

A Robust Univariate Procedure for Testing Symmetry and Normality

by

Walid S. Al-Sabah

A Thesis Presented to the

FACULTY OF THE COLLEGE OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

MATHEMATICS

June, 1978

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

A ROBUST UNIVARIATE PROCEDURE FOR TESTING
SYMMETRY AND NORMALITY

by

WALID S. AL-SABAH

A Thesis Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF PETROLEUM AND MINERALS
DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE IN MATHEMATICS

THE LIBRARY
University of Petroleum & Minerals
DHAHRAN - SAUDI ARABIA

June, 1978

UMI Number: 1381101

UMI Microform 1381101
Copyright 1996, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

UNIVERSITY OF PETROLEUM AND MINERALS
DHAHRAN, SAUDI ARABIA

THE GRADUATE SCHOOL

This thesis, written by

WALID S. AL-SABAH

under the direction of his Thesis Committee, and approved
by all its members, has been presented to and accepted by
the Dean of the Graduate School, in partial fulfilment of
the requirements for the degree of

MASTER OF SCIENCE IN MATHEMATICS

Hashim O. Yamani

Dr. Hashim A. Yamani
Dean of the Graduate School

Date May 23, 1978

A. H. Al-Moajil
Dr. Abdullah H. Al-Moajil
Department Chairman

Thesis Committee

M. David Lloyd
Dr. M. David Lloyd
Chairman

Munir Ahmad
Dr. Munir Ahmad
Member

John Stroyls
Dr. John Stroyls
Member

To Khalil Kutran with whom I started
to learn statistics

ACKNOWLEDGEMENT

I deeply thank my adviser Dr. Farouk Kazim whose constant guidance and encouragement made this work possible.

TABLE OF CONTENTS

	Page
ABSTRACT	1
INTRODUCTION	2
CHAPTER I: THE PROBLEM UNDER CONSIDERATION	4
CHAPTER II: THEORETICAL DERIVATIONS	9
CHAPTER III: GENERATION OF RANDOM SAMPLES	26
CHAPTER IV: MONTE CARLO RESULTS	33
CHAPTER V: ANALYSIS OF RESULTS	36
CHAPTER VI: SUMMARY AND CONCLUSIONS	40
APPENDIX: TABLE AND COMPUTER OUTPUT	42
REFERENCES	104

ABSTRACT

A robust univariate test of symmetry based on the correlation between the sample mean and the sample variance is presented. This test is sensitive to skewed transformations of symmetric random variables. Based on a certain characterization of the normal distribution, the procedure is extended to a test of normality.

INTRODUCTION

This work consists essentially of the development and subsequent discussion of a robust univariate procedure for testing symmetry and normality in random samples. The practical problem of testing symmetry has recently assumed a great deal of importance on account of the extensive research done on the robust estimation of the location parameters of symmetric distributions [Andrews et al. 1972]. The test of symmetry must necessarily meet two basic requirements to be acceptable, viz., it must exhibit a certain degree of robustness of validity and robustness of efficiency (or performance). The test must also be sensitive to skewed transformations of symmetric random variables since it appears that these are the most important alternatives to the hypothesis of symmetry. The problem of testing normality in random samples has been investigated extensively e.g. [Shapiro, Wilk and Chen 1968]. In this thesis, based on a fundamental characterization of the normal distribution, the procedure for testing symmetry is extended to a test of normality.

The structure of the thesis is as follows:

In Chapter 1, we explain in detail the problem under consideration and include a discussion of some of the fundamental notions of robustness. In Chapter 2, the motivation for the test is presented together with the theoretical derivations of some important results. The performance of the test is considered over a wide range of

symmetric and asymmetric distributions and the methods of generation of the random samples from each of these distributions are discussed in detail in Chapter 3. Chapter 4 is essentially a survey of the results of the computer study on the question of the robustness of the test. An in-depth analysis of the results is given in Chapter 5 and in particular, the results in the case of the normal distribution and close approximations to the normal are given special consideration. Chapter 6 concludes the thesis and gives a summary of the work done, some conclusions and recommendations for future research. The appendix contains the computer output of the results for the various distributions considered together with a few tables for the purpose of comparison.

The contribution in this thesis is, therefore, threefold, viz.,

- (i) to establish a univariate test of symmetry,
- (ii) to investigate the robustness of validity and robustness of efficiency of this test using a wide range of symmetric and asymmetric distributions,
- (iii) to extend this procedure to a test of normality.

CHAPTER I

THE PROBLEM UNDER CONSIDERATION

The problem of testing univariate symmetry has assumed a very important role in statistics because of the research currently being done on the estimation of location parameters of symmetric distributions. New estimators with very high efficiency have been suggested over a broad range of symmetric distributions [Andrews et al. 1972]. Since these estimators assume samples from symmetric distributions, a test of symmetry is necessary to investigate this assumption. Consequently, if we can establish symmetry, we may proceed to estimate the location parameters.

A test of symmetry is a decision rule which, on the basis of sample observations, decides whether the distribution of the sample is symmetric or not. A continuous univariate random variable X is symmetric about μ if and only if, for all x ,

$$F(x - \mu) = 1 - F(\mu - x)$$

where F is the cumulative distribution function and μ is the location parameter.

Various tests of symmetry have been proposed, some of which are: the Neyman-Pearson test (also known as the g_1 test); the scaled third moment test; Gupta's asymptotically nonparametric test [Hollander and Wolfe 1973]; and the adjusted V test, based on the

gaps between order statistics [Finch 1977].

The test presented in this thesis depends primarily on the relation between the sample mean \bar{X} and the sample variance S^2 . Thirty distributions are considered, fourteen of which are symmetric and sixteen asymmetric. The fourteen symmetric distributions are:

- (i) the normal (Gaussian);
- (ii) the double exponential (Laplace);
- (iii) the Cauchy;
- (iv) the slash, which is the ratio of a normal to an independent uniform (0, 1); and
- (v) Tukey's λ : for $\lambda = 1.0, 0.75, 0.5, 0.25, 0.0, -0.25, -0.5, -1.0, -1.5, -2.0$.

When $\lambda = 1.0$ we have the uniform (-1, 1) and when $\lambda = 0.0$ we have the logistic.

The sixteen asymmetric distributions are mainly skewed transformations of the normal distribution. These include:

- (i) $(1 - \alpha)Z + \alpha Z^2$ where Z is a normal (0, 1) random variable and $\alpha = 1.0, 0.5, 0.25, 0.1$.

Since Z^2 is a chi-squared random variable with one degree of freedom, we are therefore considering, in this case, a mixture of a normal and a chi-squared random variable.

- (ii) $\text{Exp}(\beta Z)$ where Z is a normal (0, 1) random variable and $\beta = 1.0, 0.5, 0.25, 0.1$.

This is the lognormal distribution.

- (iii) $Z \exp(\gamma Z)$ where Z is a normal $(0, 1)$ random variable and $\gamma = 1.0, 0.5, 0.25, 0.1$.

This random variable, like the normal, has infinite support.

- (iv) The Weibull distribution, for different values of the shape parameter c (3.0, 1.0, 0.5). $c = 1.0$ yields the exponential distribution.

- (v) The Pareto distribution.

A sample of fifty random observations is generated from each distribution. Every two distinct variates X_i and X_j are selected and the sum, $X_i + X_j = 2\bar{X}$, and the square of the difference, $(X_i - X_j)^2 = 2S^2$, are computed, thus exhausting all $\binom{50}{2}$ possible combinations. The resulting numbers are then sorted, first with respect to the sum and then with respect to the square of the difference. The sorted numbers are then divided equally into thirty five groups each containing thirty five pairs of numbers $(2\bar{X}, 2S^2)$. Thus we are actually considering a 35×35 contingency table, where each row sum and each column sum is thirty five. A chi-square test of independence based on fixed marginals is then performed [Conover 1971]. The square of the sample correlation coefficient, r^2 , is also computed.

