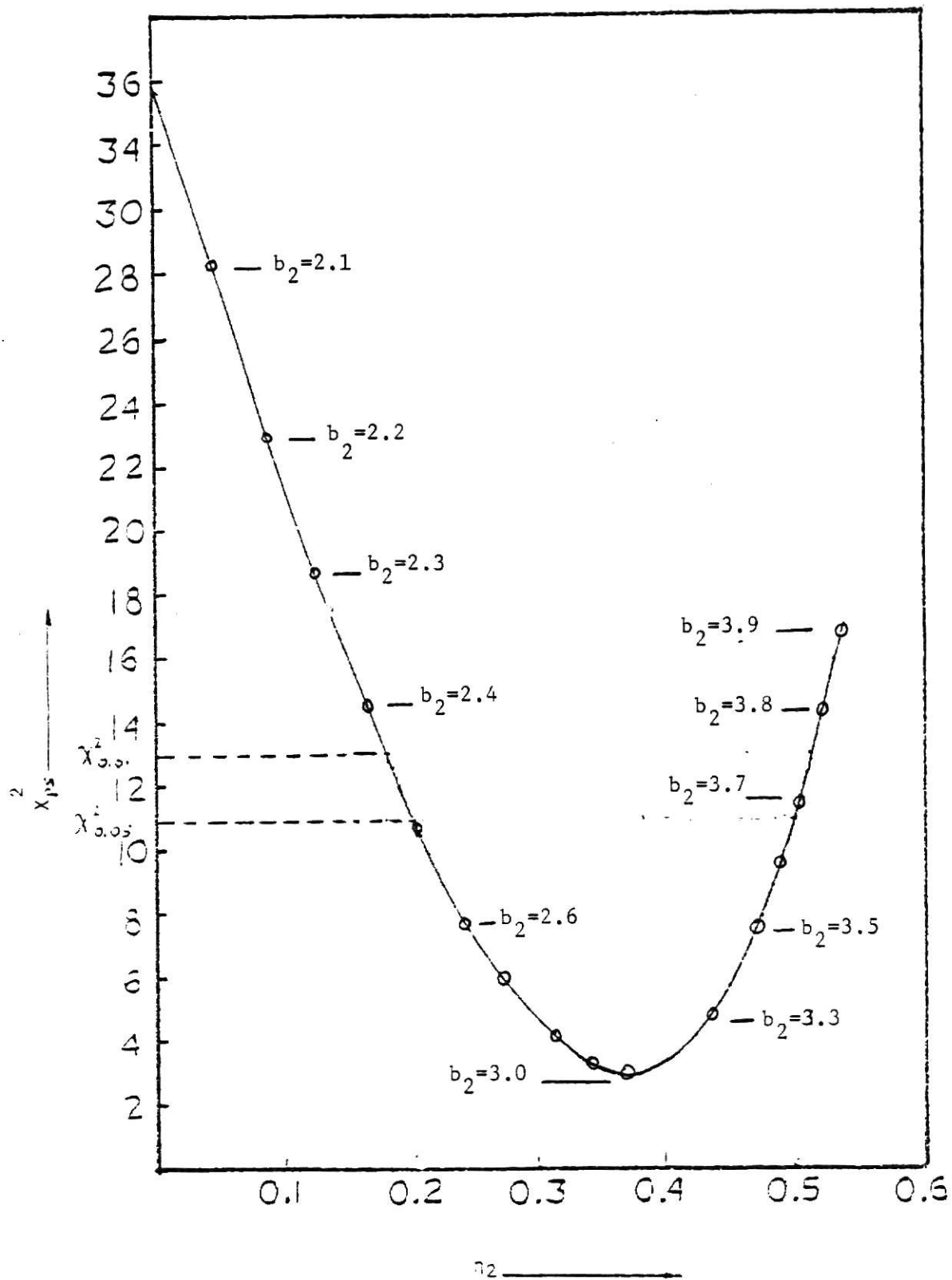
Fig. 4.4. Relationship of χ^2_{ps} and η_2.

Fig. 4.5. Relationship of δ and η_2.

