WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

PROPERTIES OF SUCCESSIVE SAMPLE MOMENT ESTIMATORS[+]

by

E. Barouch[*], S. Chow[**],
G.M. Kaufman[***], and T. Wright[****]

April 1985 #1648-85

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

PROPERTIES OF SUCCESSIVE SAMPLE MOMENT ESTIMATORS[†]

by

E. Barouch[*], S. Chow[**],
G.M. Kaufman[***], and T. Wright[****]

April 1985 #1648-85

[†] This work was supported by Brookhaven National Laboratories, Contract #119863-S, E. Kaplan, Technical Advisor.

[*] Department of Mathematics, Clarkson University
[**] Sloan School of Management
[***] M.I.T. Energy Laboratory and Sloan School of Management
[****] Educational Resources Center, Clarkson University

Abstract

Moment type estimation of characteristics of a successively sampled finite population requires finding a solution to a pair of transcendental equations. Some global properties of two structurally distinct pairs -- a symmetric and an asymmetric pair -- of such equations are presented. Results of a monte carlo experiment designed to compare performance of estimators computed by solving these transcendental equation pairs are reported.

1. INTRODUCTION

Successive sampling models have recently been used to characterize random phenomena as diverse as oil and gas field discovery and the occurrence of software bugs (Barouch and Kaufman (1976), Littlewood, B. (1981), Gordon (1981, 1983), Andreatta and Kaufman (1983)), and the draft lottery (Du Mouchel (1970)). What distinguishes these applications from the usual treatment of successive sampling in the sample survey literature is the absence of information about the sample frame which allows a priori computation of the probability that a generic element of a finite population with N elements will be included in a sample of arbitrary size $n < N$.

Gordon (1983) has suggested use of a moment-like estimator for parameters of a successively sampled finite population. The basic idea is to split the sample into two parts and then set approximate Horvitz-Thompson type estimators equal to one another in order to calculate approximations of inclusion probabilities. Two moment matching alternatives are outlined, each of which implies a pair of transcendental equations. The equation pair studied in detail by Gordon (1983) is symmetric in form, while the alternative is not. He shows that with probability one the symmetric pair possesses a unique solution in the asymptotic limit $N \to \infty$ with n/N fixed.

While intuitively one might expect the two alternative formulations to be asymptotically equivalent, their behavior for finite samples is at issue: In particular when N is finite, one questions the existence and uniqueness of solutions of each pair. What are the finite sample properties of an estimate of the number of elements in the population or of an estimate of the sum of magnitudes of population elements generated by these two alternatives?

It is the purpose of this paper to establish a sufficient condition for existence of a solution of the asymmetric pair when sample size is finite and to compare properties of estimators of population attributes based on solutions to both pairs.

1.1 SUCCESSIVE SAMPLING

Consider a finite population of N elements with labels $U = \{1,2,...,N\}$ and associated magnitudes $A = \{A_1,...,A_N\}$; i.e. the magnitude of element $j \varepsilon U$ is $A_j > 0$, $j = 1,2,...,N$. A successive sampling scheme induces a distribution on permutations of elements of U as follows: for any permutation $(i_1,...,i_N)$ of $(1,2,...,N)$,

$$P\{(i_1,...,i_N)|A\} = \prod_{j=1}^{N} A_{i_j}/(A_{i_j} + \cdots + A_{i_N}). \tag{1.1}$$

An alternative representation of (1.1) can be given in terms of exponential order statistics: let $\tilde{x}_1,...,\tilde{x}_N$ be independent, identically distributed exponential random variables with means equal to one. Then (Gordon (1983))

$$P\{(i_1,...,i_N)|A\} = P\{\frac{\tilde{x}_{i_1}}{A_{i_1}} < \frac{\tilde{x}_{i_2}}{A_{i_2}} < \ldots < \frac{\tilde{x}_{i_N}}{A_{i_N}}\}. \tag{1.2}$$

This latter representation is an analytical lever for generation of moment-type estimators of unobserved population magnitudes when sampling is incomplete.

Suppose that we observe an <u>unordered</u> sample $s_n = \{i_1,...,i_n\}$ consisting of the first n elements of $(i_1,...,i_n)$, $n<N$, and that s_n is the <u>only</u> information available about the sampling scheme. How might we use s_n to estimate properties of A? In particular, we may wish to estimate $R = \sum_{j=1}^{N} A_j$, the total sum of magnitudes in A, the number N of elements in U, or the empirical frequency function of magnitudes in A.

1.2 MOMENT MATCHING ESTIMATORS

Gordon (1983) presents a moment type method for estimating finite population properties. His method rests on three key ideas: first, if one had access to all elements of A, then it is possible to compute the probability $P\{k \varepsilon s\} \equiv \pi_k(n)$ that element k U appears in a sample s_n of size n. Let $g(A_k)$ be a given function. Armed with $\pi_k(n)$, $k = 1,2,\ldots,N$, an unbiased estimator of the sum $\sum_{j=1}^{N} g(A_j)$, is $\sum_{k \varepsilon s} g(A_k)/\pi_k(n)$, an estimator introduced by Horvitz and Thompson (1952). If all elements of A were known with certainty a priori there would be no estimation problem.

The second key idea is that when N is large and $n/N = f$ is fixed, there exists a unique solution t_f to $N-n = \sum_{k=1}^{N} \exp\{-tA_k\}$ for which $|1 - \exp\{-t_f A_k\} - \pi_k(n)| = O(N^{-1})$ (Gordon (1983), Theorem 2.2). Thus, if $A_1,\ldots,A_N$ were known, t_f could be computed and $1 - \exp\{-t_f A_k\}$ would closely approximate $\pi_k(n)$. Once t_f is obtained one can replace $\pi_k(n)$ with $1 - \exp\{-t_f A_k\}$ to obtain an approximately unbiased Horwitz-Thompson estimator. Again, the hitch is that, given a sample s_n, only $A_{i_1},\ldots,A_{i_n}$ are observed and t_f depends on all elements of A.

To overcome these difficulties Gordon proposes a third idea. If the complete sample is s_n, split it into an "early" part s_m consisting of the first $m<n$ observations and a "late" part consisting of the remaining n-m observations. In order to simplify notation and with no loss in generality, relabel elements of s_n so that $s_n = \{1,2,\ldots,n\}$ and $s_m = \{1,2,\ldots,m\}$.

Define h = m/n, and let t_h solve $N\text{-}m = \sum_{k=1}^{N} \exp\{-tA_k\}$. Then

$\hat{R}_h(\delta) = \sum_{k=1}^{m} A_k^{\delta}/(1 - \exp\{-t_h A_k\})$ and $\hat{R}_f(\delta) = \sum_{k=1}^{n} A_k^{\delta}/(1 - \exp\{1 - \exp\{-t_f A_k\})$

are approximately unbiased estimators of the characteristic $R(\delta) \equiv \sum_{k=1}^{N} A_k^{\delta}$, $0 \leq \delta \leq 1$. For two distinct choices of δ, δ_1 and δ_2, $\hat{R}_h(\delta_1) = \hat{R}_f(\delta_1)$ and $\hat{R}_h(\delta_2) = \hat{R}_f(\delta_2)$ constitute two equations in two unknowns, t_h and t_f.

In particular for the case $\delta_1 \equiv \delta$ and $\delta_2 = 1$, one obtains the symmetric pair of equations (with $t_k \equiv \alpha, t_f \equiv \alpha+\beta$).

$$\sum_{k=1}^{n} \frac{A_k}{1 - e^{-(\alpha+\beta)A_k}} = \sum_{j=1}^{m} \frac{A_k}{1 - e^{-\alpha A_k}} \tag{1.3a}$$

and

$$\sum_{k=1}^{n} \frac{A_k^{\delta}}{1 - e^{-(\alpha+\beta)A_k}} = \sum_{j=1}^{m} \frac{A_k^{\delta}}{1 - e^{-\alpha A_k}} . \tag{1.3b}$$

A different estimator arises if one utilizes s_n and the "late" portion of s_n in the following fashion: The "late" part $\{A_{m+1},\ldots,A_n\} \equiv s_{n|m}$ of s_n is generated by successively sampling $\{A_{m+1},\ldots,A_N\}$. Define $\ell = (n\text{-}m)/N$ and let t_ℓ be a solution to $\sum_{k=m+1}^{N} \exp\{-t_\ell A_k\} = N\text{-}n$. Then $\sum_{j=m+1}^{n} A_j/(1\text{-}\exp\{-t_\ell A_j\})$ estimates $R - \sum_{j=1}^{m} A_j$. By the same logic that led to (1.3a) and (1.3b), with $\beta \equiv t_\ell$ we set

$$\sum_{j=1}^{m} \frac{A_j}{1 - e^{- \;A_j}} = \sum_{j=1}^{m} A_j + \sum_{k=m+1}^{n} \frac{A_j}{1 - e^{-\beta A_j}} \, . \tag{1.3c}$$

Together with (1.3a), (1.3c) is a pair of equations that can be solved for α and β. We call (1.3a) and (1.3c) the asymmetric pair, and begin our analysis with a study of properties of this pair.

2. PROPERTIES OF THE ASYMMETRIC PAIR

For easy reference we restate the asymmetric pair as

$$\sum_{j=1}^{n} A_j/(1-\exp\{-(\alpha+\beta)A_j\}) = \sum_{j=1}^{m} A_j/1-\exp\{-\alpha A_j\}) \tag{2.1a}$$

and

$$\sum_{j=1}^{m} A_j/(1-\exp\{-\alpha A_j\}) = \sum_{j=1}^{m} A_j + \sum_{j=m+1}^{n} A_j/(1-\exp\{-\beta A_j\}). \tag{2.1b}$$

Our first task is to find conditions that guarantee existence of a solution. Consider first the existence of a solution when $\beta \to \infty$. As $\beta \to \infty$, (2.1a) and (2.1b) are redundant; i.e. both take the form

$$\sum_{j=1}^{n} A_j = \sum_{j=1}^{m} A_j/(1-\exp\{-\alpha A_j\}). \tag{2.2}$$

<u>Lemma 1</u>: Define the function g of α as

$$g(\alpha) = \sum_{j=1}^{n} A_j - \sum_{j=1}^{m} A_j/(1-\exp\{-\alpha A_j\}), \qquad \alpha>0. \tag{2.3}$$

For $A_1, \ldots, A_n$ finite and positive, a solution $\alpha_c > 0$ of $g(\alpha) = 0$ exists and is unique.

<u>Proof</u>: (Compare with Gordon (1983), Lemma 5.8(a)) With g as defined in (2.3),

(i) $\lim_{\alpha \to \infty} g(\alpha) = \sum_{j=m+1}^{n} A_j > 0.$

(ii) $g(\alpha)$ is continuous for $0 < \alpha < \infty$.

(iii) For α small, $1 - \exp\{-\alpha A_j\} = \alpha A_j - \frac{1}{2}\alpha^2 A_j^2 + O(\alpha^3)$, so

$$g(\alpha) = \sum_{j=1}^{n} A_j - \frac{1}{x} \sum_{j=1}^{m} \frac{1}{(1 - \frac{1}{2}\alpha A_j + O(\alpha)^2)} = \sum_{j=1}^{n} A_j - \frac{1}{\alpha} \sum_{j=1}^{m} (1 + \frac{1}{2}\alpha A_j + O(\alpha^2))$$

whereupon

$$g(\alpha) = -\frac{m}{\alpha} + (\sum_{j=1}^{n} A_j - \frac{1}{2}\sum_{j=1}^{m} A_j) + O(\alpha). \qquad (2.4)$$

(iv) For $\alpha > 0$ and small, (iii) implies $g(\alpha) < 0$.

The function $g(\alpha)$ is positive for large α, negative for small α and continuous. Thus a zero α_c of $g(\alpha) = 0$ must exist.

(v) As $\frac{\partial g}{\partial \alpha} = \sum_{j=1}^{m} \frac{A_j^2 e^{-\alpha A_j}}{[1 - e^{-\alpha A_j}]^2} > 0$, $\partial g/\partial\alpha$ at $\alpha = \alpha_c$ is positive,

so the solution (α_c, ∞) is unique. ∎

Lemma 2: No solution (α_0, β_0) to (2.1a) and (2.1b) with $\alpha_0 > \alpha_c$ exists.

Proof: First consider a solution of the form (α_0, β), $\beta \to \infty$. Lemma 1 shows that (α_c, ∞) with $g(\alpha_c) = 0$ is a unique solution. In addition $\partial g/\partial\alpha > 0$. For any α, $\beta < \infty$, the left side of (2.1a) is greater than $\sum_{j=1}^{n} A_j$. Consequently, there can be no solution (α_0, β_0) with $\alpha_0 > \alpha_c$. ∎

We next establish a sufficient condition for existence of a solution (α_0, β_0) to (2.1a) and (2.1b) with $0 < \alpha_0 < \alpha_c$.