The scatter plot of the points for each distribution is then carefully investigated. If the points are concentrated along a line, the distribution of the sample is definitely asymmetric. If the

concentration has a certain pattern but not along one line, the distribution is symmetric and more information about the tails of the distribution can be inferred, depending on the way the concentration appears. If there is a complete spread of points, in other words no concentration, the distribution is either symmetric or almost symmetric. We may note at this point that many special forms of asymmetric distributions resemble symmetric distributions, and in particular give extremely good approximations to the normal distribution [Johnson and Kotz 1970].

For any statistical test to be acceptable, the procedures used must be reasonably powerful, i.e., sensitive to specific factors under test; and robust, meaning that it is fairly insensitive to changes in extraneous factors not under test. We thus require robustness of validity, and robustness of performance or efficiency, i.e., stability of the power of the test. This test proves to be sensitive to skewed transformations of symmetric random variables, in other words, powerful. It also satisfies the criterion of validity since it can identify the symmetric distributions in a wide variety of cases.

The normal distribution is the most important single distribution in statistics. In the applications of probability theory we are often concerned with random variables that may be regarded as sums of a large number of independent components, each of which contributes only an insignificant amount of the total. Thus according to the "hypothesis of elementary errors" the total error committed at

a physical measurement is regarded as being produced by a large number of elementary errors, which add their effects. It often seems reasonable to apply a similar hypothesis to the variables which express the results of various kinds of industrial, biological and medical experiments. In all cases of this type we might expect the random variables considered to be at least approximately normally distributed. In fact, this expectation is largely corroborated by statistical experience.

Due to its importance in both physical and theoretical work, the normal distribution has been extensively studied and many characterizations have been established. The most fundamental of these characterizations is the following:

Let X_1, X_2, \dots, X_n be a random sample from some parent distribution. Then the sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ are independent if and only if the parent population is normal. Based on the computed χ^2 statistic and the r^2 value, normality can be established.

CHAPTER II

THEORETICAL DERIVATIONS

In Chapter I, a comprehensive idea about the test was given, but nothing was mentioned about the motivation for such a test. It is well-known that if the distribution function of a random variable X is symmetric about the mean, μ , then the odd central moments, if they exist, are all zero, where the r -th central moment of a random variable X is defined to be $E(X-\mu)^r$ and is denoted by μ_r . It will be shown, later on in this chapter, that the correlation coefficient between the sample mean \bar{X} and the sample variance S^2 is proportional to the third moment, μ_3 , and thus it is zero if the distribution function of the sample is symmetric. The converse is not true [Churchill 1946] but if the third moment is non-vanishing, as is the case with commonly encountered asymmetric distributions, \bar{X} and S^2 are necessarily correlated. On a scatter diagram, a high concentration of points along a line suggests that the two variables in question are highly linearly correlated. Since an asymmetric distribution function implies that \bar{X} and S^2 are correlated, the scatter plot of the two should show a high concentration along a line.

Now, we are ready to give an explicit expression for the correlation between \bar{X} and S^2 . Let X_1, X_2, \dots, X_n be a random sample of size n from a distribution function $F_X(x)$.

Let $E(X) = \mu_1' = \mu$ (mean);
 $E(X^2) = \mu_2' = \sigma^2 + \mu^2$;
 $E(X^3) = \mu_3'$;
 $E(X^4) = \mu_4'$ be the first four moments.

Let $E(X - \mu) = \mu_1 = 0$;
 $E(X - \mu)^2 = \mu_2 = \sigma^2$;
 $E(X - \mu)^3 = \mu_3$;
 $E(X - \mu)^4 = \mu_4$ be the first four central moments about μ .

Define $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

Then:

(i) $E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$
 $= \frac{1}{n} E\left(\sum_{i=1}^n X_i\right)$
 $= \frac{1}{n} \sum_{i=1}^n E(X_i)$
 $= \frac{1}{n} \sum_{i=1}^n \mu$
 $= \frac{1}{n} (n\mu)$
 $= \mu$

(ii) $\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$

$$\begin{aligned}
&= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \\
&= \frac{1}{n^2} (n\sigma^2) \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

(iii) To find $E(S^2)$; we first prove the following identity:

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

Proof:

$$\begin{aligned}
\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\
&= \sum_{i=1}^n \left[(X_i - \bar{X}) + (\bar{X} - \mu) \right]^2 \\
&= \sum_{i=1}^n \left[(X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2 \right] \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) + \sum_{i=1}^n (\bar{X} - \mu)^2 \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2
\end{aligned}$$

$$\begin{aligned}
\text{Hence, } E(S^2) &= E \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] \\
&= \frac{1}{n-1} E \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n-1} \left[\sum_{i=1}^n E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n-1} \left[\sum_{i=1}^n \sigma^2 - n \text{Var}(\bar{X}) \right] \\
&= \frac{1}{n-1} \left(n\sigma^2 - n \frac{\sigma^2}{n} \right) \\
&= \sigma^2
\end{aligned}$$

$$\begin{aligned}
\text{(iv) } \text{Var}(S^2) &= \text{Var} \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] \\
&= \frac{1}{(n-1)^2} \text{Var} \left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] \\
&= \frac{1}{(n-1)^2} \text{Var} \left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{(n-1)^2} \left[\text{Var} \left[\sum_{i=1}^n (X_i - \mu)^2 \right] + \text{Var} \left[n(\bar{X} - \mu)^2 \right] \right. \\
&\quad \left. - 2 \text{Cov} \left[\sum_{i=1}^n (X_i - \mu)^2, n(\bar{X} - \mu)^2 \right] \right]
\end{aligned}$$

$$\begin{aligned}
\text{Var} \left[\sum_{i=1}^n (X_i - \mu)^2 \right] &= n \text{Var}(X_1 - \mu)^2 \\
&= n \left[E(X_1 - \mu)^4 - \left[E(X_1 - \mu)^2 \right]^2 \right] \\
&= n(\mu_4 - \sigma^4) .
\end{aligned}$$

$$\begin{aligned}
\text{Var} \left[n(\bar{X} - \mu)^2 \right] &= \frac{1}{n^2} \text{Var} \left[n^2(\bar{X} - \mu)^2 \right] \\
&= \frac{1}{n^2} \text{Var}(n\bar{X} - n\mu)^2 \\
&= \frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n (X_i - \mu) \right]^2 \\
&= \frac{1}{n^2} \text{Var} \left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (X_i - \mu)(X_j - \mu) \right] \\
&= \frac{1}{n^2} \left[n \text{Var}(X_1 - \mu)^2 + \binom{n}{2} \text{Var} \left[2(X_i - \mu)(X_j - \mu) \right] \right] ; i \neq j \\
&= \frac{1}{n^2} \left[n(\mu_4 - \sigma^4) + 2n(n-1) \left[E(X_i - \mu)^2 (X_j - \mu)^2 \right. \right. \\
&\quad \left. \left. - \left[E(X_i - \mu)(X_j - \mu) \right]^2 \right] \right] ; i \neq j
\end{aligned}$$

$$= \frac{1}{n^2} \left[n(\mu_4 - \sigma^4) + 2n(n-1)\sigma^4 \right]$$

$$= \frac{\mu_4}{n} + \frac{2n-3}{n} \sigma^4.$$

$$\text{Cov} \left[\sum_{i=1}^n (X_i - \mu)^2, n(\bar{X} - \mu)^2 \right] = E \left[\left[\sum_{i=1}^n (X_i - \mu)^2 \right] \left[n(\bar{X} - \mu)^2 \right] \right]$$