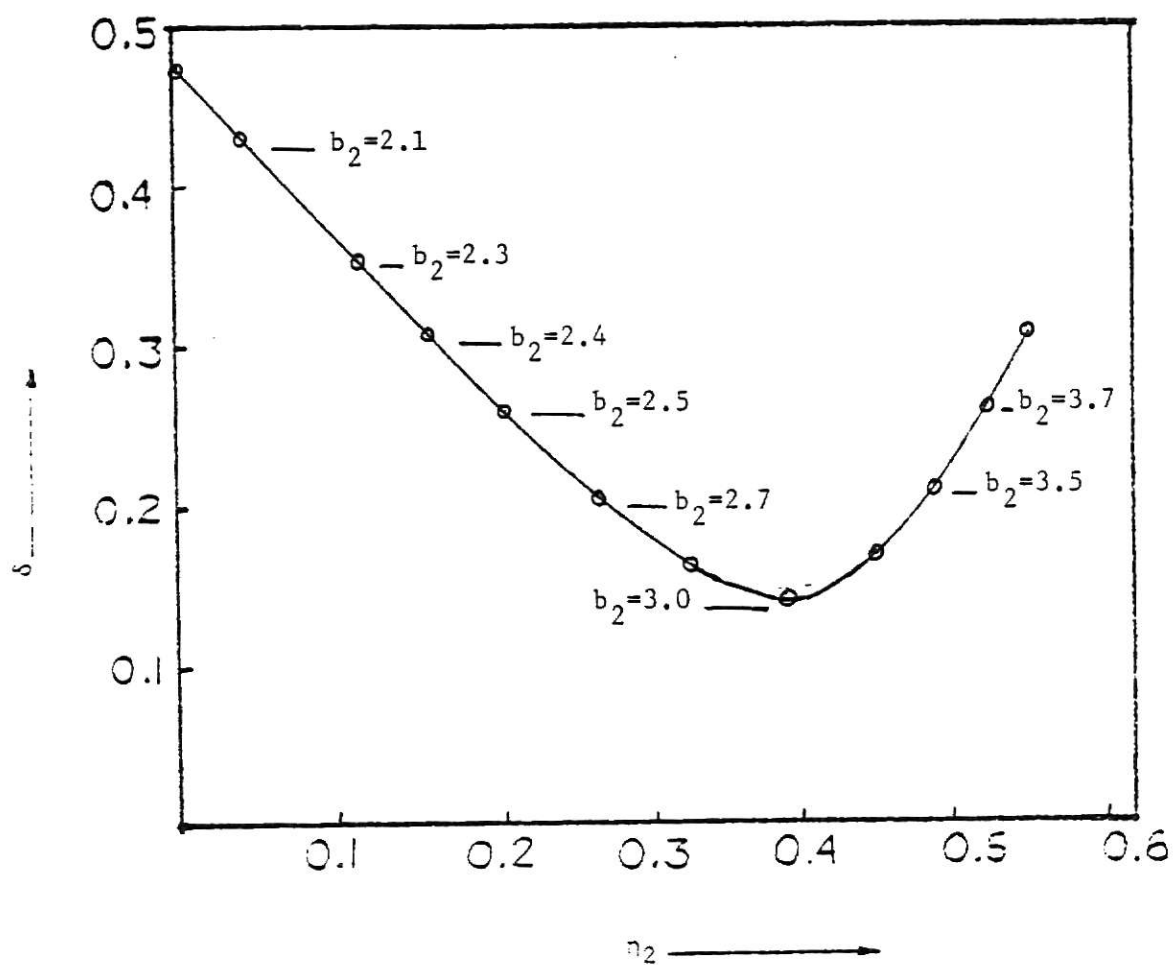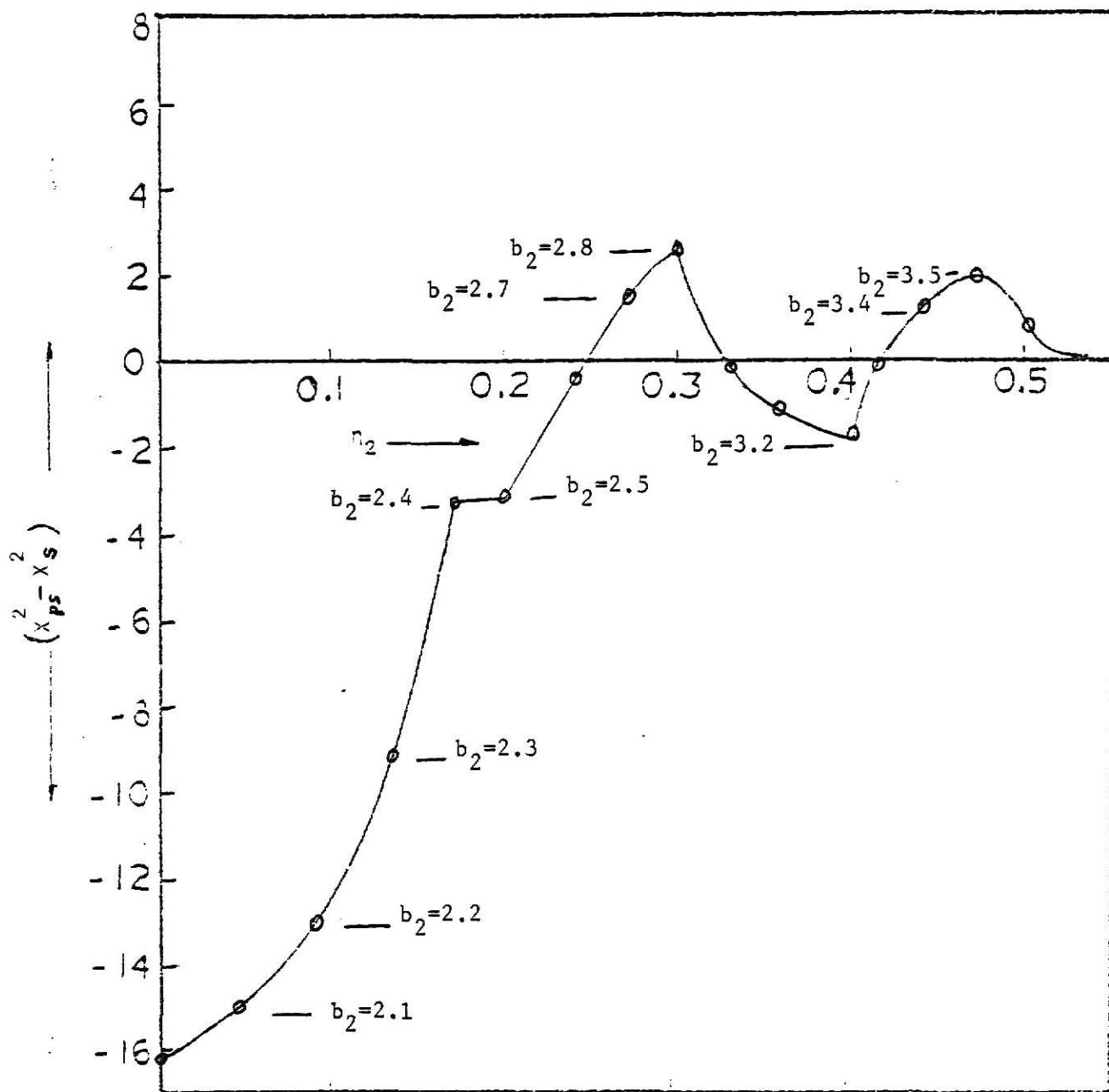Fig. 