Theorem 1: Defining $\frac{1}{n}\sum_{j=1}^{n} A_j = \bar{A}_n$ and $\frac{1}{m}\sum_{j=1}^{m} A_j = \bar{A}_m$, a sufficient condition for existence of a solution (α_0, β_0) with $0 < \alpha_0 < \alpha_c$ is $\bar{A}_m > \bar{A}_n$. When $\bar{A}_m > \bar{A}_n$ the number of solutions is odd.

We set the stage for proof of Theorem 1 with three propositions based on consideration of two functions: the Implicit Function Theorem allows us to define a function $\beta_1(\alpha)$ via (2.1a) and a function $\beta_2(a)$ via (2.1b). A solution of (2.1a) and (2.1b) obtains when $\beta_1(\alpha) = \beta_2(\alpha)$. It is evident that $\lim \beta_1(\alpha) = \lim \beta_2(\alpha) = \infty$ as $\alpha \to \alpha_c$ from below. We shall show that $\beta_1(\alpha)$ and $\beta_2(\alpha)$ approach zero as $\alpha \to 0$, that when $\bar{A}_m > \bar{A}_n$, $\beta_2(\alpha) > \beta_1(\alpha)$ for α in a neighborhood of zero and that $\beta_2(\alpha) < \beta_1(\alpha)$ for α in a neighborhood of α_c. These facts imply that $\beta_1(\alpha)$ and $\beta_2(\alpha)$ must cross an odd number of times in the open interval $(0, \alpha_c)$.

Proposition 1: $\lim_{\alpha\to 0} \beta_1(\alpha) = \lim_{\alpha\to 0} \beta_2(\alpha) = 0.$

Proof: Compute a Taylor expansion of $\beta_1(\alpha)$ and $\beta_2(\alpha)$ for $\alpha > 0$ and small.

Equation (2.1a) takes the form

$$\frac{n}{\alpha + \beta_1(\alpha)} + \frac{1}{2}\sum_{j=1}^{n} A_j \doteq \frac{m}{\alpha} + \frac{1}{2}\sum_{j=1}^{m} A_j . \tag{2.5}$$

Since (2.5) possesses a singularity on both sides, equating singular parts yields the leading term of the expansion of $\beta_1(\alpha)$. This is $\beta_1(\alpha) \doteq (\frac{n-m}{m})\alpha$. The second term of the expansion involves a constant C_1 such that

$$\beta_1(\alpha) = (\frac{n-m}{m})\alpha[1 + C_1\alpha],$$

so

$$\alpha + \beta_1(\alpha) \cong (\frac{n}{m}\alpha)[1 + (\frac{n-m}{n})C_1\alpha].$$

This allows us to write

$$\frac{n}{\alpha + \beta_1(\alpha)} \cong \frac{m}{\alpha} - \frac{m}{n}(n-m)C_1.$$

Substitution of this relation in the Taylor expansion of (2.1a) yields

$$C_1 = \frac{n}{2m(n-m)}[\sum_{j=m+1}^{n} A_j].$$

Therefore,

$$\beta_1(\alpha) \cong (\frac{n-m}{m})\alpha[1 + \frac{n\alpha}{2m(n-m)}(\sum_{j=m+1}^{n} A_j)].$$

Similarly,

$$\beta_2(\alpha) \cong (\frac{n-m}{m})\alpha[1 + \frac{\alpha}{2m}\sum_{j=1}^{n} A_j].$$

Thus $\beta_1(\alpha) \to 0$ and $\beta_2(\alpha) \to 0$ as $\alpha \to 0$. ■

Proposition 2: For $\alpha > 0$ and $\alpha \sim 0$ and $\bar{A}_m > \bar{A}_n$, $\beta_2(\alpha) > \beta_1(\alpha)$.

Proof: Subtract the two Taylor expansions of $\beta_1(\alpha)$ and $\beta_2(\alpha)$. The leading term cancels and

$$[\beta_2(\alpha) - \beta_1(\alpha)]\frac{m}{(n-m)\alpha^2} = \frac{1}{2m(n-m)}[(n-m)\sum_{j=1}^{n} A_j - n\sum_{j=m+1}^{n} A_j]$$

$$= \frac{n}{2(n-m)}[\bar{A}_m - \bar{A}_n] > 0.$$ ■

Proposition 3: For $\alpha_c - \alpha > 0$ and $\alpha \sim \alpha_c$, $\beta_2(\alpha) < \beta_1(\alpha)$.

Proof: In the vicinity of α_c, $\beta_1(\alpha)$ and $\beta_2(\alpha)$ are unbounded and increase indefinitely. Since α_c is finite and since m and n are finite, we can neglect α_c in comparison with $\beta_1(\alpha)$ in some terms on the left hand side of (2.1a). The system (2.1a) and (2.1b) takes a slightly modified form, valid for $\alpha \sim \alpha_c$,

$$\sum_{j=1}^{m} \frac{A_j}{1 - e^{-(\alpha+\beta_1(\alpha))A_j}} + \sum_{j=m+1}^{n} \frac{A_j}{1 - e^{-\beta_1(\alpha)A_j}} \simeq \sum_{j=m+1}^{m} \frac{A_j}{1 - e^{-\alpha A_j}} \tag{2.6a}$$

and

$$\sum_{j=1}^{m} \frac{A_j}{1 - e^{-\alpha A_j}} = \sum_{j=1}^{m} A_j + \sum_{j=m+1}^{n} \frac{A_j}{1 - e^{-\beta_2(\alpha)A_j}} . \tag{2.6b}$$

Since at $\alpha = \alpha_c$ we have a solution and since the left side of (2.6a) equals the right side (2.6b), we equate them. Doing so yields,

$$\sum_{j=1}^{m} \left[\frac{A_j}{1 - e^{-(\alpha+\beta_1(\alpha))A_j}} - A_j\right] + \sum_{j=m+1}^{n} A_j\left[\frac{1}{1 - e^{-\beta_1(\alpha)A_j}} - \frac{1}{1 - e^{-\beta_2(\alpha)A_j}}\right] = 0. \tag{2.7}$$

The first term is positive. Thus the second term is negative implying $\beta_1(\alpha) > \beta_2(\alpha)$ for $\alpha_c - \alpha > 0$ and $\alpha \sim \alpha_c$. ■

Theorem 1 states that if $\bar{A}_m > \bar{A}_n$, then (2.1a) and (2.1b) have at least one solution in $(0, a_c)$ and the number of solutions is odd.

<u>Proof</u>: A solution takes place when $\beta_1(\alpha) = \beta_2(\alpha)$. Define $\delta(\alpha) = \beta_2(\alpha) - \beta_2(\alpha)$ $\delta(\alpha)$ is continuous and differentiable in $(0, a_c)$. Proposition 2 implies $\delta(\alpha) > 0$ for $\alpha > 0$ and $\alpha - 0$, and Proposition 3 implies $\delta(\alpha) < 0$ for $\alpha_c - \alpha > 0$ and $\alpha - \alpha_c$. This guarantees existence of at least one solution α_0 to $\delta(\alpha_0) = 0$ in the open interval 0 to α_c and excludes an even number of zeros of $\delta(\alpha)$ in this interval. ∎

The theorem provides a simple, sample based sufficient condition for existence of a solution; i.e. $\bar{A}_m$ and $\bar{A}_n$ are sample statistics.

An alternative formulation of the problem of demonstrating existence and uniqueness of a solution to the equations studied in the preceding subsection is to study the character of solutions to

$$G(\alpha) = F_1(\alpha + \beta_2(\alpha)) - F_2(\alpha) \tag{2.8}$$

where

$$F_1(x) = \sum_{j=1}^{n} A_j/(1 - e^{-xA_j}), \tag{2.9}$$

$$F_2(x) = \sum_{j=1}^{m} A_j/(1 - e^{-xA_j}), \tag{2.10}$$

and $\beta_2(\alpha)$ is defined implicitly by (2.1b); i.e. with

$$F_4(x) = \sum_{j=m+1}^{n} A_j/(1 - e^{-xA_j}), \tag{2.11}$$

$\beta_2(\alpha)$ is the value of β satisfying

$$F_2(\alpha) = mA_m + F_4(\beta). \tag{2.12}$$

The equivalence becomes transparent upon differentiating G and both sides of (2.12) with respect to α. Then

$$G'(\alpha) = \frac{dG(\alpha)}{d\alpha} = (1 + \frac{d\beta_2(\alpha)}{d\alpha}) \frac{\partial F_1}{\partial \alpha} - \frac{dF_2}{d\alpha} \tag{2.13}$$

and via the definition of $\beta_1(\alpha)$,

$$1 + \frac{d\beta_1(\alpha)}{d\alpha} = \frac{dF_2}{d\alpha} / \frac{\partial F_1}{\partial \alpha} , \tag{2.14}$$

so that

$$G'(\alpha) = \frac{\partial F_1}{\partial \alpha} \{\frac{d\beta_2(\alpha)}{d\alpha} - \frac{d\beta_1(\alpha)}{d\alpha}\}. \tag{2.15}$$

It has been shown that $G(0) = \frac{1}{2} m[\bar{A}_n - \bar{A}_m]$. When $\bar{A}_n < \bar{A}_m$ a solution in $(0, \alpha_c)$ exists (Theorem 1), so demonstration that at any zero of $G(\alpha)$ occurring in $(0, \alpha_c)$ $G'(\alpha) > 0$ is sufficient to establish uniqueness of the solution. (If $G(0) < 0$, $G'(\alpha)$ at the left-most α satisfying $G(\alpha) = 0$ must be positive. If $G'(\alpha) > 0$ at <u>any</u> $\alpha \in (0, \alpha_c)$ satisfying $G(\alpha) = 0$, then the solution is unique).

3. PROPERTIES OF THE PAIR (1.3a) AND (1.3b)

We turn now to the symmetric pair

$$\sum_{k=1}^{n} \frac{A_k}{1-e^{-(\alpha+\beta)A_k}} = \sum_{j=1}^{m} \frac{A_j}{1-e^{-\alpha A_j}} \tag{3.1}$$

and

$$\sum_{k=1}^{n} \frac{A_k^{\delta}}{1-e^{-(\alpha+\beta)A_k}} = \sum_{j=1}^{m} \frac{A_j^{\delta}}{1-e^{-\alpha A_j}} \tag{3.2}$$

with $0 \leq \delta < 1$. In what follows we shall assume that $A_j > 0$, $j=1,2,\ldots,n$ and bounded.

Our first observation is that a solution to (3.1) and (3.2) for unbounded β does not exist. If $\beta \to \infty$, (3.1) and (3.2) become

$$\sum_{j=1}^{n} A_j = \sum_{j=1}^{m} \frac{A_j}{1-e^{-\alpha A_j}} \tag{3.3}$$

and

$$\sum_{j=1}^{n} A_j^{\delta} = \sum_{j=1}^{m} \frac{A_j^{\delta}}{1-e^{-\alpha A_j}} \, . \tag{3.4}$$

A necessary condition for a unique solution to (3.3) and (3.4) is that they be redundant. Since (3.3) and (3.4) are in general not redundant, the solution to (3.3) will in differ from that for (3.4), so no unique value of α may solve both (3.3) and (3.4). Consequently, in general, no solution to (3.1) and (3.2) of the form $(\alpha,\beta) = (\alpha_0,\infty)$ exists.

In addition, the solution in α to (3.3) is bounded, so we conclude that if a solution (α_0,β_0) to (3.1) and (3.2) exists, both α_0 and β_0 must be bounded.

We next examine the behavior of the symmetric pair in a neighborhood of the origin for an illustrative value $\frac{1}{2}$ of δ. A Laurent expansion of the right- and left-hand sides of (3.1) and (3.2) yields

$$\frac{n}{\alpha + \beta} \cong \frac{m}{\alpha} \tag{3.5}$$

and

$$\frac{1}{\alpha + \beta} \sum_{j=1}^{n} A_j^{-\frac{1}{2}} \cong \frac{1}{\alpha} \sum_{j=1}^{m} A_j^{-\frac{1}{2}}, \tag{3.6}$$

so that a solution at $(\alpha,\beta) = (0,0)$ may exist only if $\frac{1}{n} \sum_{j=1}^{n} A_j^{-\frac{1}{2}} = \frac{1}{m} \sum_{k=1}^{m} A_k^{-\frac{1}{2}}$, a very special constraint on values of $A_1,\ldots,A_n$.