$$- E \left[\sum_{i=1}^n (X_i - \mu)^2 \right] E \left[n(\bar{X} - \mu)^2 \right]$$

$$E \left[\left[\sum_{i=1}^n (X_i - \mu)^2 \right] \left[n(\bar{X} - \mu)^2 \right] \right] = \frac{1}{n} E \left[\left[\sum_{i=1}^n (X_i - \mu)^2 \right] \left[\sum_{i=1}^n (X_i - \mu) \right]^2 \right]$$

$$= \frac{1}{n} E \left[\left[\sum_{i=1}^n (X_i - \mu)^2 \right] \left[\sum_{i=1}^n (X_i - \mu)^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (X_i - \mu)(X_j - \mu) \right] \right]$$

$$= \frac{1}{n} E \left[\sum_{i=1}^n (X_i - \mu)^4 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (X_i - \mu)^2 (X_j - \mu)^2 + \right.$$

$$\left. \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n (X_i - \mu)^3 (X_j - \mu) + \frac{1}{2} \sum_{i=1}^n \sum_{\substack{j=1 \\ j \neq i}}^n \sum_{\substack{k=1 \\ k \neq j \\ k \neq i}}^n (X_i - \mu)^2 (X_j - \mu)(X_k - \mu) \right]$$

$$= \frac{1}{n} \left[n\mu_4 + n(n-1)\sigma^4 \right]$$

$$= \mu_4 + (n-1)\sigma^4$$

$$E \left[\sum_{i=1}^n (X_i - \mu)^2 \right] = \sum_{i=1}^n E(X_i - \mu)^2 = n\sigma^2.$$

$$E \left[n(\bar{X} - \mu)^2 \right] = nE(\bar{X} - \mu)^2 = \sigma^2$$

Hence,

$$\begin{aligned} \text{Cov} \left[\sum_{i=1}^n (X_i - \mu)^2, n(\bar{X} - \mu)^2 \right] &= \mu_4 + (n-1)\sigma^4 - n\sigma^4 \\ &= \mu_4 - \sigma^4. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{Var}(S^2) &= \frac{1}{(n-1)^2} \left[n\mu_4 - n\sigma^4 + \frac{\mu_4}{n} + \frac{2n-3}{n} \sigma^4 - 2\mu_4 + 2\sigma^4 \right] \\ &= \frac{1}{(n-1)^2} \left[\frac{\mu_4}{n} (n^2 - 2n + 1) - \frac{\sigma^4}{n} (n^2 - 4n + 3) \right] \\ &= \frac{\mu_4}{n} - \frac{n-3}{n(n-1)} \sigma^4 \end{aligned}$$

(v) $\text{Cov}(\bar{X}, S^2) = E(\bar{X}S^2) - E(\bar{X})E(S^2)$

$$\begin{aligned} E(\bar{X}S^2) &= E \left[\left[\frac{1}{n} \sum_{i=1}^n X_i \right] \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] \right] \\ &= \frac{1}{n(n-1)} E \left[\left[\sum_{i=1}^n X_i \right] \left[\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right] \right] \\ &= \frac{1}{n-1} E \left[X_1 \left[\sum_{i=1}^n X_i^2 \right] - \frac{1}{n} \left[\sum_{i=1}^n X_i \right]^2 \right] \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n-1} E \left[X_1 \left(\sum_{i=1}^n X_i^2 \right) \right] - \frac{1}{n(n-1)} E \left[X_1 \left(\sum_{i=1}^n X_i \right)^2 \right] \\
&= \frac{1}{n-1} E \left[X_1 (X_1^2 + X_2^2 + \dots + X_n^2) \right] \\
&\quad - \frac{1}{n(n-1)} E \left[X_1 (X_1 + X_2 + \dots + X_n) (X_1 + X_2 + \dots + X_n) \right] \\
&= \frac{1}{n-1} \left[\mu_3' + (n-1)\mu(\sigma^2 + \mu^2) \right] - \frac{1}{n(n-1)} E \left[X_1^2 (X_1 + X_2 + \dots + X_n) \right] \\
&\quad - \frac{n-1}{n(n-1)} E \left[X_1 X_2 (X_1 + X_2 + \dots + X_n) \right] \\
&= \frac{\mu_3'}{n-1} + \mu\sigma^2 + \mu^3 - \frac{1}{n(n-1)} \left[\mu_3' + (n-1)\mu(\sigma^2 + \mu^2) \right] \\
&\quad - \frac{1}{n} \left[2\mu(\sigma^2 + \mu^2) + (n-2)\mu^3 \right] \\
&= \frac{\mu_3'}{n-1} \left(1 - \frac{1}{n} \right) + \mu\sigma^2 \left(1 - \frac{1}{n} - \frac{2}{n} \right) + \mu^3 \left(1 - \frac{1}{n} - \frac{2}{n} - \frac{n-2}{n} \right) \\
&= \frac{\mu_3'}{n} + \frac{n-3}{n} \mu\sigma^2 - \frac{\mu^3}{n} \\
&= \frac{1}{n} (\mu_3' + (n-3)\mu\sigma^2 - \mu^3)
\end{aligned}$$

Hence, $\text{Cov}(\bar{X}, S^2) = \frac{\mu_3'}{n} + \mu\sigma^2 - \frac{3}{n} \mu\sigma^2 - \frac{\mu^3}{n} - \mu\sigma^2$

$$= \frac{1}{n} (\mu_3' - 3\mu\sigma^2 - \mu^3)$$

$$= \frac{\mu_3}{n}$$

Therefore

$$\begin{aligned} \text{Corr}(\bar{X}, S^2) &= \frac{\text{Cov}(\bar{X}, S^2)}{\sqrt{\text{Var}(\bar{X})} \sqrt{\text{Var}(S^2)}} \\ &= \frac{\mu_3}{\sigma \sqrt{\mu_4 - \frac{n-3}{n-1} \sigma^4}} \end{aligned}$$

In the estimation problem, it is desirable to obtain an indicator of the strength of the linear relationship between two variables X and Y which will be independent of their respective scales of measurement and an estimator of the population correlation coefficient. The measure of linear correlation commonly used in statistics is called the Pearson product moment coefficient of correlation. This statistic, denoted by the symbol r , is defined as follows:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sqrt{\left[n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right] \left[n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2 \right]}}$$

But it sometimes seems more meaningful to use r^2 , the coefficient of determination, since this gives a better indication of the strength of the relation between X and Y .

To investigate the independence of \bar{X} and S^2 , the test performed is the chi-square test with fixed marginal totals [Conover

1971]. The data are summarized in a 35×35 contingency table where each row sum and column sum is 35. The assumptions are that the observations are random and that each is placed into exactly one cell, each having the same probability of being placed into cell (i, j) as any other observation. The hypothesis tested is that \bar{X} and S^2 , which are in this case the components of each observation, are independent of each other. The test statistic T is given by

$$T = \sum_{i=1}^{35} \sum_{j=1}^{35} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the actual number of observations in cell (i, j) and E_{ij} is the expected number of observations in cell (i, j) . In our case, the test statistic reduces to

$$T = \sum_{i=1}^{35} \sum_{j=1}^{35} (O_{ij} - 1)^2 \text{ since } E_{ij} = 1 \text{ for all } i, j.$$

The decision rule is to reject the hypothesis if T exceeds the $1 - \alpha$ quantile of a chi-squared random variable with $34 \times 34 = 1156$ degrees of freedom.