4.6. Relationship of $(x_{ps}^2 - x_s^2)$ and n_2.
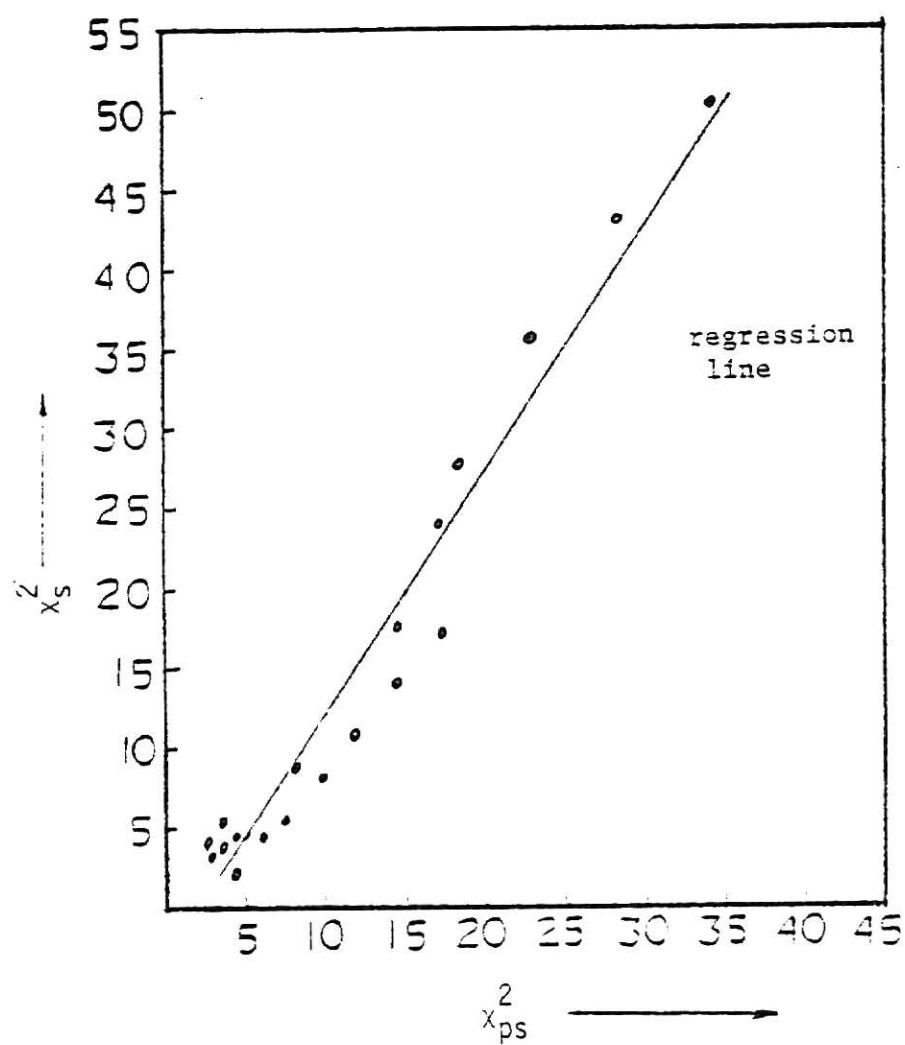