By keeping one more term in the series expansion of the form (3.5) and (3.6) and solving the resulting equations, a lower bound can be established on possible solutions to the symmetric pair. However, the approximate solution so obtained does not provide any information about rates of convergence of numerical methods for computing solutions. To this end define the averages

$$\bar{B}_\ell = \frac{1}{\ell} \sum_{j=1}^{\ell} A_j^{\frac{1}{2}} \qquad \ell = m,n \tag{3.7}$$

$$\bar{C}_\ell = \frac{1}{\ell} \sum_{j=1}^{\ell} A_j^{-\frac{1}{2}} \qquad \ell = m,n, \tag{3.8}$$

and approximate (3.1) and (3.2) by

$$\frac{n}{\alpha + \beta} + \frac{1}{2}\, n\, \bar{A}_n \cong \frac{1}{\alpha}\, m + \frac{1}{2}\, m\, \bar{A}_m \tag{3.9}$$

and

$$\frac{n}{\alpha + \beta}\, \bar{C}_n + \frac{1}{2}\, n\, \bar{B}_n \cong \frac{m}{\alpha}\, \bar{C}_m + \frac{1}{2}\, m\, \bar{B}_m\ . \tag{3.10}$$

An approximate solution is then

$$\alpha \cong \frac{2m(\bar{C}_n - \bar{C}_m)}{n[\bar{A}_n \bar{C}_n - \bar{B}_n] - m[\bar{A}_m \bar{C}_n - \bar{B}_m]} \tag{3.11}$$

and, as with $\beta_1(\alpha)$ (cf. (2.5)),

$$\beta \cong \frac{n - m}{m} \alpha . \tag{3.12}$$

4. COMPUTATION METHODS

Consider the system of transcendental equations

$$F_1(\alpha,\beta) \equiv \sum_{j=1}^{n} A_j[1-\exp(-\{\alpha+\beta\}A_j)]^{-1} - \sum_{j=1}^{m} A_j[1-\exp(-\alpha A_j)]^{-1} = 0 \qquad (4.1)$$

$$F_2(\alpha,\beta) \equiv \sum_{j=1}^{m} A_j[1-\exp(-\alpha A_j)]^{-1} - \sum_{j=1}^{m} A_j - \sum_{j=m+1}^{n} A_j[1-\exp(-\beta A_j)]^{-1} = 0 \qquad (4.2)$$

$$F_3(\alpha,\beta) \equiv \sum_{j=1}^{n} A_j^{\frac{1}{2}}[1-\exp(-\{\alpha+\beta\}A_j)]^{-1} - \sum_{j=1}^{m} A_j^{\frac{1}{2}}[1-\exp(-\alpha A_j)]^{-1} = 0 \qquad (4.3)$$

From (4.1), (4.2) and (4.3), one obtains two independent pairs. The asymmetric pair (4.1) and (4.2) and the symmetric pair (4.1) and (4.3).

As pointed out earlier, each of the sums in (4.1), (4.2), and (4.3) has a pole at the origin $(\alpha,\beta) = (0,0)$, and $\lim F(\alpha,\beta) = 0$ as α, $\beta \to 0$. The range of possible solutions to $F_1(\alpha,\beta) = 0$ and $F_2(\alpha,\beta) = 0$ with $0 \leq \beta \leq \infty$ restricts the range of α to $0 \leq \alpha \leq \alpha_c \equiv$ ACRIT. Furthermore, the fact that β can be unbounded opens up the possibility of exponential underflows. Accounting for underflows may in turn give rise to inaccuracies in numerically computed solutions, so care must be exercised in implementing a numerical scheme to solve the pairs $F_1(\alpha,\beta) = 0$, $F_2(\alpha,\beta) = 0$, and $F_1(\alpha,\beta) = 0$, $F_3(\alpha,\beta) = 0$.

Since we have an arbitrary choice of scale, we chose to set $\sum_{j=1}^{n} A_j = 1$. The computer program for solving both symmetric and asymmetric pairs begins by solving

$$\sum_{j=1}^{n} A_j - \sum_{j=1}^{m} A_j[1-\exp(-\alpha A_j)]^{-1} = 0 \qquad (4.4)$$

for $\alpha_c \equiv$ ACRIT. To avoid exponential underflows e^{-170} is approximated by zero. This protection is used throughout the entire numerical scheme.

To eliminate poles at the origin, a lower limit of $(10^{-12}, 10^{-12})$ is set on (α,β) and an upper limit of ACRIT-10^{-10} is set on α.

Once ACRIT is computed by the use of a modified regula falsi (MRF) routine (GTRANS), equation (4.2) is solved for a given α by a second MRF routine (GMRF) to yield β as a function of α. This $\beta(\alpha)$ is substituted in equation (4.1) and GTRANS is used to find the desired $\alpha=\alpha_0$, which is then substituted back into (4.2) to yield $\beta(\alpha_0)$. The range allowed for α is $(10^{-12}$,ACRIT-$10^{-10})$ and for β is $(10^{-12},10^{70})$. Since this range is large, the process is repeated with $\alpha \varepsilon (\alpha_0-1, \alpha_0+1)$, $\beta \varepsilon (\beta_0-1, \beta_0+1)$ to narrow the range on which the regula falsi iteration takes place with 10^{-14} tolerance. As an accuracy check, a two-dimensional Newton-Raphson (N-R) scheme is implemented utilizing (α_0,β_0) as the starting solution with a tolerance of 10^{-14}. When the results from both methods agree to desired accuracy, the solution (α_0, β_0) is accepted and estimates $\hat{N}$ of N and $\hat{R}$ of R are computed.

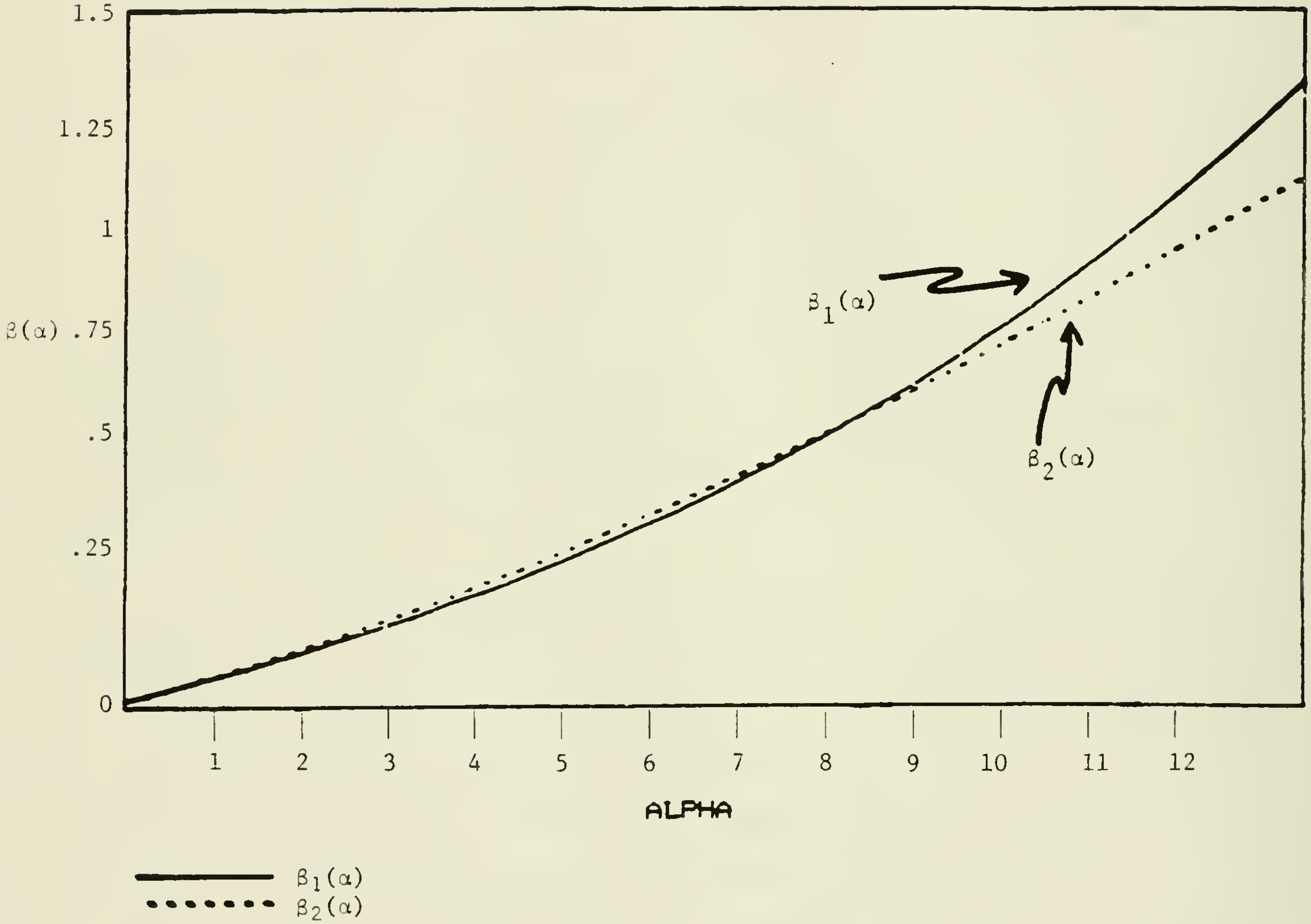
This procedure was used to solve both symmetric and asymmetric pairs for 400 Monte Carloed successive samples (cf. section 5). In most cases it works well. However, there are cases for which the GMRF and GTRANS equation solvers do not bracket the roots. Then an initial guess of ACRIT/2 is used for α, GTRANS finds the corresponding β, and the 2-dimensional N-R scheme is utilized with this initial guess. This procedure worked quite well for the asymmetric pair but failed several times for the symmetric pair. In each such instance a solution was found by brute force: $[0, \alpha_c]$ was divided into 30,000 subintervals and the initial root bracketing interval found by identification of change of sign of values of $\beta_1(\alpha) - \beta_3(\alpha)$, $\beta_1(\alpha)$ a solution to $F_1(\alpha, \beta) = 0$ for given α, and $\beta_3(\alpha)$ a solution to $F_3(\alpha, \beta) = 0$.

A unique solution to the asymmetric pair was found in each of six cases where the sufficient conditon $\bar{A}_m > \bar{A}_n$ is not satisfied.

The nature of the difficulties experienced in computing solutions is reflected in the character of the graph of $\beta_1(\alpha)$ and $\beta_2(\alpha)$ vs. α. For many cases, $|\beta_2(\alpha) - \beta_1(\alpha)|$ is small in value on the interval $(0,\alpha_0)$, but grows rapidly as $\alpha \to \alpha_c$.

[Figure 4.1 Here]

Figure 4.1.



5. MONTE CARLO STUDY OF SOLUTIONS

In order to study the comparative behavior of solutions to both the symmetric and asymmetric equation pairs, a Monte Carlo simulation consisting of 400 successive samples drawn from a fixed finite population was performed. Samples were generated from a population of N=170 elements and magnitudes $A_1,\dots,A_{170}$, with A_k being the (k/N+1)st fractile of a lognormal population with parameter $(\mu,\sigma^2) = (.83, 1.62)$. This choice of values for μ, σ^2, and N matches estimates provided by Meisner and Demirmen (1980) for a segment of the North Sea, based on n=58 discoveries. These particular values of N, μ, and σ^2 correspond to $R = \sum_{j=1}^{N} A_j = 818$. For each of 400 successive samples of size n=58, a solution to the symmetric pair (1.3a) and (1.3b) and a solution to the asymmetric pair (1.3a) and (1.3c) were computed by the methods described in section 4. Each sample was split at m=38.

Tables and graphs describing properties of Monte Carloed sampling distributions of $\hat{N}$ and $\hat{R}$ for this particular case are presented in section 5.1.

5.1 SAMPLING PROPERTIES OF $\hat{R}$ AND OF $\hat{N}$

Summary displays are in the form of:

(1) Parallel boxplots of $\hat{N}$ and of $\hat{R}$ values generated by symmetric pair and by asymmetric pair solutions.

(2) Quantile-quantile plot of $\hat{R}$-R quantiles vs. unit normal quantiles for the symmetric pair and for the asymmetric pair.

(3) Quantile-quantile plot of $\hat{N}$-N quantiles vs. unit normal quantiles for the symmetric pair and for the asymmetric pair.

(4) Measures of location and of spread for $\hat{N}$ and for $\hat{R}$ values.

(5) Scatterplot of $\hat{N}$ values generated by the symmetric pair vs. $\hat{N}$ values generated by the asymmetric pair. A similar scatterplot for $\hat{R}$ values.

Some tentative conclusions about the behavior of estimates generated by moment matching as described in earlier sections emerge. When the particular finite population used in this Monte Carlo experiment is successively sampled with sample size n=58 and a split at m=38:

(I) Estimators of N and of R derived by solving either the symmetric or the asymmetric pair of equations are negatively biased.

The parallel boxplots for $\hat{N}$ and $\hat{R}$ values (Figures 5.1 and 5.2), the quantile-quantile plots for $\hat{N}$-N and $\hat{R}$-R values (Figures 5.3 to 5.6), and the summary statistics in Table 5.1 display this bias in different ways.

(II) The sampling distributions for both $\hat{N}$ and $\hat{R}$ values generated by solving the symmetric pair exhibit larger spreads and more (right tail) outliers than the corresponding distributions generated by solving the asymmetric pair.

The quantile-quantile plots of Figures 5.3 and 5.4 for $\hat{R}$-R values exhibit fatter than normal right tails. Figure 5.4, in particular, is a display of $\hat{R}$-R values for the symmetric pair, which exhibits substantial right tail skewedness. If, however, the eight largest of the four hundred values plotted are disregarded, the graph defined by the remaining points appears very close to a straight line.

> (III) While extreme right tails (above .98 quantiles) of distributions of $\hat{R}$-R values are fatter than normal, these distributions appear to be close to normal in shape elsewhere.