In various statistical applications, it is often necessary to generate values of a given random variable X . To generate a value of a random variable X having a cumulative distribution function $F_X(\cdot)$, it suffices to generate a value of a random variable U that is uniformly distributed over the unit interval $(0, 1)$. Then by the "Probability Integral Transform," $X = F_X^{-1}(U)$ is a random

variable having $F_X(\cdot)$ as its distribution function.

Theorem: Probability Integral Transform

If X is a random variable with continuous cumulative distribution function $F_X(\cdot)$, then $U = F_X(X)$ is uniformly distributed over the interval $(0, 1)$. Conversely, if U is uniformly distributed over the interval $(0, 1)$, then $X = F_X^{-1}(U)$ has cumulative distribution function $F_X(\cdot)$.

Proof:

(The proof can be found in any standard text on mathematical statistics); e.g. [Mood, Graybill and Boes 1974].

Since $F_X(\cdot)$ is a non-decreasing function, the inverse function $F_X^{-1}(\cdot)$ may be defined for any value of u between 0 and 1 as $F_X^{-1}(u)$ is the smallest x satisfying $F_X(x) \geq u$.

$$\begin{aligned} P(U \leq u) &= P[F_X(X) \leq u] \\ &= P[X \leq F_X^{-1}(u)] \\ &= F_X[F_X^{-1}(u)] \\ &= u \quad \text{for } 0 < u < 1. \end{aligned}$$

$$\begin{aligned} \text{Conversely, } P(X \leq x) &= P[F_X^{-1}(U) \leq x] \\ &= P[U \leq F_X(x)] = F_X(x). \end{aligned}$$

We close this chapter by proving a fundamental characterization of the normal distribution.

Theorem: [Kagan, Linnik and Rao 1973].

Let $X^{(n)} = (X_1, X_2, \dots, X_n)$ be a random sample vector with $n \geq 2$. For the independence of $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, the first two sample moments, it is necessary and sufficient that the sample be normal.

Proof:

To prove the sufficiency of the condition, we prove that the degenerate normal vector $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$ is independent of \bar{X} by showing that the joint moment generating function of \bar{X} and $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$ is the product of their respective moment generating functions.

$$\begin{aligned} M(t, t_1, \dots, t_n) &= E \left[\exp \left[t\bar{X} + t_1(X_1 - \bar{X}) + \dots + t_n(X_n - \bar{X}) \right] \right] \\ &= E \left[\exp \left[\sum_{i=1}^n t_i X_i - \left(\sum_{i=1}^n t_i - t \right) \bar{X} \right] \right] \\ &= E \left[\exp \left[\sum_{i=1}^n \left(t_i - \frac{t_1 + \dots + t_n - t}{n} \right) X_i \right] \right] \end{aligned}$$

$$\begin{aligned}
&= E \left[\prod_{i=1}^n \exp \left[\frac{X_i(nt_i - n\bar{t} + t)}{n} \right] \right] \quad \text{where } \bar{t} = \frac{1}{n} \sum_{i=1}^n t_i \\
&= \prod_{i=1}^n E \left[\exp \left[\frac{X_i(nt_i - n\bar{t} + t)}{n} \right] \right] \\
&= \prod_{i=1}^n \exp \left[\frac{\mu}{n} \left[t + n(t_i - \bar{t}) \right] + \frac{\sigma^2}{2n^2} \left[t + n(t_i - \bar{t}) \right]^2 \right] \\
&= \exp \left[\frac{\mu}{n} \left[nt + n \sum_{i=1}^n (t_i - \bar{t}) \right] + \frac{\sigma^2}{2n^2} \sum_{i=1}^n \left[t + n(t_i - \bar{t}) \right]^2 \right] \\
&= \exp(\mu t) \exp \left[\frac{\sigma^2}{2n^2} \left[nt^2 + n^2 \sum_{i=1}^n (t_i - \bar{t})^2 \right] \right] \\
&= \exp(\mu t + \frac{\sigma^2}{2n^2} t^2) \exp \left[\frac{\sigma^2}{2} \sum_{i=1}^n (t_i - \bar{t})^2 \right] \\
&= M(t) \cdot M(t_1, \dots, t_n).
\end{aligned}$$

To prove the necessary condition, define the function f by:

$$f(t) = E \left[\exp(itX_1 - X_1^2) \right]$$

Clearly f is an entire function of t . In view of the independence of the X_j , we have

$$E \left[\exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right] = [f(t)]^n$$

Consider the expression

$$E \left[nS^2 \exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right]$$

In view of the independence of \bar{X} and S^2 , and of the identity:

$$\sum_{j=1}^n X_j^2 = n\bar{X}^2 + nS^2$$

$$\begin{aligned} E \left[nS^2 \exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right] &= E(nS^2 e^{-nS^2}) E \left[\exp \left[it \sum_{j=1}^n X_j - n\bar{X}^2 \right] \right] \\ &= a \cdot E \left[\exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right] \\ &= a \left[f(t) \right]^n \end{aligned}$$

where $a = \frac{E(nS^2 e^{-nS^2})}{E(e^{-nS^2})} > 0$

$$\begin{aligned} E \left[nS^2 \exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right] &= E \left(\sum_{j=1}^n X_j^2 \exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right) \\ &\quad - E \left[n\bar{X}^2 \exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right] \\ &= E \left[\left(\sum_{j=1}^n X_j^2 \right) \exp \left[\sum_{j=1}^n (itX_j - X_j^2) \right] \right] \end{aligned}$$

$$\begin{aligned}
& -\frac{1}{n} E \left[\left(\sum_{j=1}^n x_j^2 \right) \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& = E \left[\left(\sum_{j=1}^n x_j^2 \right) \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& \quad - \frac{1}{n} E \left[\left(\sum_{j=1}^n x_j^2 \right) \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& \quad - \frac{1}{n} E \left[\left(\sum_{j=1}^n \sum_{\substack{k=1 \\ k \neq j}}^n x_j x_k \right) \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& = \frac{n-1}{n} E \left[\left(\sum_{j=1}^n x_j^2 \right) \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& \quad - \frac{1}{n} E \left[\left(\sum_{j=1}^n \sum_{\substack{k=1 \\ k \neq j}}^n x_j x_k \right) \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& = (n-1) \left[f(t) \right]^{n-1} E \left[x_1^2 \exp \left[\sum_{j=1}^n (itX_j - x_j^2) \right] \right] \\
& \quad + (n-1) E \left[ix_1 \exp(itX_1 - x_1^2) \right] E \left[ix_2 \exp(itX_2 - x_2^2) \right] \\
& \quad \quad \quad E \left[\exp \left[\sum_{j=3}^n (itX_j - x_j^2) \right] \right] \\
& = -(n-1) [f(t)]^{n-1} f''(t) + (n-1) [f(t)]^{n-2} [f'(t)]^2
\end{aligned}$$

Hence,

$$-(n-1)[f(t)]^{n-1}f''(t) + (n-1)[f(t)]^{n-2}[f'(t)]^2 = a[f(t)]^n$$

or

$$[f(t)]^{n-1}f''(t) - [f(t)]^{n-2}[f'(t)]^2 = -a_1[f(t)]^n$$

Here $a_1 = \frac{a}{n-1} > 0$; f is an entire function with

$$f(0) = \int_{-\infty}^{\infty} e^{-u^2} dF(u) > 0$$

Hence in some neighborhood of the origin, we have

$$\frac{f''(t)}{f(t)} - \left[\frac{f'(t)}{f(t)} \right]^2 = -a_1$$

or $[nf(t)]'' = -a_1$

The solution of this differential equation is

$$f(t) = k \exp(i\gamma t - \frac{1}{2} a_1 t^2) \quad (1)$$

where $k = f(0)$ and $\gamma = -if'(0)$.