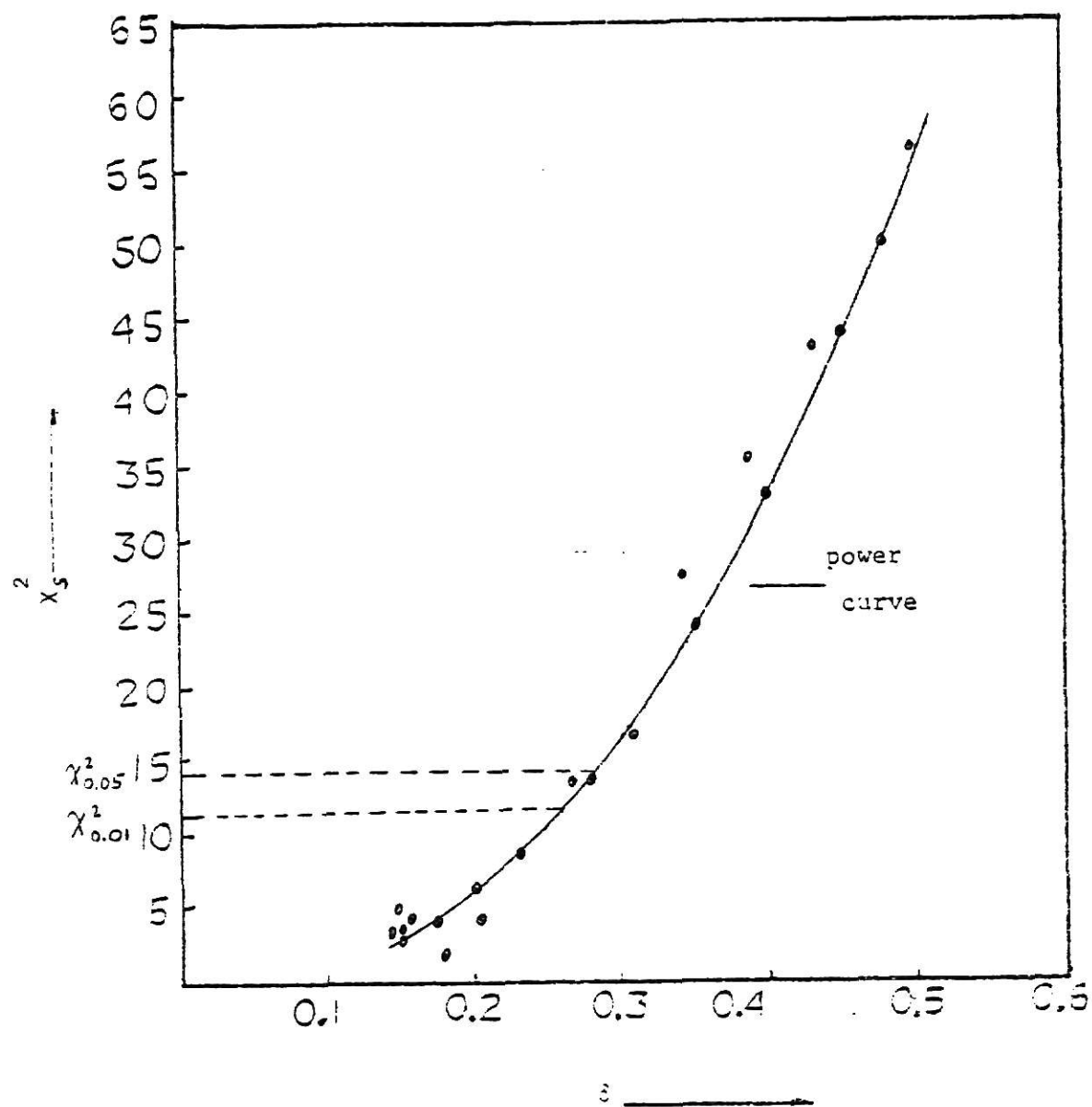
Fig. 4.7. Relationship of x^2_{ps} and x^2_s.