In contrast, a visual examination of Figures 5.5 and 5.6 shows that:

> (IV) Distributions of $\hat{N}$-N generated by solving either the asymmetric pair or the symmetric pair are decisively not normal in shape.

TABLE 5.1

SUMMARY MEASURES FOR MONTE CARLOED SAMPLING DISTRIBUTIONS OF ESTIMATORS $\hat{N}$ AND $\hat{R}$

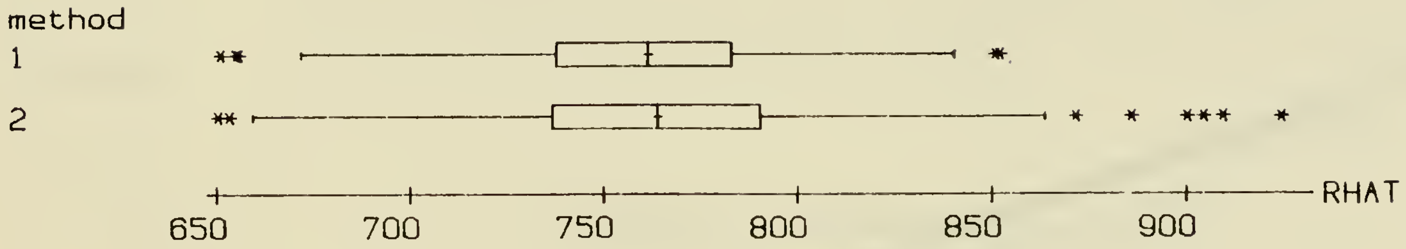
	SYMMETRIC PAIR		ASYMMETRIC PAIR	
Measures of Location	$\hat{N}$	$\hat{R}$	$\hat{N}$	$\hat{R}$
Median	146	764	145	762
Mean	151	765	148	760
Lower quartile	134	737	134	738
Upper quartile	163	791	160	783
Minimum	104	651	108	651
Maximum	255	925	232	852
Measures of Spread				
Standard Deviation	22.3	41.4	19.3	33.1
Interquartile Range	29	54	26	45
Range	151	274	124	201

True value of N is 170

True Value of R is 818

FIGURE 5.1.

PARALLEL BOXPLOTS FOR $\hat{R}$-VALUES[*]

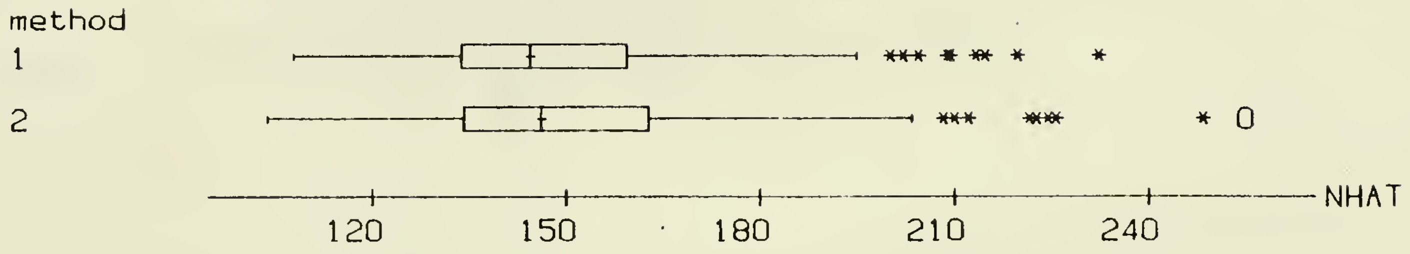


[*] 1 denotes the boxplot for asymmetric pair,

2 denotes the boxplot for the symmetric pair.

FIGURE 5.2.

PARALLEL BOXPLOTS FOR $\hat{N}$-VALUES*



*1 denotes the boxplot for asymmetric pair,

2 denotes the boxplot for the symmetric pair.

FIGURE 5.3.

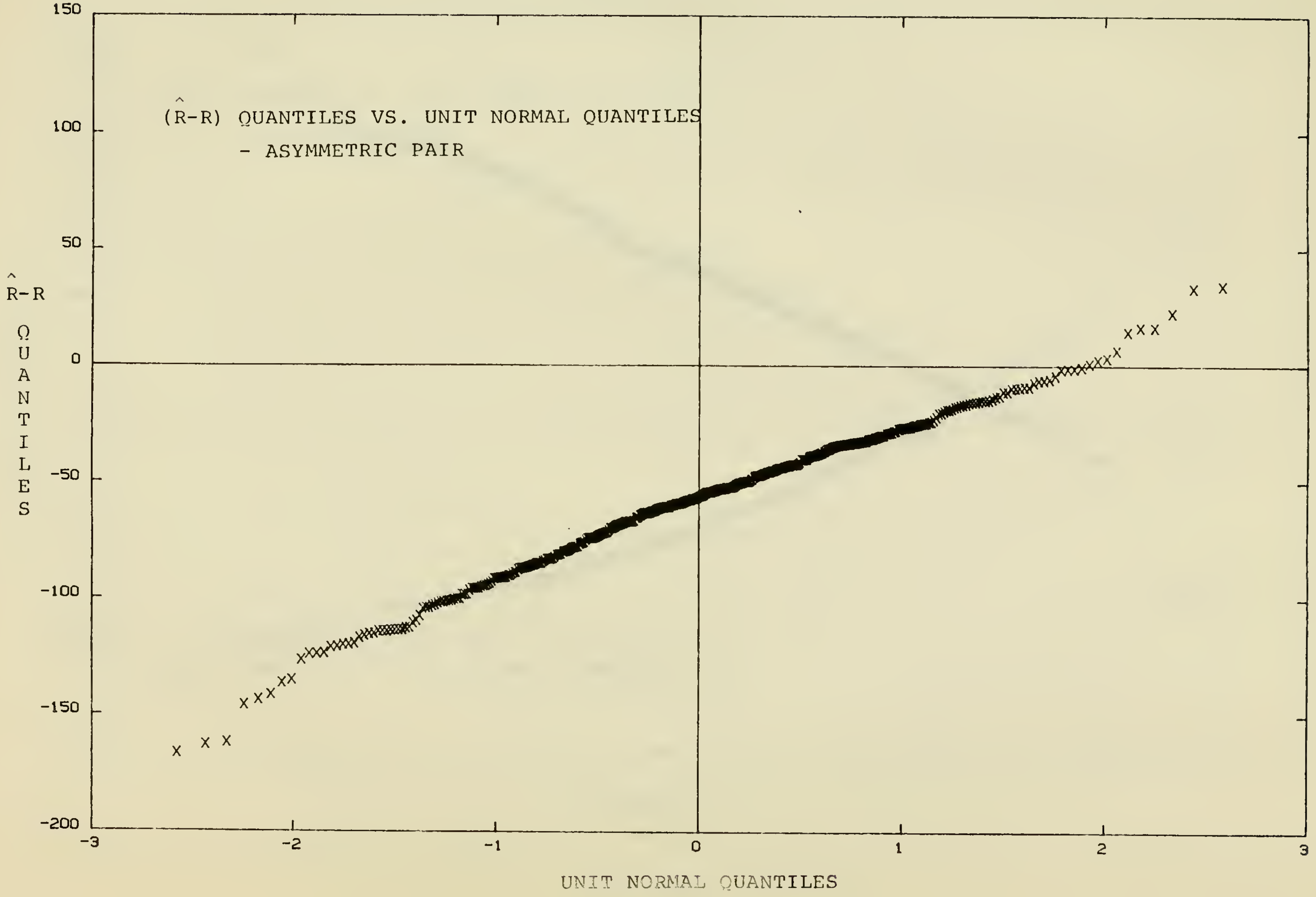

(R̂-R) QUANTILES VS. UNIT NORMAL QUANTILES
- ASYMMETRIC PAIR
R̂-R QUANTILES
UNIT NORMAL QUANTILES
150
100
50
0
-50
-100
-150
-200
-3
-2
-1
0
1
2
3

FIGURE 5.4.

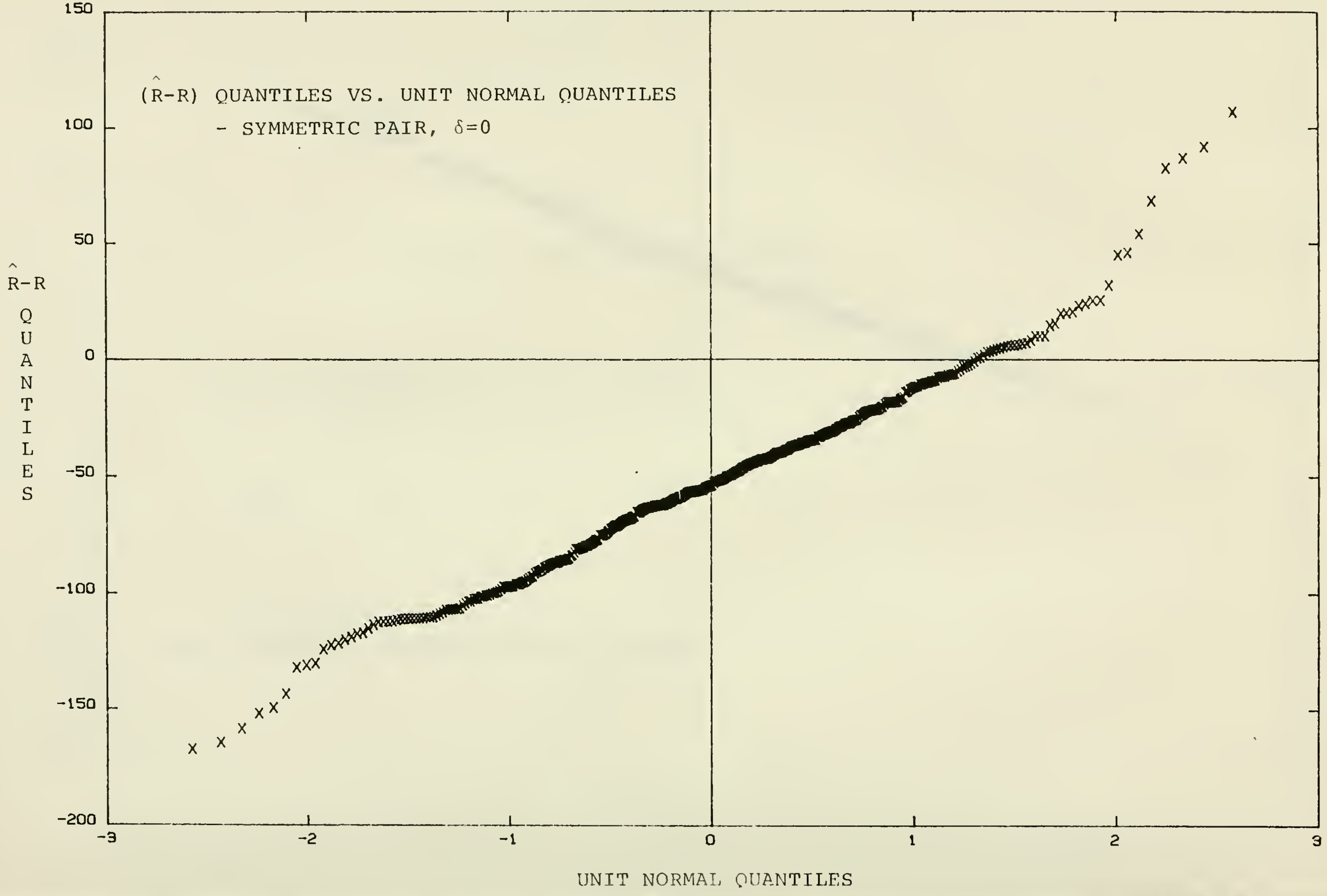

$(\hat{R}-R)$ QUANTILES VS. UNIT NORMAL QUANTILES
- SYMMETRIC PAIR, $\delta=0$
$\hat{R}-R$ QUANTILES
150
100
50
0
-50
-100
-150
-200
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

FIGURE 5.5.