Since f is an entire function, it coincides with the right hand side of (1) on $(-\infty < t < \infty)$ if it does so on a neighborhood of the origin.

We claim that F is a normal distribution function.

Set $G(x) = \int_{-\infty}^x e^{-u^2} dF(u)$. Since $f = k \cdot f_0$ where f_0 is the characteristic function of a normal distribution, $G(x)$ can be represented in the form

$$G(x) = \frac{k}{\sqrt{2\pi a_1}} \int_{-\infty}^x \exp \left[-\frac{(x-\gamma)^2}{2a_1} \right] dx$$

so that, for any x ,

$$\int_{-\infty}^x e^{-u^2} dF(u) = \frac{k}{\sqrt{2\pi a_1}} \int_{-\infty}^x \exp \left[-\frac{(x-\gamma)^2}{2a_1} \right] dx \quad (2)$$

We shall show that F is everywhere differentiable and calculate its derivative.

Let $x_0 > 0$ and $h > 0$, then from (2)

$$\begin{aligned} \frac{k}{\sqrt{2\pi a_1}} \frac{e^{x_0^2}}{h} \int_{x_0}^{x_0+h} \exp \left[-\frac{(x-\gamma)^2}{2a_1} \right] dx &< \frac{F(x_0+h) - F(x_0)}{h} \\ &< \frac{k}{\sqrt{2\pi a_1}} e^{(x_0+h)^2} \int_{x_0}^{x_0+h} \exp \left[-\frac{(x-\gamma)^2}{2a_1} \right] dx \end{aligned}$$

and similarly for $h < 0$.

Hence it follows that

$$F'(x_0) = \frac{k}{\sqrt{2\pi a_1}} \exp \left[x_0^2 - \frac{(x_0 - \gamma)^2}{2a_1} \right]$$

so that F is a normal distribution function.

CHAPTER III

GENERATION OF RANDOM SAMPLES

Many techniques for the generation of pseudo-random numbers have been suggested in recent years. The most commonly used algorithm and the most suitable generator of uniform variates for most applications seems to be the multiplicative congruential generator. This method, which is based on calculating residues modulo some integer N , generates sequences whose statistical characteristics are good enough for most purposes.

After generating the uniform $(0, 1)$ random numbers using the congruential method, we proceed to generate the random variates from each of the thirty distributions. Random variables with a standard normal distribution, i.e., mean $\mu = 0$ and variance $\sigma^2 = 1$, have the following density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad -\infty < x < \infty$$

and thus a cumulative distribution function

$$F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}t^2} dt.$$

Since it is not possible to evaluate this integral analytically and then find the inverse distribution $F^{-1}(x)$, many methods and approximations have been suggested. The method we use to generate normal $(0, 1)$

variates is the following [Abramowitz and Stegun 1970].

$$\text{Let } Q(x_p) = p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_p} e^{-\frac{1}{2}t^2} dt$$

for values of $0 < p \leq 0.5$

$$x_p = \frac{c_0 + c_1 t + c_2 t^2}{d_1 t + d_2 t^2 + d_3 t^3} - t + \epsilon(p)$$

where $t = (-\ln p^2)^{1/2}$

$$|\epsilon(p)| < 4.5 \times 10^{-4}$$

$$c_0 = 2.515517 \quad d_1 = 1.432788$$

$$c_1 = 0.802853 \quad d_2 = 0.189269$$

$$c_2 = 0.010328 \quad d_3 = 0.001308;$$

for values of $0.5 < p \leq 1$

$$x_p = t - \frac{c_0 + c_1 t + c_2 t^2}{d_1 t + d_2 t^2 + d_3 t^3} + \epsilon(p)$$

x_p : is the required normal (0, 1) variate.

Variables with a double exponential (Laplace) distribution have the following density function

$$f_X(x) = \frac{1}{2} \alpha \exp(-\alpha|x - \beta|); \quad -\infty < x < \infty$$

$$-\infty < \beta < \infty, \alpha > 0$$

with $E(X) = \beta$ and $\text{Var}(X) = \frac{2}{\alpha^2}$

The cumulative distribution function of these variables is

$$F_X(x) = \begin{cases} \frac{1}{2} \exp[\alpha(x - \beta)] & , x \leq \beta \\ 1 - \frac{1}{2} \exp[-\alpha(x - \beta)] & , x > \beta \end{cases}$$

Hence given $0 < F_X(x) < 1$

$$x = \begin{cases} \beta + \frac{1}{\alpha} \log[2F_X(x)] & \text{if } F_X(x) \leq 0.5 \\ \beta - \frac{1}{\alpha} \log[2 - 2F_X(x)] & \text{if } F_X(x) > 0.5 \end{cases}$$

The parameters β and α are chosen to ensure that the mean μ and the variance σ^2 of this distribution are zero and one respectively, in agreement with those of the standard normal. Then $\beta = 0$ and $\alpha = \sqrt{2}$.

Variables that are Cauchy distributed have the following density function [Hastings and Peacock 1975]

$$f_X(x) = \frac{\alpha}{\pi} \left[\frac{1}{\alpha^2 + (x - \beta)^2} \right], -\infty < x < \infty, -\infty < \beta < \infty, \alpha > 0.$$

None of the moments of this distribution exists, and thus the mean and the variance do not exist, while the median is β . The Cauchy cumulative distribution function is

$$F_X(x) = \frac{1}{\pi} \text{Arctan} \left(\frac{x - \beta}{\alpha} \right) + \frac{1}{2}$$

Hence given $0 < F_X(x) < 1$

$$x = \beta - \alpha \cot(\pi F_X(x))$$

The parameters β and α are chosen in the following way. The median β is matched with the mean μ of a standard normal and the tails are matched at the 95% points, and thus

$$\alpha = \frac{-1.96}{\tan(-0.475\pi)}$$

The slash distribution is the ratio of a normal to an independent uniform $(0, 1)$. Random variates having a slash distribution are obtained by first generating normal variates and then using another seed, and generating an independent sequence of uniform $(0, 1)$ variates. Then by forming the ratios of the former to the latter we get the required slash distributed random variates.

The inverse cumulative distribution function of Tukey's λ family is given by

$$F_X^{-1}(x) = \frac{x^\lambda - (1-x)^\lambda}{\lambda}, \quad 0 < x < 1$$

For different values of the shape parameter λ we get a variety of symmetric distributions. When $\lambda = 1$,

$$F_X^{-1}(x) = 2x - 1, \quad 0 < x < 1$$

Hence
$$F_X(x) = \frac{x+1}{2}, \quad -1 < x < 1$$

which is the cumulative distribution function of a uniform $(-1, 1)$ random variable. When $\lambda = 0$,

$$\lim_{\lambda \rightarrow 0} F_X^{-1}(x) = \frac{\lim_{\lambda \rightarrow 0} \frac{d}{d\lambda} [x^\lambda - (1-x)^\lambda]}{\lim_{\lambda \rightarrow 0} \frac{d}{d\lambda} (\lambda)}$$

$$\begin{aligned}
&= \lim_{\lambda \rightarrow 0} [x^\lambda \log x - (1-x)^\lambda \log(1-x)] \\
&= \log x - \log(1-x) \\
&= \log \frac{x}{1-x}
\end{aligned}$$

which is the inverse cumulative distribution function of a logistic random variable. The equation above is the formula used to generate random variates from a logistic distribution.