Fig. 4.8. Relationship of χ_s^2 and δ.

5.0 ACKNOWLEDGEMENTS

6.0 REFERENCES

1. Johnson, N. L., and Kotz, S., Distributions in Statistics: Continuous Univariate Distributions - 2. Wiley, Chapter 24, Section 5 (1970).

2. Ali, S. M. and Silvey, S. D., "A General Class of Coefficients of Divergence of One Distribution from Another," Jour. Roy Stat. Soc. 28, 131-142 (1966).

3. Adhikan, B. P., and Joshi, D. D., "Distance, Discrimination et Resumé Exhaustif," Publ. Inst. Statist. Univ. Paris, 5, 57-74 (1956).

4. Rao, C. R., Advanced Statistical Methods in Biometric Research, New York: Wiley, (1952).

5. Chernoff, H., "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," Ann. Math. Statist. 23, 493-507 (1952).

6. Kulback, S., "Information Theory and Statistics, New York: Wiley (1959).

7. Kolmogorov, A. N., "On the Approximation of Distributions of Sums of Independent Summands by Infinitely Divisible Distributions," Sankhya, 25, 159-174 (1963).

8. Jeffrey, S. H., Theory of Probability (2nd ed.), Oxford Univ. Press, (1948).

9. Mahalanobis, P. C., "On the Generalized Distance in Statistics," Proc. Nat. Inst. Sci. India, 12, 49-55 (1936).

10. Tildesley, M. L., A First Study of the Burmese Skull, Biometrika, 13, 176 (1921).

11. Freund, J. E., Mathematical Statistics (2nd ed.) Prentic Hall (1971).

12. IBM Corporation, "System/360 Scientific Subroutine Package Version III Programmer's Manual," 5th ed. (1970).

13. Fishman, G. S., "Concepts and Methods in Discrete Event Digital Simulation," Wiley, New York (1973).

14. Marsaglia, G., et al., "How to Use the McGill Random Number Package "Super Duper"", School of Computer Science, McGill University, Unpublished Manuscript.

APPENDIX I

This appendix contains the computer program used to compute η_1, η_2, δ_1 and χ^2 for a pair of beta distributions. Also included is an explanation of the symbols used in the program.