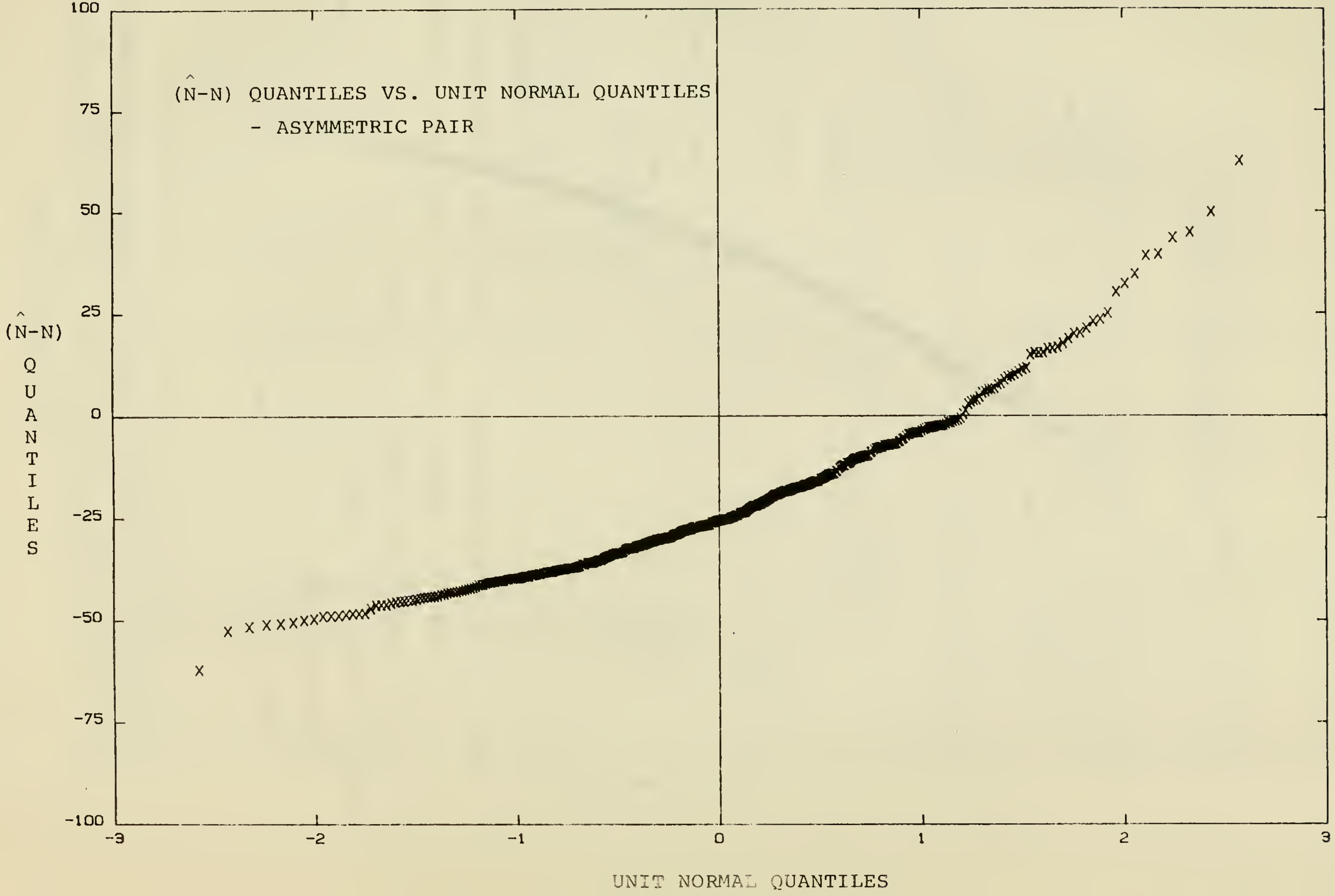
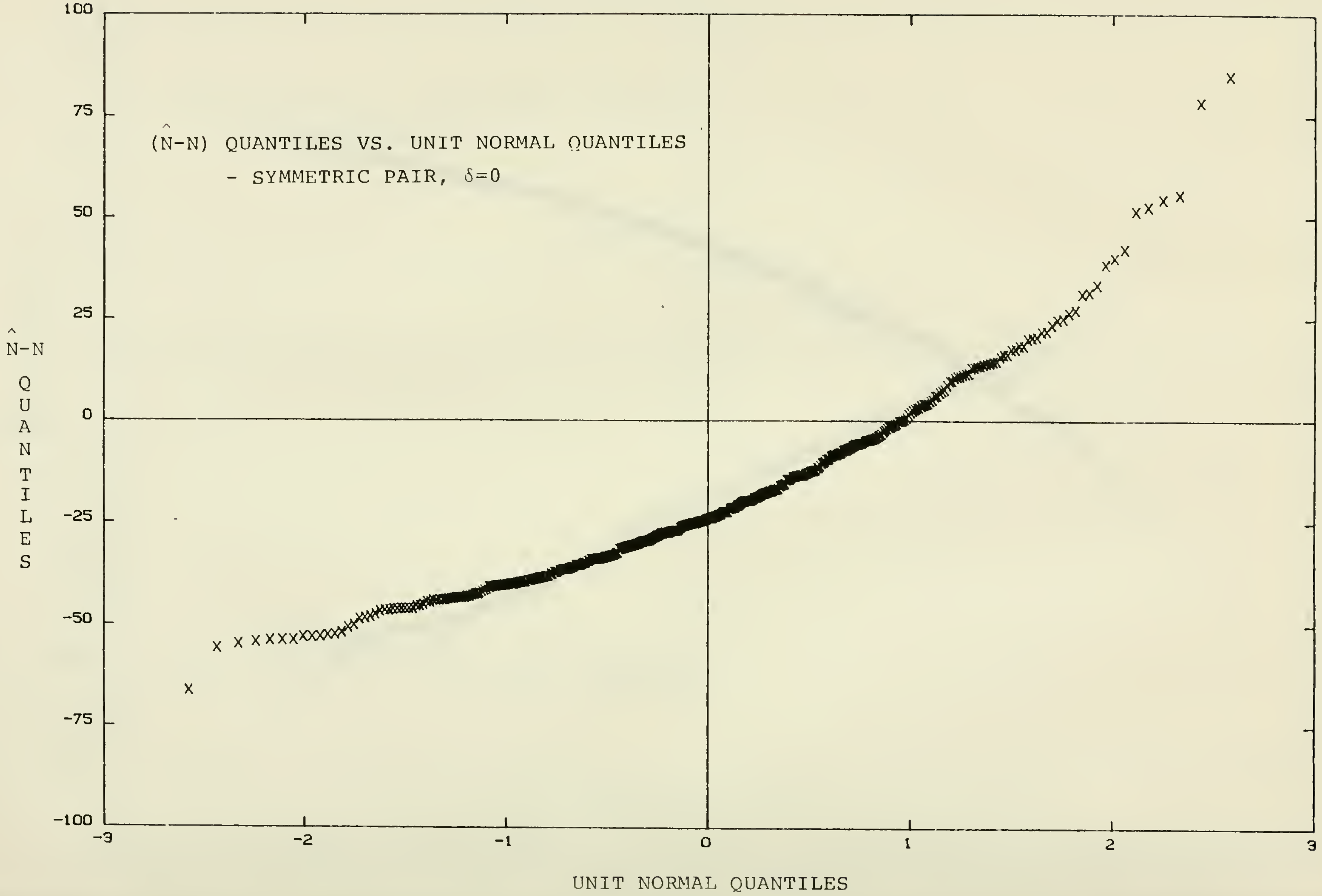

$(\hat{N}-N)$ QUANTILES VS. UNIT NORMAL QUANTILES
- ASYMMETRIC PAIR
$(\hat{N}-N)$ QUANTILES
100
75
50
25
0
-25
-50
-75
-100
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

FIGURE 5.6.


(N̂-N) QUANTILES VS. UNIT NORMAL QUANTILES
- SYMMETRIC PAIR, δ=0
N̂-N QUANTILES
100
75
50
25
0
-25
-50
-75
-100
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

5.2 SAMPLING PROPERTIES OF SOLUTIONS TO ASYMMETRIC AND SYMMETRIC PAIRS

Properties of the sampling distribution of solutions to symmetric and to asymmetric pairs are displayed in a fashion similar to that for $\hat{N}$ and for $\hat{R}$ values. Table 5.2 presents some summary statistics.

Examination of parallel boxplots for α and for β values (Figures 5.7 and 5.8) shows that

(I) The distribution of β values for the symmetric pair is spread over a much larger range than corresponding β values for the asymmetric pair.

While, in accord with asymptotic theory, quantiles of α values and quantiles of β values plot as straight lines against unit normal quantiles throughout a range of -2.0 to +2.5 (Figures 5.9 to 5.13),

(II) Extreme left tails (.02-.025 fractiles and smaller) of distributions of α and of β values clearly deviate from normality.

This finding accords with deviations from normality seen in the right tails of sampling distributions for $\hat{N}$ and for $\hat{R}$ values. For α values generated by the asymmetric pair, a quantile transformation $\exp\{\alpha/40\}$ of α (Figure 5.12) appears slightly closer to normal than quantiles of α (Figure 5.11).

Figure 5.14 is a scatterplot for symmetric vs. asymmetric pair α values and Figure 5.15 a similar scatterplot for β values. That α values generated by symmetric and by asymmetric pairs are positively correlated (sample correlation = .652) is obvious. The same is true for β values (sample correlation = .751).

More interesting is the heteroscedasticity displayed by symmetric vs asymmetric pair α values. This feature of Figure 5.14 can be captured by fitting the ratio $\alpha_S^{(j)}/\alpha_A^{(j)}$ of the j^{th} monte carloed sample symmetric pair α

value, $\alpha_S^{(j)}$, to the corresponding asymmetric pair α value, $\alpha_A^{(j)}$, with a model of the form "constant plus error". To wit,

$$\frac{\alpha_S^{(j)}}{\alpha_A^{(j)}} = \beta + \varepsilon^{(j)}, \qquad j=1,2,\ldots,400. \tag{5.1}$$

We may interpret (5.1) as a variance stabilizing transformation of a linear model of the form $\alpha_S = \beta\alpha_A + \eta$ with observed values of error term $\eta = \alpha_A \varepsilon$ exhibiting increasing spread as the value of α_A increases. The behavior of the equation systems leads us to expect that $\beta = 1.0$. and that when $\beta = 1.0$ the empirical distribution of $\varepsilon^{(j)}$ values is approximately normal with mean zero.

Our expectations are borne out: the mean $\frac{1}{400} \sum_{j=1}^{400} \alpha_S^{(j)}/\alpha_A^{(j)} = \hat{\beta} = 1.004$ and the graph in Figure 5.16 of quantiles of the empirical distribution of residuals $\hat{\varepsilon}^{(j)} = [\alpha_S^{(j)}/\alpha_A^{(j)}] - \hat{\beta}$ versus unit normal quantiles appears reasonably close to a straight line.

A slight improvement is afforded by fitting a model

$$\frac{\alpha_S^{(j)}}{\alpha_A^{(j)}} = \gamma_1 + \gamma_2 Z^{(j)} + w^{(j)}, \qquad j=1,2,\ldots,400 \tag{5.2}$$

with $\bar{\alpha}_A = \frac{1}{400} \sum_{j=1}^{n} \alpha_A^{(j)}$ and $Z^{(j)} = \alpha_A^{(j)} - \bar{\alpha}_A$. Upon doing so we find $\hat{\gamma}_1 = 1.004$ and $\hat{\gamma}_2 = .00667$ (t statistic = -7.2, standard error .00093), but the addition of the linear term in Z explains only about 11.5% of the total variance of observed values of the ratio of α values. A graph of quantities of the empirical distribution of residuals $\hat{w}^{(j)}$ versus unit normal quantities given in Figure 5.17 shows this slight improvement.

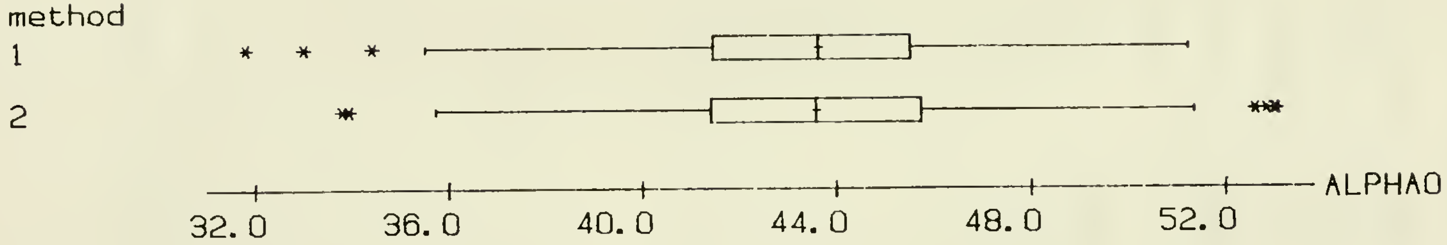
TABLE 5.2

SUMMARY MEASURES FOR MONTE CARLOED SAMPLING DISTRIBUTIONS OF ESTIMATORS α AND β

	SYMMETRIC PAIR		ASYMMETRIC PAIR	
<u>Measures of Location</u>	$\underline{\alpha}$	$\underline{\beta}$	$\underline{\alpha}$	$\underline{\beta}$
Median	43.59	42.40	43.65	43.97
Mean	43.60	42.69	43.49	43.83
Lower quartile	41.42	39.02	41.48	42.94
Upper quartile	45.75	46.68	45.54	44.90
Minimum	33.80	22.46	31.82	37.88
Maximum	53.05	58.13	51.24	47.34
<u>Measures of Spread</u>				
Standard Deviation	3.25	5.97	2.99	1.50
Interquartile Range	4.33	7.66	4.06	1.86
Range	19.25	25.67	9.42	9.45

FIGURE 5.7.