Pareto distributed random variables have the following density function

$$f_X(x) = \alpha \beta^\alpha x^{-(\alpha+1)}, \quad x > \beta, \alpha > 0, \beta > 0$$

with $E(X) = \frac{\alpha \beta}{\alpha - 1}, \alpha > 1$

and $\text{Var}(X) = \frac{\alpha \beta^2}{(\alpha - 1)^2 (\alpha - 2)}, \alpha > 2$

The cumulative distribution function of these variables is

$$F_X(x) = 1 - \left(\frac{\beta}{x}\right)^\alpha$$

Hence for $0 < F_X(x) < 1$

$$x = \beta [1 - F_X(x)]^{-\frac{1}{\alpha}}$$

By setting the mean equal to μ and the variance equal to σ^2 one finds that

$$\begin{aligned}
\alpha &= 1 + \sqrt{1 + \left(\frac{\mu}{\sigma}\right)^2} \\
\beta &= \frac{\mu \sqrt{1 + \left(\frac{\mu}{\sigma}\right)^2}}{1 + \sqrt{1 + \left(\frac{\mu}{\sigma}\right)^2}}
\end{aligned}$$

This distribution cannot have a zero mean for then we will have a degenerate random variable $X = 0$ with probability one. For this reason we set $\mu = 0.01$.

A random variable X is said to have a Weibull distribution if its density function is given by [Hastings and Peacock 1975]

$$f_X(x) = \frac{cx^{c-1}}{b^c} \exp\left(-\frac{x}{b}\right)^c, \quad x > 0,$$

$b > 0$ is the scale parameter and

$c > 0$ is the shape parameter.

The values of the parameters we considered are $b = 1$ and $c = 0.5, 1.0, 3.0$ for which the shape of the density changes drastically. [Hastings and Peacock 1975]. The cumulative distribution function of this variable is

$$F_X(x) = 1 - \exp(-x^c)$$

Hence given $0 < F_X(x) < 1$

$$x = -\log[1 - F_X(x)]$$

Random variables that are transformations of symmetric ones are generated using their respective transformations. For example, to generate random variates $X = (1 - \alpha)Z + \alpha Z^2$ where Z is a standard normal random variable, we first generate the standard normal variates and then by using the above transformation we get the required random variates. Similarly for the other two transformed random variables

$X = \exp(\beta Z)$ and $X = Z \exp(\gamma Z)$. For different values of α , β and γ we get the other twelve asymmetric distributions.

CHAPTER IV

MONTE CARLO RESULTS

The purpose of this chapter is to give a general description of the results. It is a survey of the computer study of the problem of testing symmetry. The number of random variates generated from each distribution is fifty; this gives rise to a sufficiently large number of points to consider. Note that all $\binom{50}{2}$ possible samples of size two are used. Moreover, this number of observations, $\binom{50}{2} = 1225$, can be classified into a 35×35 contingency table, which is easier to handle, and from which we compute the chi-square statistic. A number of computer runs are performed utilizing different input seeds for the random number generator. Thus we have, for each distribution, a number of computer output runs corresponding to different sets of data points. Furthermore, different values for the mean and variance are used to study the effects of these extraneous factors on the test. The results obtained for each distribution are all consistent in the sense that there is no noticeable change in the general shape of the scatter diagrams, and in the computed χ^2 and r^2 values. This asserts the robustness and the validity of this test. A complete set of contingency tables appears in the appendix together with the χ^2 and r^2 values for each individual distribution.

The $\binom{50}{2}$ computed pairs of statistics $(2\bar{X}, 2S^2)$ are plotted against one another and the scatter plot is thoroughly investigated.

Data from most of the asymmetric distributions are concentrated along a single line. Though the degree of concentration may differ from one distribution to the other, it is quite obvious that the data points are clustered around a single line. The values of the χ^2 statistic and r^2 for each of these distributions are considerably higher than those of the normal. Some other asymmetric distributions do not show this property of concentration. In fact they behave very much like the normal and many other symmetric distributions where the points are uniformly dispersed over the whole diagram. On the other hand, a number of symmetric distributions do not show concentration along a single line but along two distinct lines, the orientation of which gives an indication as to how the tails of these distributions fall off. A more detailed discussion of this phenomenon is given in the next chapter. The computed χ^2 and r^2 values for these distributions are higher than those of the normal.

The appendix contains the computer output of thirty 35×35 contingency tables. The exact number of points in each cell is shown, giving a general overview of how the points are spread. The χ^2 and r^2 values are given at the bottom of each table. In order to give a clearer picture of the concentration pattern, the following display technique is adopted. If the number of points in a cell is less than five, a blank replaces that number. If the number of points is between five and fifteen, a star is printed in that cell, while two stars are printed if the number of points exceeds fifteen. Each

contingency table, in the appendix, is followed by its corresponding concentration display picture.

CHAPTER V

ANALYSIS OF RESULTS

The test under consideration is performed on thirty sets of data, each generated from a different parent distribution. Since this test is basically a test of symmetry, it is therefore of interest to know how well it performs when different symmetric and asymmetric distributions are considered. The results obtained and discussed in the previous chapter indicate the quality of performance of the test when the underlying assumptions are slightly modified.

Data generated from most of the asymmetric distributions behave similarly since all show the one line concentration property. This and the fact that data from symmetric distributions do not have a single line concentration, assert the validity of this test to distinguish symmetric from asymmetric distributions. Thus this test is effective in the sense that even if the assumption of symmetry is modified slightly, or more explicitly, even in the case of slightly skewed symmetric distributions the test is capable of discriminating between symmetric and asymmetric distributions.

On the other hand, as long as the distribution is symmetric the results obtained are all consistent, independent of its shape. Hence the effect of extraneous factors not under test, e.g., shape, is negligible. This asserts the robustness of validity of the test.

Furthermore, this test is not only effective in distinguishing symmetric from asymmetric distributions, but it also gives an indication of how symmetric distributions behave away from the center, precisely, whether the distribution has long or short tails.

Distributions like the Laplace, the Cauchy, the slash, Tukey's λ for $\lambda = -0.5, -1.0, -1.5, -2.0$, that have fat (long) tails compared to those of the normal, show a two-line concentration starting to the left of the scatter diagram and opening to the right with an acute angle between the two. (See figures 2-B, 3-B, 4-B, 11-B, 12-B, 13-B, 14-B). Distributions with shorter tails than the normal, e.g., the uniform, Tukey's λ where $\lambda = 0.75, 0.5, 0.25$, show concentration along two lines that intersect to the right and then branch out to the left subtending an acute angle. (See figures 5-B, 6-B, 7-B, 8-B)

Some distributions, in particular the logistic, the Weibull with shape parameter = 3, two forms of the lognormal with coefficients of variation 0.25 and 0.1 respectively, $(0.9)Z + (0.1)Z^2$ and $Ze^{(0.1)Z}$ do not show any concentration. Moreover, the scatter plots of these distributions very much resemble that of the normal (See figures 9-B, 16-B, 22-B, 25-B, 26-B, 30-B). To justify the behavior of these distributions we note that:

- (a) A particular form of the logistic distribution with cumulative

distribution function $F_X(x) = \frac{1}{1 + e^{-\frac{\pi}{\sqrt{3}}x}}$ is very close to

a normal (0, 1) and the exact form used in this study is

$$F_X(x) = \frac{1}{1 + e^{-x}}.$$

- (b) A lognormal distribution which has a small absolute value (say less than 0.5) of the coefficient of variation can give a good representation of a normal distribution.
- (c) A form of the Weibull distribution with shape parameter ≈ 3.25 is almost identical with the unit normal. Hence the Weibull (1, 3) is close to a normal (0, 1). [Johnson and Kotz 1970]
- (d) $(0.9)Z + (0.1)Z^2$ is expected to behave very much like a normal since it is a mixture of a standard normal and chi-squared random variable in the ratio of nine to one.
- (e) $Ze^{\gamma Z} \approx Z(1 + \gamma Z)$ for small γ
 $= Z + \gamma Z^2$

Thus for $\gamma = 0.1$, we have $Z + (0.1)Z^2$ which is even closer to a normal than $(0.9)Z + (0.1)Z^2$.