SYMBOLS USED IN PROGRAM

ALPHA(L), BETA(L)	input parameters, a and b, of the beta distributions. L=1 (model, L-2 (alternative).
CC, CD CE	symbols used in the calculation of $$B(a,b) = \frac{\Gamma(a) \ P(b)}{r(a+b)}$$
PMU(I)	mean of Ith distribution, I=1 (model),I=2 (alternative).
SIG(I)	std. deviation I^{th} distribution, I=1 (model), I=2 (alternative).
CON	B(a,b)
DELTA	index of non-congruity
DEV(L,I)	midpoint of class intervals for the Lth distribution, I^{th} class
DENS(L,I)	ordinates at the DEV(L,I) points.
PRBEL(L,I)	probability element for the Lth distribution, I^{th} class.
PRBE(L,I,M)	frequency at the Ith class for sample size M.
PRB(1,J,M)	"corrected" frequency taking into consideration the requirement that $E_i \geq 5$ for the Jth class and sample size M.
CHSQ(M)	the Chi-square statistic for sample size M.
J-1	degrees of freedom after computing the "effective" number of classes to satisfy the requirement that $E_i \geq 5$.
PROBAB(M)	right hand tail area of suitably chosen chi-square distribution for sample size M.

```
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C                                                      C
C     PURPOSE                                          C
C                                                      C
C     THE PURPOSE OF THIS PROGRAM IS TO CCMPUTE        C
C     INDICES OF DIFFERENCES BETWEEN SEVERAL           C
C     PAIRS OF BETA DISTRIBUTIONS, AND RELATE          C
C     THEM TO A GOODNESS OF FIT TEST. A DESCRIPTION    C
C     OF PARAMETERS IS INCLUDED SEPARATELY             C
C                                                      C
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
C
C
C
      DIMENSION PROB(2),PRO(2),CC(2),CD(2),CE(2),R(2),PMU(2),SIG(2),
     XPROBA(2,2),DEV(2,10),PRBEL(2,10),DENS(2,10),PRBE(2,10,500),
     XPRB(2,10,500),CHSQ(500),PROBAB(500)
      COMMON ALPHA(2),BETA(2),CON
      EXTERNAL FCT
C
C   INITIALIZE VALUES OF ALL VARIABLES
C
      DO 2 I=1,2
      PROB(I)=0.0
      PRO(I)=0.0
      CC(I)=0.0
      CD(I)=0.0
      CE(I)=0.0
      R(I)=0.0
      PMU(I)=0.0
    2 SIG(I)=0.0
      DO 3 J=1,2
      DO 4 K=1,2
    4 PROBA(J,K)=0.0
      DO 5 L=1,10
      DEV(J,L)=0.0
      PRBEL(J,L)=0.0
      DENS(J,L)=0.0
      DO 6 M=1,500
      PRBE(J,L,M)=0.0
    6 PRB(J,L,M)=0.0
    5 CONTINUE
    3 CONTINUE
      DO 7 N=1,500
      CHSQ(N)=0.0
    7 PROBAB(N)=0.0
C
C   READ IN VALUES OF THE PARAMETERS ALPHA AND BETA OF THE TWO DISTRIBUTIONS
C   THE CLASS WIDTH DINCR,THE NUMBER OF CLASS INTERVALS INUM,THE MAXIMUM SAMPLE
C   SIZE MMAX, AND THE STEP SIZE FOR COMPUTATIONS,N
C
      READ (5,1) ALPHA(1),BETA(1),ALPHA(2),BETA(2),DINCR,INUM,MMAX,N
    1 FORMAT (5F10.4,3I10)
C
C   DETERMINE THE VALUE OF THE CONSTANT (CON)
C
      DO 102 I=1,2
      XX=ALPHA(I)
      CALL GMMA (XX,GX,IER)
      CC(I)=GX
```

```
      XX=BETA(1)
      CALL GMMMA (XX,GX,IER)
      CD(1)=GX
      XX=ALPHA(1)+BETA(1)
      CALL GMMMA (XX,GX,IER)
  102 CE(1)=GX
      CON=(((CC(1)*CD(1))/(CC(2)*CD(2)))*CE(2)/CE(1)
      WRITE (6,103) CON
  103 FORMAT ('1','THE VALUE OF THE CONSTANT IS',2X,F9.4)
C
C   DETERMINE THE NUMBER OF INTERSECTION POINTS
C
      AA=ALPHA(1)-ALPHA(2)
      BB=BETA(1)-BETA(2)
      IF (AA.EQ.0.AND.ABS(BB).GT.0) GO TO 121
      IF (ABS(AA).GT.0.AND.BB.EQ.0) GO TO 122
      IF (AA.LT.0.AND.BB.GT.0) GO TO 123
      IF (AA.GT.0.AND.BB.LT.0) GO TO 124
      IF (AA.GT.0.AND.BB.GT.0) GO TO 125
C
C   DETERMINE THE INTERSECTION POINTS R(1) AND R(2)
C