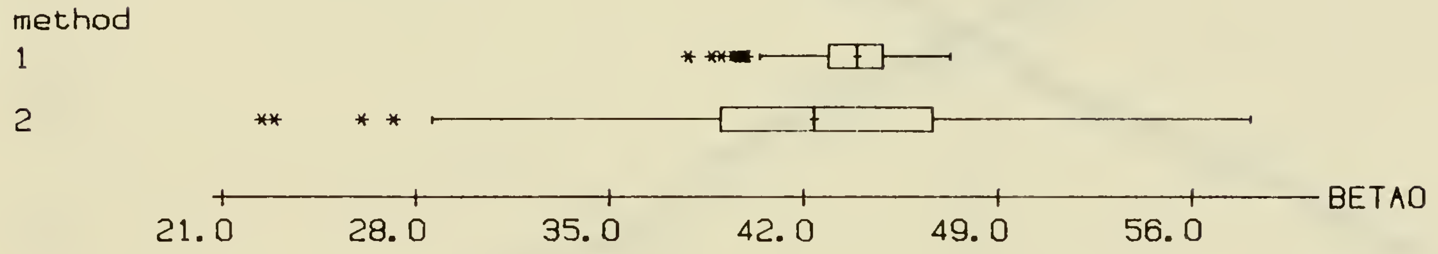
PARALLEL BOXPLOTS OF α-VALUES*



*1 denotes the boxplot for the asymmetric pair,
2 denotes the boxplot for the symmetric pair.

FIGURE 5.8.

PARALLEL BOXPLOTS OF β-VALUES*



*1 denotes the boxplot for the asymmetric pair,

2 denotes the boxplot for the symmetric pair.

FIGURE 5.9.

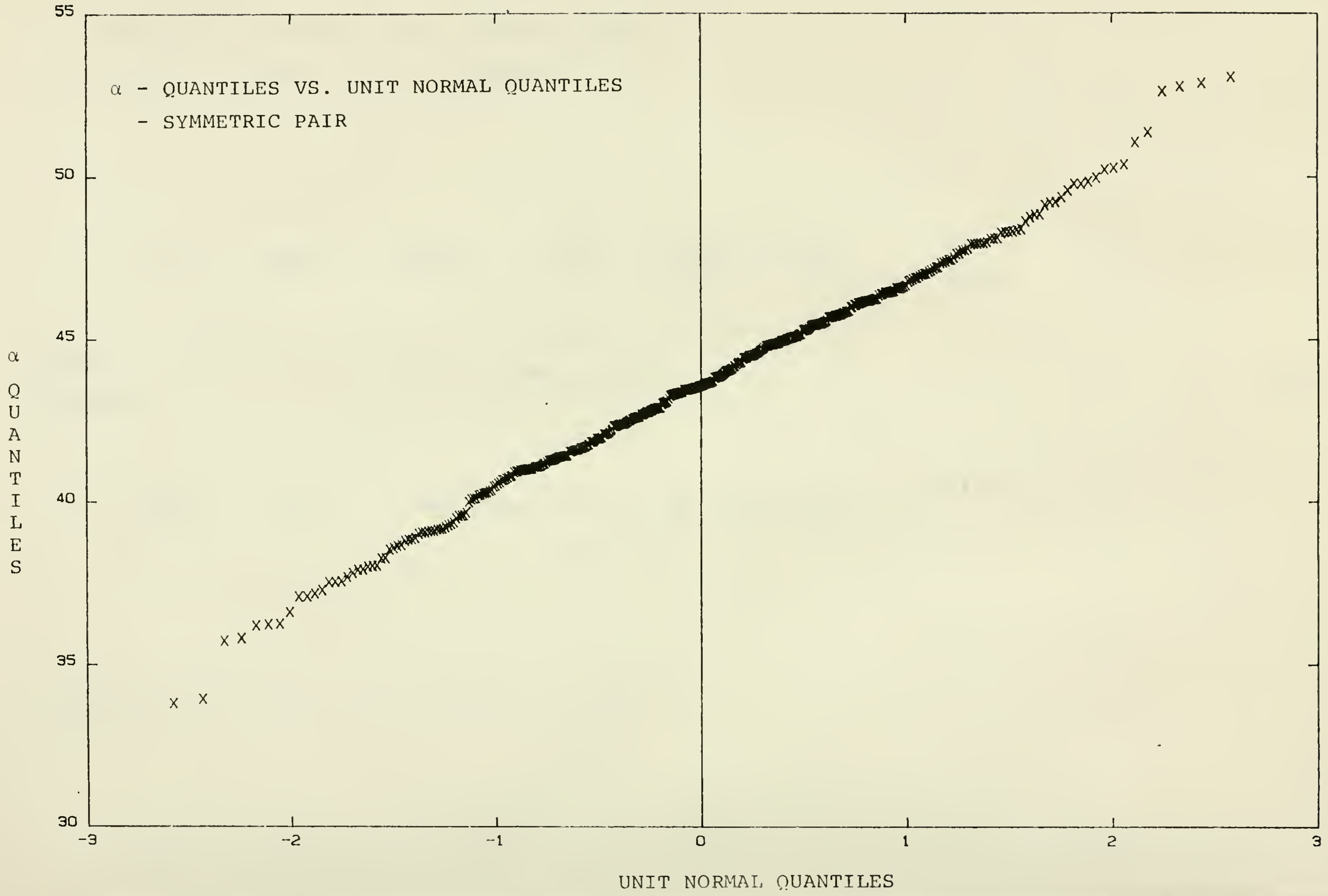

α - QUANTILES VS. UNIT NORMAL QUANTILES
- SYMMETRIC PAIR
α QUANTILES
UNIT NORMAL QUANTILES
55
50
45
40
35
30
-3
-2
-1
0
1
2
3

FIGURE 5.10.

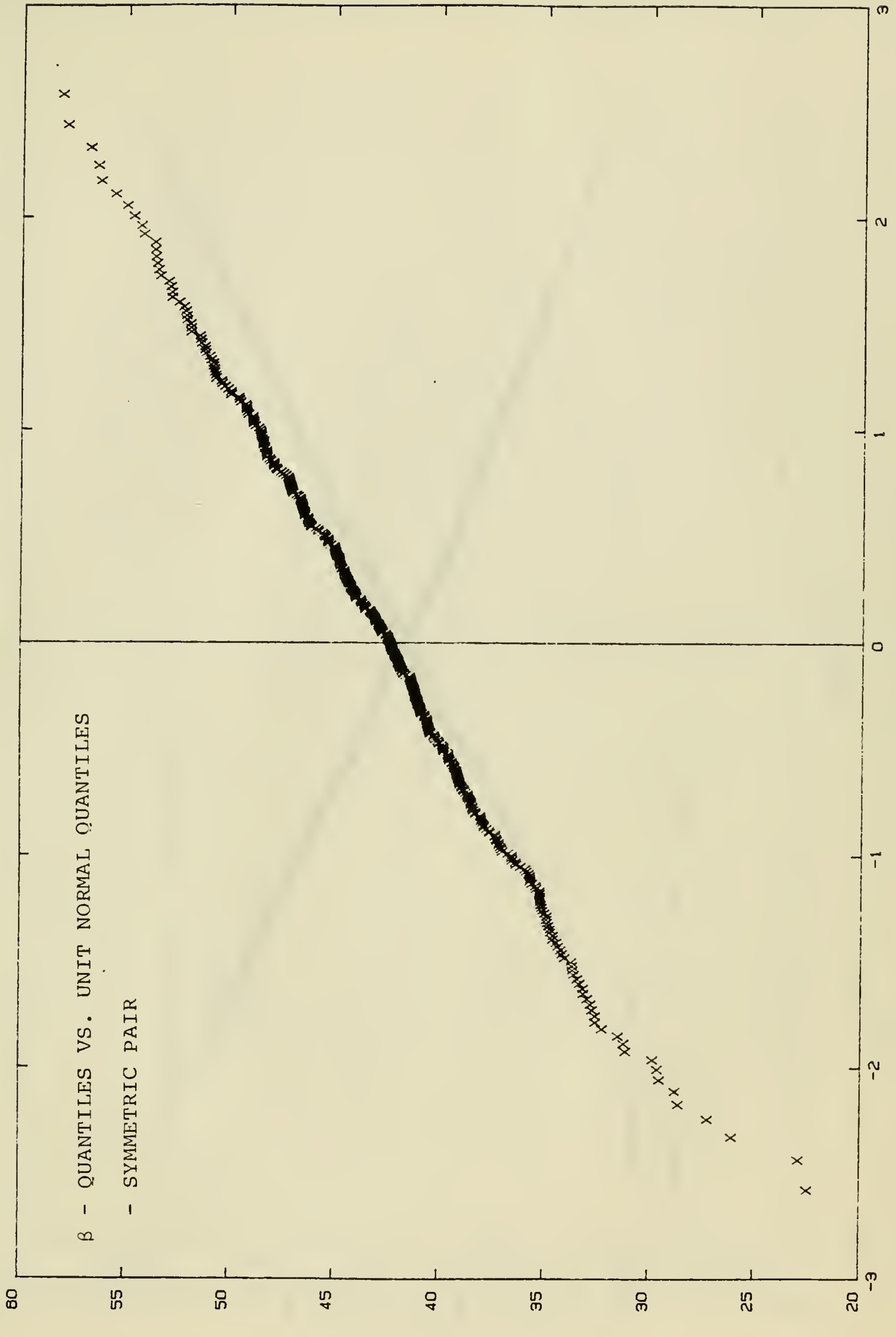

β - QUANTILES VS. UNIT NORMAL QUANTILES
- SYMMETRIC PAIR
β QUANTILES
UNIT NORMAL QUANTILES
60
55
50
45
40
35
30
25
20
-3
-2
-1
0
1
2
3

FIGURE 5.11.

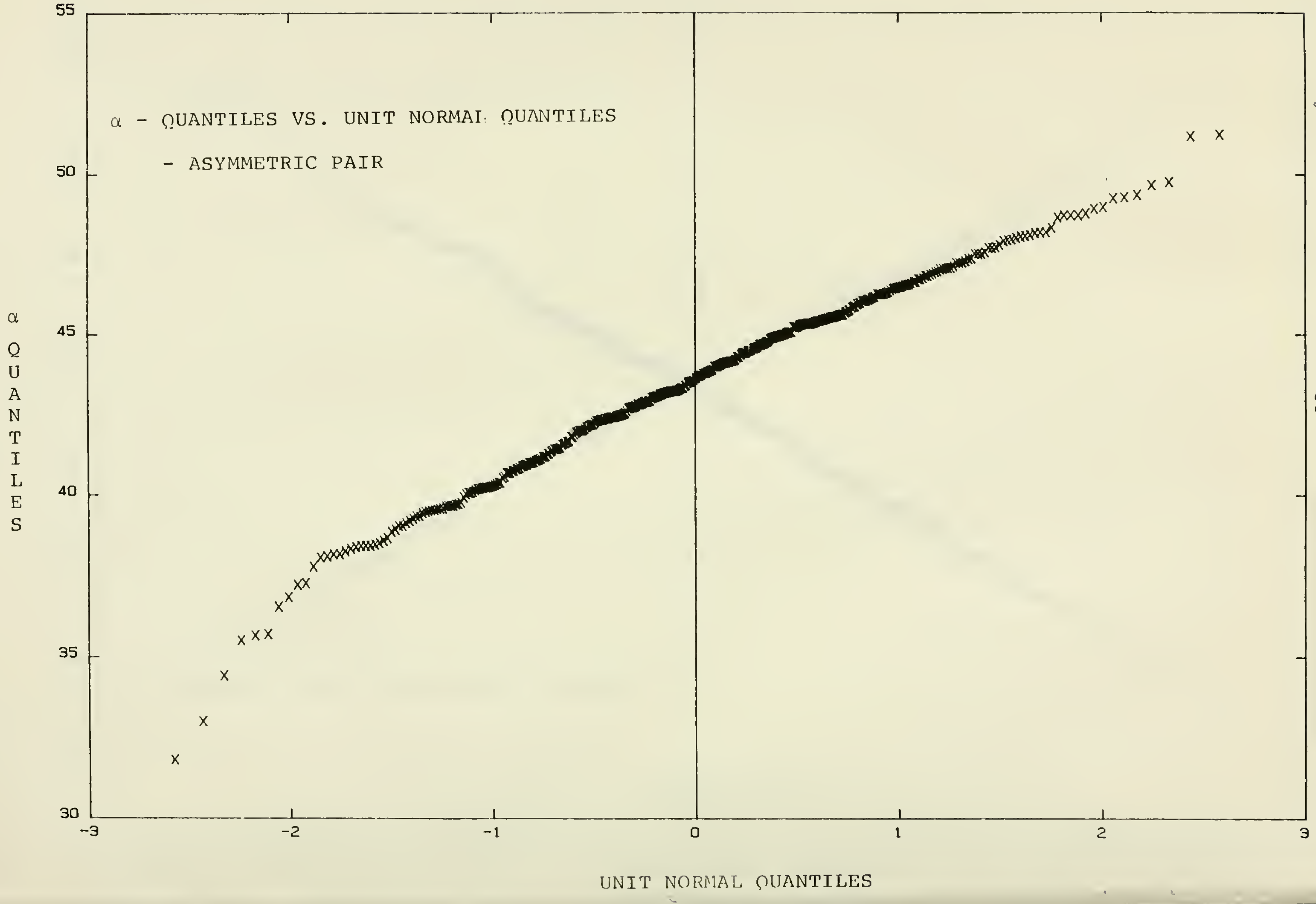

α - QUANTILES VS. UNIT NORMAL QUANTILES
- ASYMMETRIC PAIR
α QUANTILES
55
50
45
40
35
30
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

FIGURE 5.12.