The normal distribution is characterized by the independence of \bar{X} and S^2 . Hence it is expected that the chi-squared test of independence would give better results for the normal and thus help identify the normal among all other distributions. However, due to the fact that many of the distributions considered give good approximations to the normal, the values of the χ^2 statistics computed for these are almost equal to that of the normal. On the

other hand, the r^2 value computed, which is a measure of association of the two variables \bar{X} and S^2 is very much smaller for the normal than for the other thirty distribution considered. Hence based on the r^2 value, normality is established.

CHAPTER VI

SUMMARY AND CONCLUSIONS

In the preceding five chapters we constructed and discussed the behavior of a new test of symmetry based on the relation between the sample mean \bar{X} and the sample variance S^2 .

We began chapter 1 with a discussion of the notion of symmetry and symmetric distributions, and noted exactly what tests of symmetry had been performed. We also gave a list of the distributions considered for the test. Finally we discussed the notion of robustness and gave an idea of the importance of the normal distribution and the test of normality. Theoretical derivations and basic formulae in support of the methods used were given in chapter 2 and the motivation for the test was considered. In chapter 3 we presented the techniques used in generating the random variates from each of the thirty distributions considered. Density functions, cumulative distribution functions and inverse distribution functions were given, together with the methods of generating random variates that are skewed transformations of the normal. A survey of the results was given in chapter 4. Following this was a discussion of the performance of the test and its robustness of validity and efficiency. The display techniques were also discussed. Chapter 5 concluded the work and gave more detailed information about the results, justifi-

cation for some of the less intuitive results and the method used to test for normality.

Essentially, all the results of this thesis are derived from empirical sampling studies. They are accordingly subject to errors and uncertainties. The experience of this study has been that empirical sampling can provide a very useful general guide on sensitivity properties even with a few Monte Carlo runs. One way to elaborate on this work and perhaps improve on it is to consider a larger number of distributions and exhaust all possible size samples of the chosen population. The immediate problem that will arise is the increase in the amount of computer time needed to do the computations.

APPENDIX

CONCENTRATION PATTERNS OF DISTRIBUTIONS UNDER TEST

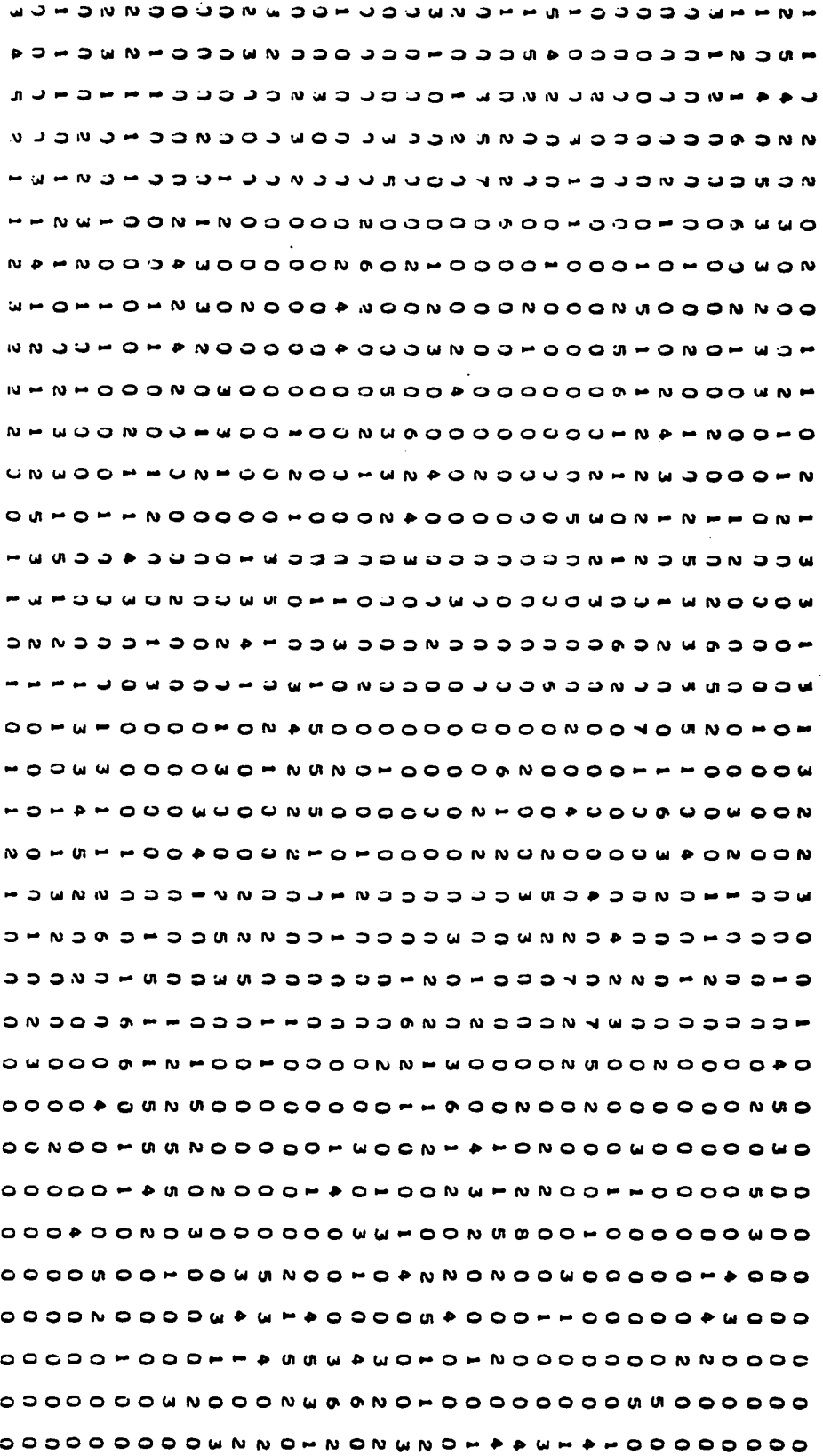
No Concentration	One-Line Concentration	Left-Right* Concentration	Right-Left** Concentration
Normal	Pareto	Laplace	Uniform
Weibull(1,3)	Weibull(1,1)	Cauchy	Tukey's(0.75)
Logistic	Weibull(1,0.5)	Slash	Tukey's(0.5)
Tukey's(-0.25)	Z^2	Tukey's(-0.5)	Tukey's(0.25)
$(0.9)Z+(0.1)Z^2$	$(0.5)Z+(0.5)Z^2$	Tukey's(-1.0)	
$e^{(0.25)Z}$	$(0.75)Z+(0.25)Z^2$	Tukey's(-1.5)	
$e^{(0.1)Z}$	e^Z	Tukey's(-2.0)	
$Ze^{(0.1)Z}$	$e^{(0.5)Z}$		
	Ze^Z		
	$Ze^{(0.5)Z}$		
	$Ze^{(0.25)Z}$		