  121 XO=1-(CON**(1/BB))
      GO TO 104
  122 XO = CON**(1/AA)
      GO TO 104
  123 AAA=0.01
      B=1.0
      EPS=0.01
      IEND=30
      CALL RTMI (X,F,FCT,B,AAA,EPS,IEND,IER)
      R(1)=X
      WRITE (6,107) R(1)
      GO TO 115
  124 AAA=0.0
      B=0.99
      EPS=0.01
      IEND=30
      CALL RTMI (X,F,FCT,B,AAA,EPS,IEND,IER)
      R(1)=X
      WRITE (6,107) R(1)
      GO TO 115
  125 XO=AA/(AA+BB)
      FXO=(XO**(ALPHA(1)-ALPHA(2)))*((1-XO)**(BETA(1)-BETA(2)))-CON
      IF (FXO) 104,104,105
  104 WRITE (6,107) XO
  107 FORMAT('1','THERE IS ONE INTERSECTION POINT'/' ','THIS POINT IS X=
     X',F7.4)
      R(1)=XO
      GO TO 115
  105 AAA=0.0
      BBB=1.0
      B=XO
      EPS=0.01
      IEND=20
      CALL RTMI (X,F,FCT,B,AAA,EPS,IEND,IER)
      R(1)=X
      CALL RTMI (X,F,FCT,B,BBB,EPS,IEND,IER)
      R(2)=X
      WRITE (6,108) R(1),R(2)
```

```
   42 FORMAT('1','FREQUENCY DISTRIBUTICN FCR MODEL B(',F7.4,',',F7.4,')
     XAND ALTERNATIVE B(',F7.4,',',F7.4,')'/' ','THE MODEL   FREU.      AR
     XE DENOTED BY EF AND THE ALTERNATIVE FREJUENCIES BY UF'////)
      WRITE (6,43)
   43 FORMAT ('0',T32,'CLASS MARK',8X,'UF',14X,'EF'/T32,12('-'),5X,4('-'
     X),12X,4('-'))
      DO 44 I=1,INUM
   44 WRITE (6,45) DEV(1,I),PRBEL(2,I),PRBEL(1,I)
   45 FORMAT('0',T35,F5.2,8X,F7.4,8X,F7.4)
C
C COMPUTE THE PSEUDO CHI-SQUARE FOR SAMPLES OF SIZE M
C
      WRITE (6,48)
   48 FORMAT('1',T30,'SAMPLE SIZE',8X,'CHI-SQUARE',8X,'DEGREES UF FREEDO
     XM',8X,'ALPHA'/T30,11('-'),8X,10('-'),8X,18('-'),8X,5('-'))
      DO 130 M=50,MMAX,N
      DO 131 L=1,2
      DO 132 I=1,INUM
  132 PRBE(L,I,M)=PRBEL(L,I)*M
  131 CONTINUE
      J=0
      YY=PRBE(2,1,M)
      Y=PRBE(1,1,M)
      CHSU(M)=0.0
      I=1
   52 IF(Y.GE.5.0) GO TO 53
      Y=Y+PRBE(1,I+1,M)
      YY=YY+PRBE(2,I+1,M)
      I=I+1
      IF(I+1.EQ.INUM) GO TO 54
      GO TO 52
   53 I=I+1
      IF(I+1.EO.INUM) GO TO 54
      J=J+1
      PRB(1,J,M)=Y
      PRB(2,J,M)=YY
      Y=PRBE(1,I,M)
      YY=PRBE(2,I,M)
      GO TO 52
   54 IF(PRBE(1,I+1,M).GE.5.0) GO TO 55
      J=J+1
      PRB(1,J,M)=Y+PRBE(1,I+1,M)+PRBE(1,I,M)
      PRB(2,J,M)=YY+PRBE(1,I+1,M)+PRBE(2,I,M)
      GU TO 56
   55 IF(PRBE(1,I,M).GE.5.0) GO TO 75
      J=J+1
      PRB(1,J,M)=Y+PRBE(1,I+1,M)
      PRB(2,J,M)=YY+PRBE(2,I+1,M)
      GU TO 56
   75 J=J+1
      PRB(1,J,M)=PRBE(1,I,M)
      PRB(2,J,M)=PRBE(2,I,M)
      J=J+1
      PRB(1,J,M)=PRBE(1,I+1,M)
      PRB(2,J,M)=YY+PRBE(2,I+1,M)
   56 DO 133 K=1,J
  133    CHSU(M)=CHSU(M)+(((PRB(2,K,M)-PRB(1,K,M))**2)/PRB(1,K,M))
      DF=J-1.
      CHSUE=CHSU(M)
C
```

```
108 FORMAT('1','THERE ARE TWO INTERSECTION POINTS'/'0','THESE ARE X(1)
   X=',F7.4,2X,'X(2)=',F7.4)
    GO TO 120
115 XXX=R(1)
C
C   DETERMINE THE INDEX OF NON-CONGRUITY,DELTA
C
    DO 116 I=1,2
    ALPH=ALPHA(I)
    BET=BETA(I)
    CALL MDBETA (XXX,ALPH,BET,P,IER)
    PROB(I)=P
116 PRO(I)=1.-P
    DELT1=ABS(PROB(1)-PROB(2))
    DELT2=ABS(PRO(2)-PRO(1))
    DELTA=DELT1+DELT2
    GO TO 300
120 DO 118 I=1,2
    XXX=R(I)
    DO 118 J=1,2
    ALPH=ALPHA(J)
    BET=BETA(J)
    CALL MDBETA (XXX,ALPH,BET,P,IER)
118 PROBA(I,J)=P
    DELT1=PROBA(1,1)-PROBA(1,2)
    DELT2=PROBA(2,1)-PROBA(1,1)
    DELT3=PROBA(2,2)-PROBA(1,2)
    DELT4=1-PROBA(2,2)
    DELT5=1-PROBA(2,1)
    DELTA=ABS(DELT1)+ABS(DELT3-DELT2)+ABS(DELT5-DELT4)
300 WRITE (6,101) DELTA
101 FORMAT('0','THE VALUE OF DELTA(THE INDEX OF NON-CONGRUITY) IS',F9.6)
   X6)
C
C   DETERMINE PROBABILITY INTERVALS FOR EACH CLASS INTERVAL FOR EACH DISTRIBUTION