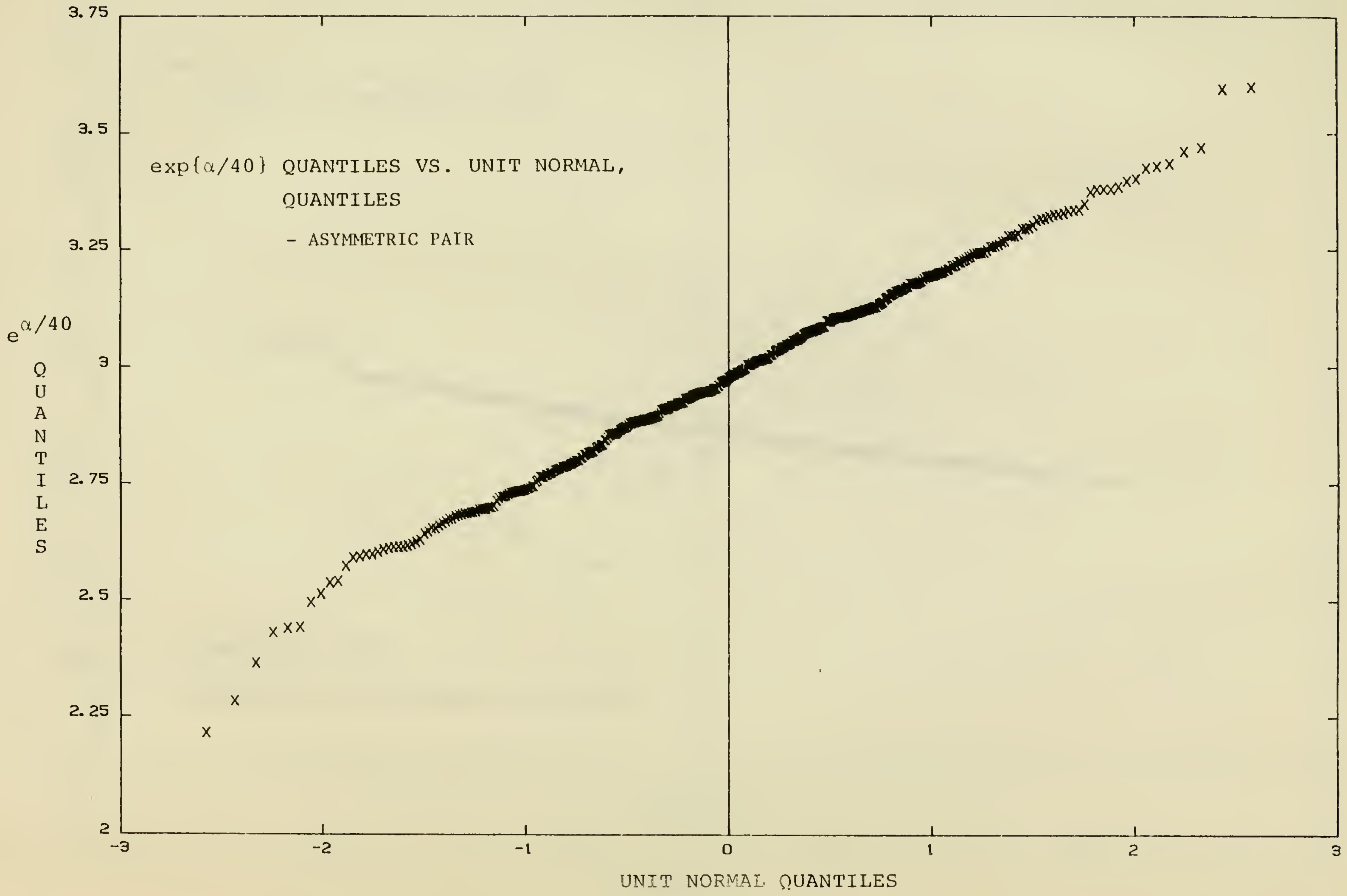

exp{α/40} QUANTILES VS. UNIT NORMAL,
QUANTILES
- ASYMMETRIC PAIR
$e^{\alpha/40}$
QUANTILES
3.75
3.5
3.25
3
2.75
2.5
2.25
2
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

FIGURE 5.13.

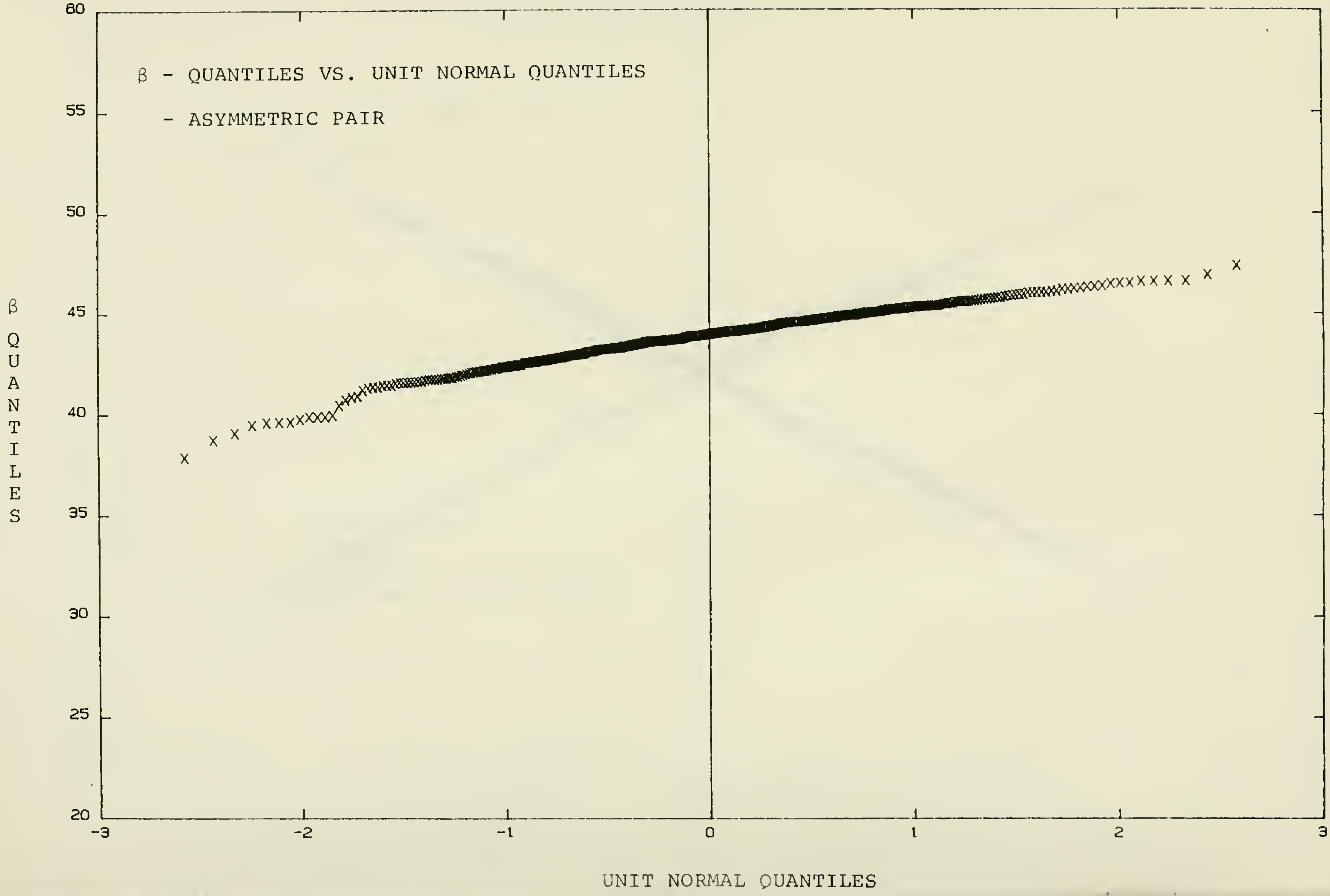
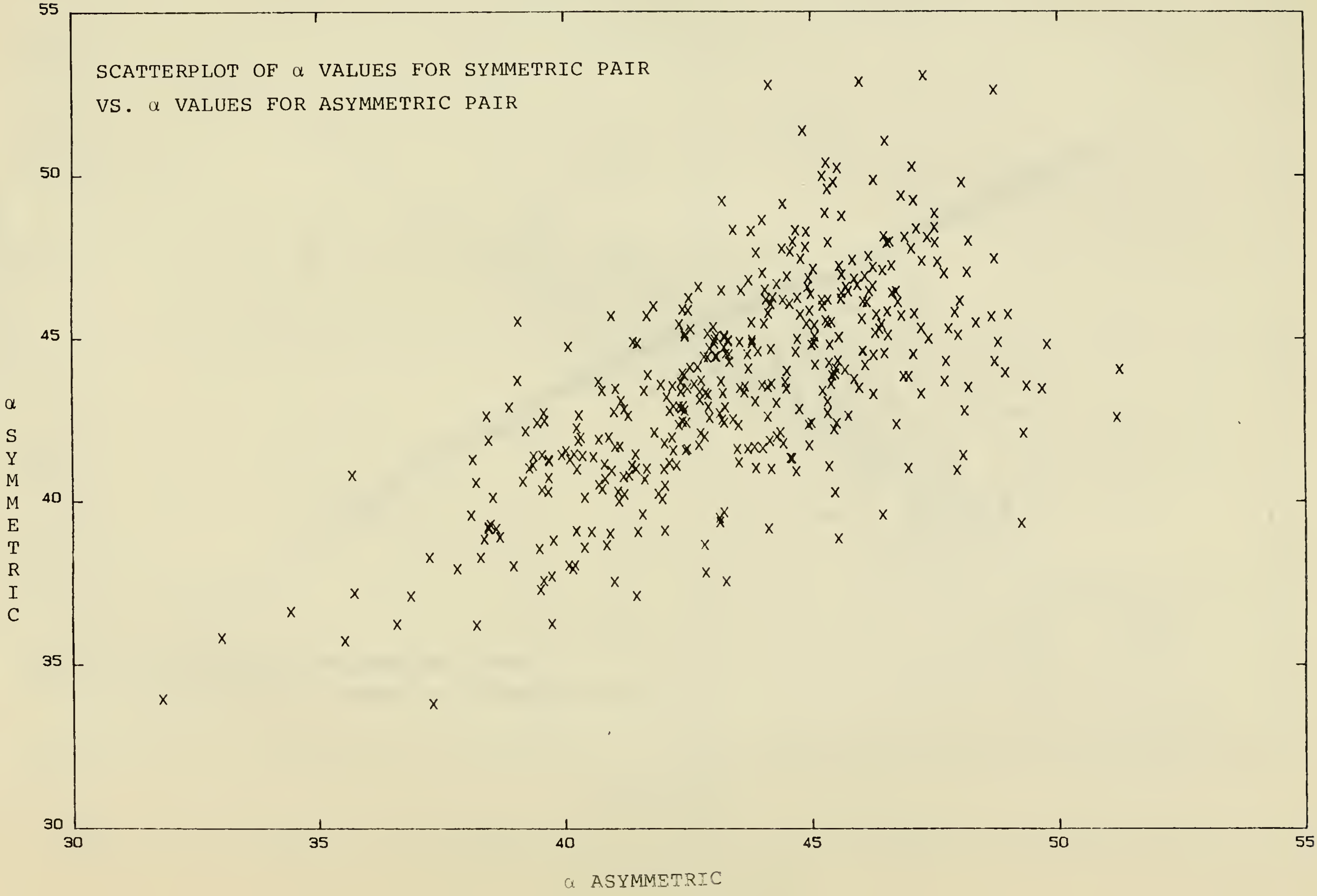
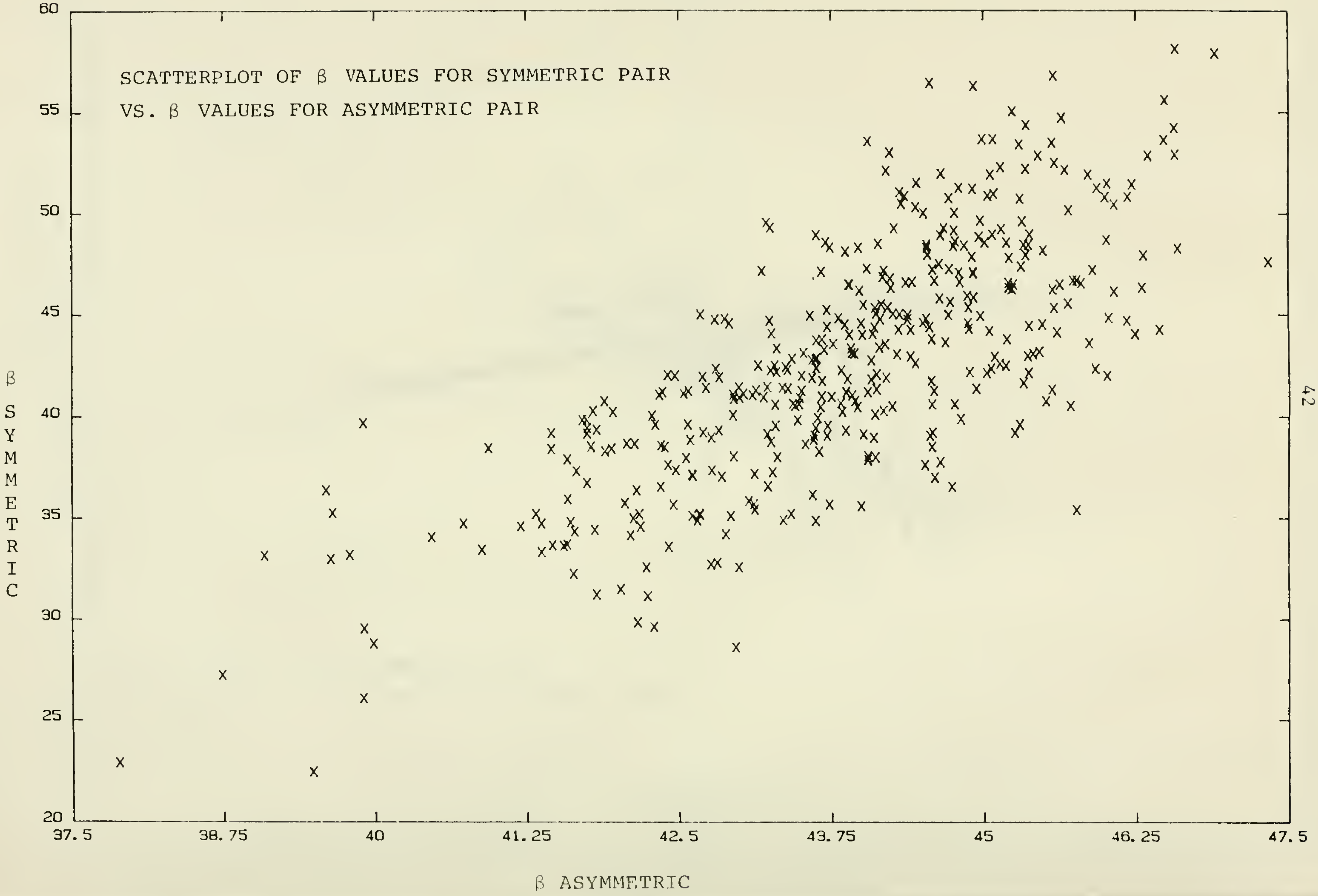