C
    DO 301 L=1,2
    DEV(L,1)=0.0
    PRBEL(L,1)=0.0
    DO 10 I=2,INUM
    DEV(L,I)=DEV(L,I-1)+DINCR
 10 CONTINUE                                                  00000108
    I=2
 20 XXX=DEV(L,I)
    BET=BETA(L)
    ALPH=ALPHA(L)
    CALL MDBETA (XXX,ALPH,BET,P,IER)
    PRBEL(L,I)=P-PRBEL(L,I-1)
    IF (IER.EQ.0) GO TO 30                                    00000114
    WRITE (6,22) I,IER
 22 FORMAT (1H0,'ERROR IN ITERATION',I4,'. BOTH ERROR CODE IS',I4,  00000116
   1'.  END OF PROGRAM.')                                    00000117
    WRITE (6,23).                                            00000118
 23 FORMAT (1H0)                                              00000119
    GO TO 200
 30 IF(I.EQ.(INUM+1))GO TO 301
    I=I+1                                                     00000122
    GO TO 20                                                 00000123
301 CONTINUE
    WRITE (6,42) ALPHA(1),BETA(1),ALPHA(2),BETA(2)
```

```
C   DETERMINE THE RIGHT HAND TAIL AREA (PROBAB(M)) OF A SUITABLY CHOSEN
C   CHI-SQUARED DISTRIBUTION. M DENOTES THE SAMPLE SIZE.
C
      PROBAB(M)=CADTR(CHSQE,DF)
C
C   DETERMINE MEAN AND STANDARD DEVIATION,(PMU AND SIG RESPECTIVELY) OF THE
C   TWO DISTRIBUTIONS.
      DO 134 L=1,2
      PMU(L)=ALPHA(L)/(ALPHA(L)+BETA(L))
134   SIG(L)=SQRT(ALPHA(L)*BETA(L)/(((ALPHA(L)+BETA(L))**2)*(ALPHA(L)+B
     XETA(L)+1.)))
      WRITE(6,57) M,CHSQ(M),DF,PROBAB(M)
      IF(PROBAB(M).EQ.0.0) GO TO 200
57    FORMAT('0',I32,I3,13X,E13.4,13X,F5.2,13X,E13.4)
130   CONTINUE
C
C   COMPUTE THE VALUES OF THE PARAMETERS ETA(1) AAD ETA(2)
C
200   D=PMU(1)-PMU(2)
      DD=SIG(1)-SIG(2)
      E1=ABS(D)+ABS(DD)
      E2=(PMU(1)/SIG(1))-(PMU(2)/SIG(2))
      DDDD=D/DD
C
C   PRINT RESULTS
C
      WRITE(6,98)
98    FORMAT('1',T15,'I',T17,'PMU(1)-PMU(2)',T31,'I',1X,'SIG(1)-SIG(2)',
     XT47,'I',1X,'PMU(1)-PMU(2)',T63,'I',1X,'DELTA',T71,'I',1X,'ALPHA',T79,'I',T
     X79,'I',T82,'CHI',T88,'I',1X,'ETA(1)',T97,'I',1X,'ETA(2)',T105,
     X'I'/'0',T15,'I',T31,'I',T47,'I',1X,13('-')/'0',T15,'I',T31,'I',
     XT47,'I',1X,'SIG(1)-SIG(2)',T63,'I',T71,'I',T79,'I',1X,'SQUARE',
     XT88,'I',T97,'I',T105,'I'/'0',T15,100('-'))
      WRITE(6,99) D,DD,DDDD,DELTA,PROBAB(50),CHSQ(50),E1,E2
99    FORMAT('0',T19,F6.3,T35,F6.3,T51,F7.3,T64,F6.3,T72,F7.3,T80,F6.3,
     XT89,F6.3,T98,F6.3)
      STOP
      END
```

ESTIMATING STATISTICALLY SIGNIFICANT DIFFERENCES
BETWEEN A PAIR OF BETA DISTRIBUTIONS

by

KRISHNAN LAKSHMINARAYAN

B. Tech. (Mechanical Engineering)
Indian Institute of Technology, Madras, 1974

———————————

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements of the degree

MASTER OF SCIENCE

Department of Industrial Engineering
KANSAS STATE UNIVERSITY
Manhattan, Kansas

1980

ABSTRACT

Computational procedures have been developed for estimating measures of separation between a pair of beta distributions. These measures were variously defined as:

   (a)   the amount of non-congruity of the two density curves

   (b)   the measure of separation of the two means and standard deviations and

   (c)   the discrepancy of the μ/σ ratios.

They have been compared with each other and with a χ^2 goodness-of-fit test for a specific group of model and alternative distributions. The results obtained were not conclusive, i.e., the measures of separation developed were shown to be only moderately good for the particular cases studied, and are probably not promising enough to warrant further investigations.