β - QUANTILES VS. UNIT NORMAL QUANTILES
- ASYMMETRIC PAIR
β QUANTILES
60
55
50
45
40
35
30
25
20
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

FIGURE 5.14.



SCATTERPLOT OF α VALUES FOR SYMMETRIC PAIR
VS. α VALUES FOR ASYMMETRIC PAIR

FIGURE 5.15.



SCATTERPLOT OF β VALUES FOR SYMMETRIC PAIR
VS. β VALUES FOR ASYMMETRIC PAIR
β SYMMETRIC
60
55
50
45
40
35
30
25
20
37.5
38.75
40
41.25
42.5
43.75
45
46.25
47.5
β ASYMMETRIC

FIGURE 5.16.

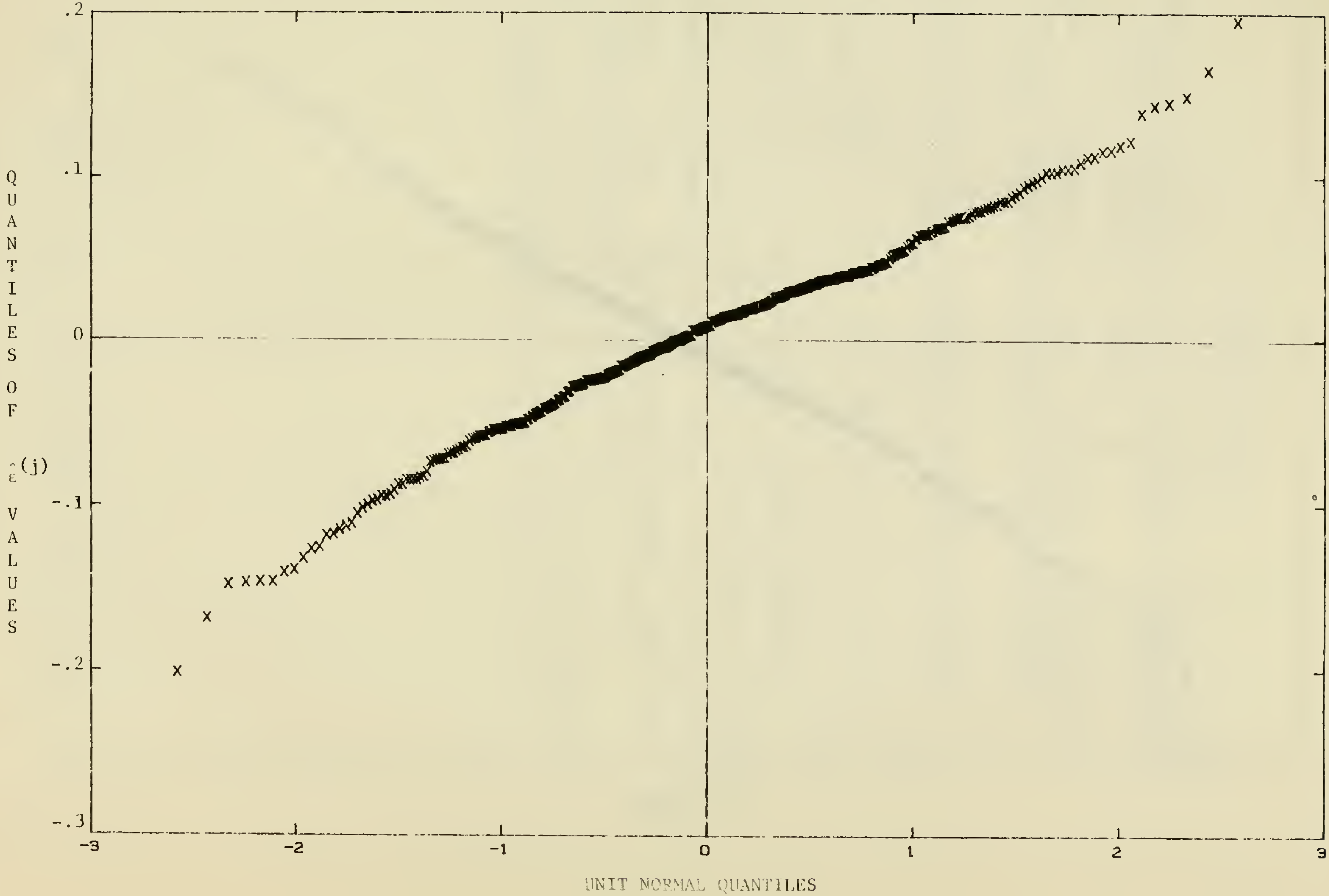
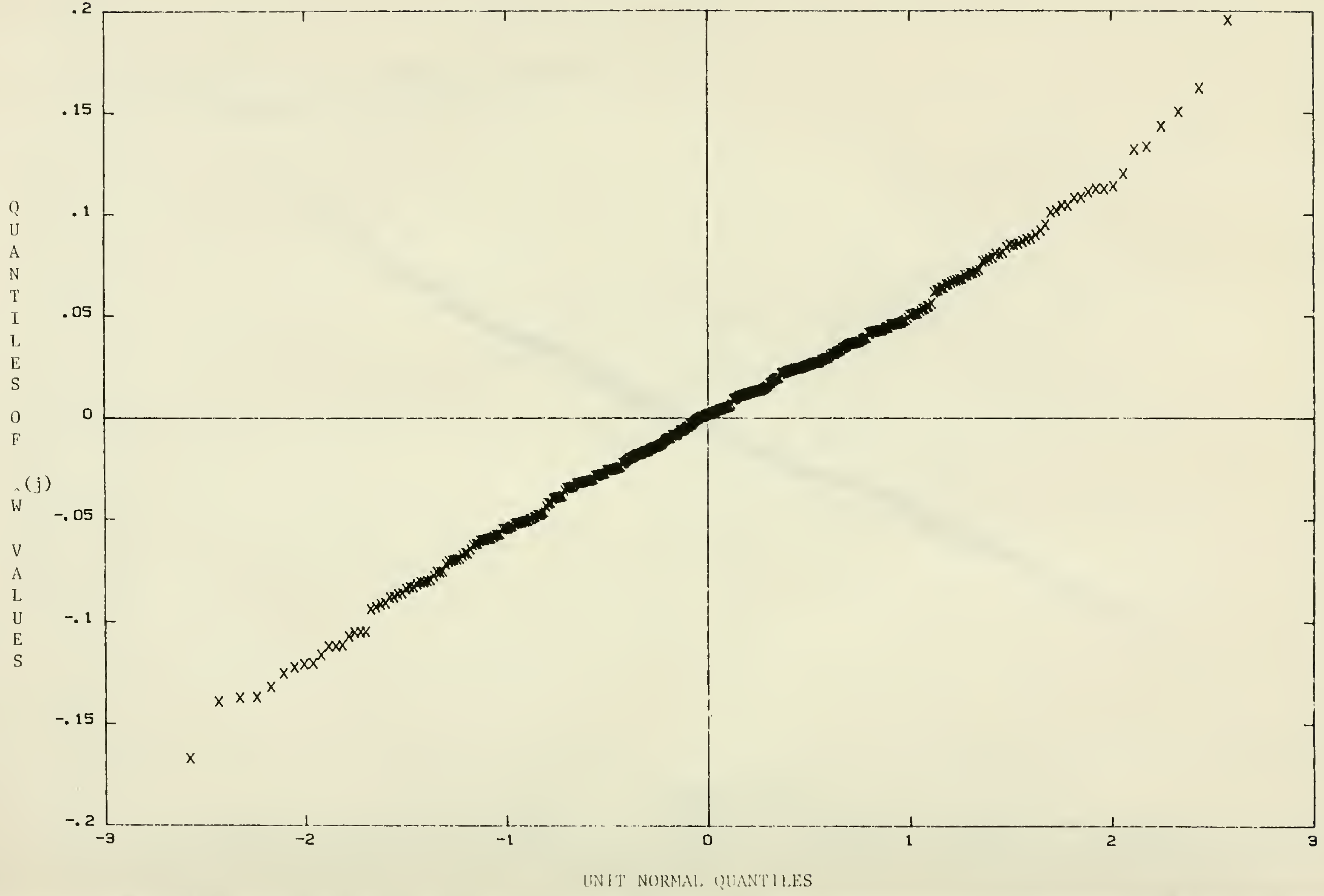

.2
.1
0
-.1
-.2
-.3
QUANTILES OF $\hat{\varepsilon}(j)$ VALUES
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

FIGURE 5.17.



QUANTILES OF $\hat{W}^{(j)}$ VALUES
.2
.15
.1
.05
0
-.05
-.1
-.15
-.2
-3
-2
-1
0
1
2
3
UNIT NORMAL QUANTILES

Andreatta, G., and Kaufman, G.M. (1985). "Estimation of Finite Population Properties When Sampling is Without Replacement and Proportional to Magnitude. M.I.T. Working Paper MIT-EL80-027WP. (Massachusetts Institute of Technology: Cambridge, MA).

Barouch E., and Kaufman, G.M. (1976). "Estimation of Undiscovered Oil and Gas," in Proceedings of Symposia in Applied Mathematics, Vol. XXI, "Mathematical Aspects of Production and Distribution of Energy," p. 77.

DuMouchel, W.H. (1970). "An Analysis of the December 1969 Selective Service Draft Lottery," Department of Statistics, University of California, Berkeley, California.

Gordon, L. (1981). "Successive Sampling in Large Finite Populations," Ann. Statist. Vol. 11, No. 2, pp. 702-706.

Gordon L. (1983). "Estimation for Large Successive Samples with Unknown Inclusion Probabilities," (submitted to Annals of Statistics).

Horvitz, D.G., and Thompson, D.J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe." J. Amer. Statist. Assoc., Vol. 47, pp. 663-685.

Littlewood, B. (1981). "Stochastic Reliability-growth: A Model for Fault-removal in Computer Programs and Hardward Designs. IEEE Trans. Reliability, R-30, pp. 313-320.

Meisner, J. and Demirmen, F. (1980). "The Creaming Method: A Bayesian Procedure to Forecast Future Oil and Gas Discoveries in Nature Exploration Provinces, JRSS, Series A, Vol. 143.

Barcode is on back cover.