* Left-Right concentration means that concentration looks like <

** Right-Left concentration means that concentration looks like >

NORMALLY DISTRIBUTED RANDOM VARIABLE

FIGURE 1 - A



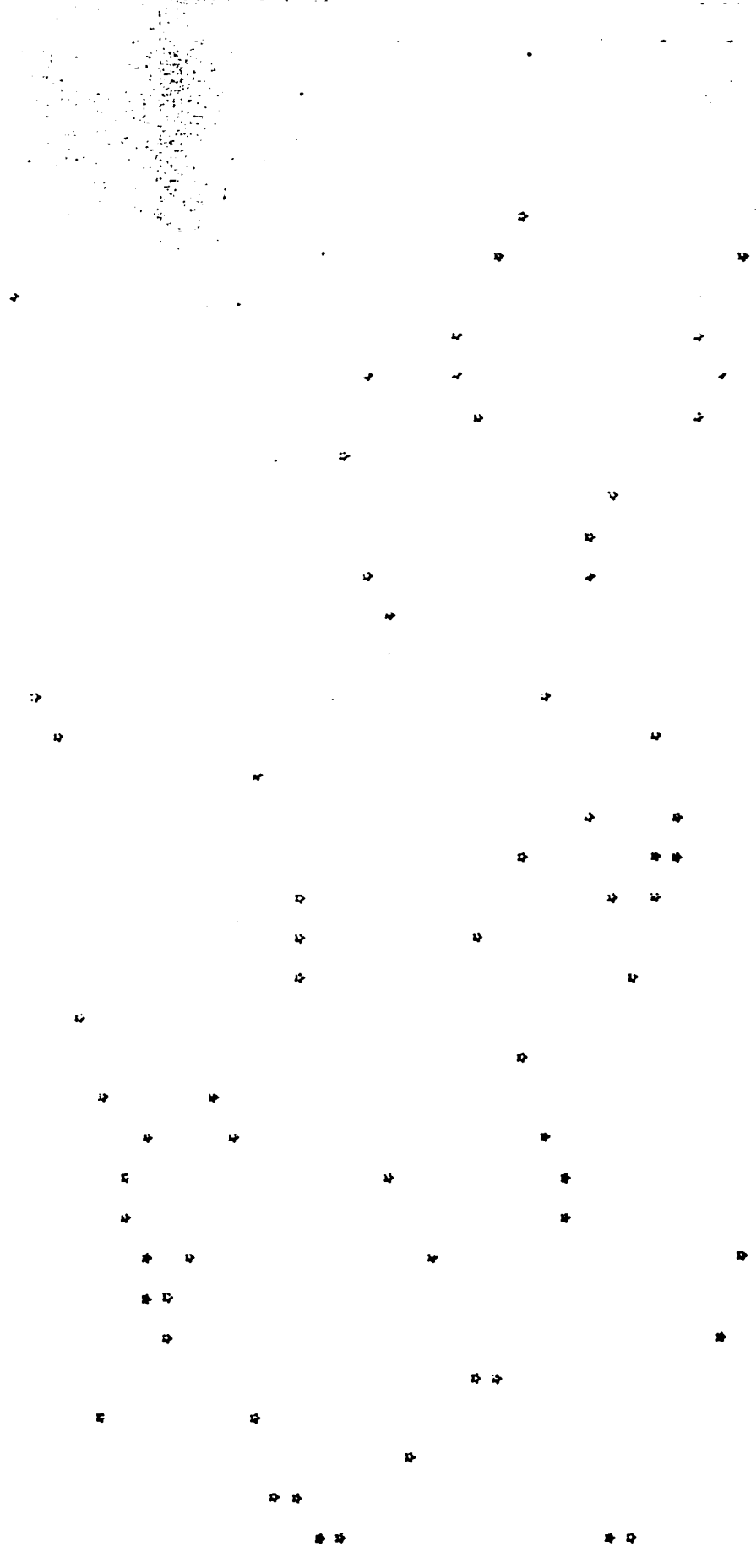
CHISQUARE = 2976

SQUARE OF CORRELATION COEFFICIENT = 0.000004942

NORMALLY DISTRIBUTED RANDOM VARIABLE

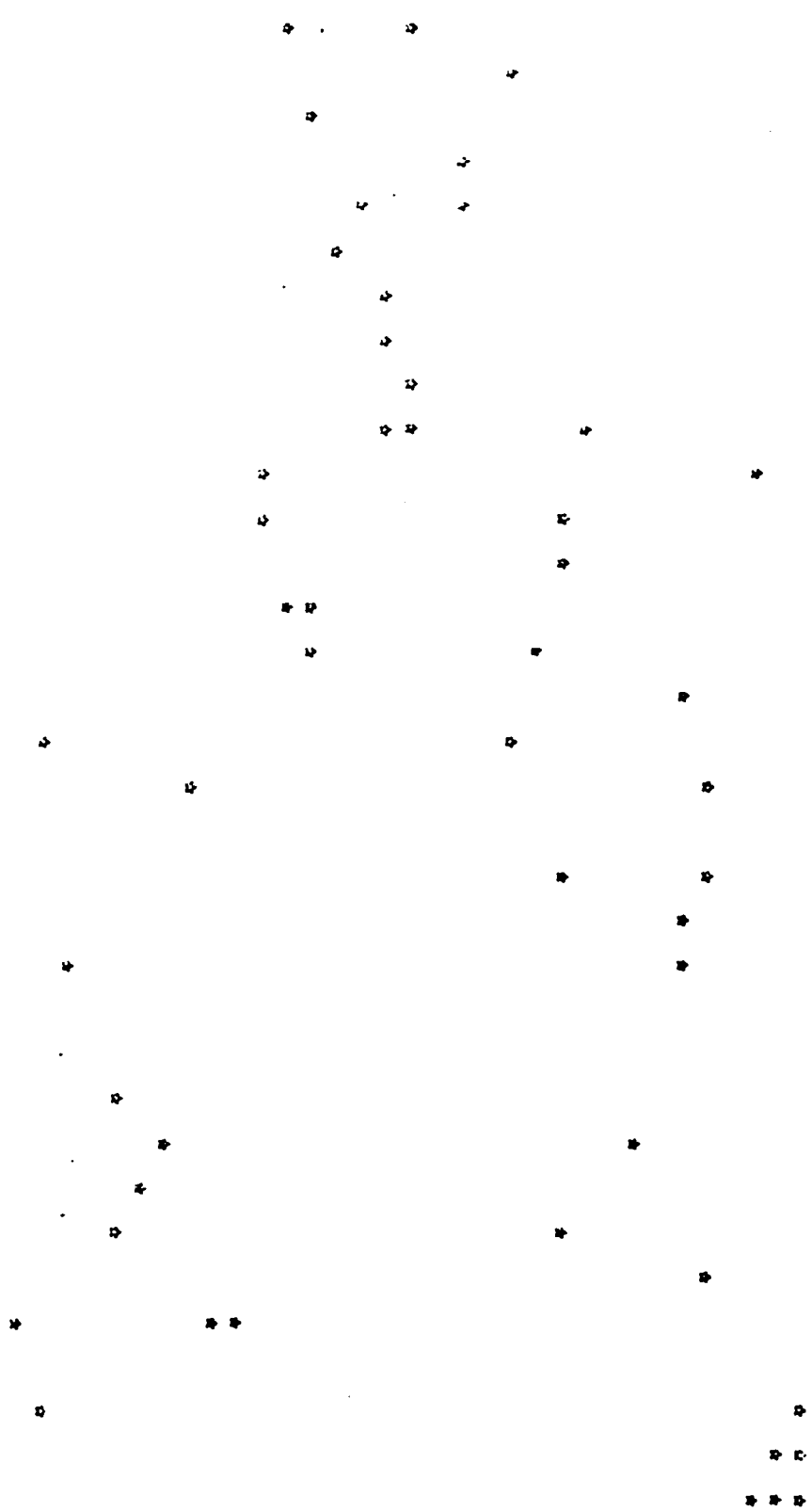
FIGURE 1 - B

CATENOID = 2475
SQUARE OF CORRELATION COEFFICIENT = 0.000004842



DOUBLE EXPONENTIALLY DISTRIBUTED RANDOM VARIABLE

FIGURE 2 - B

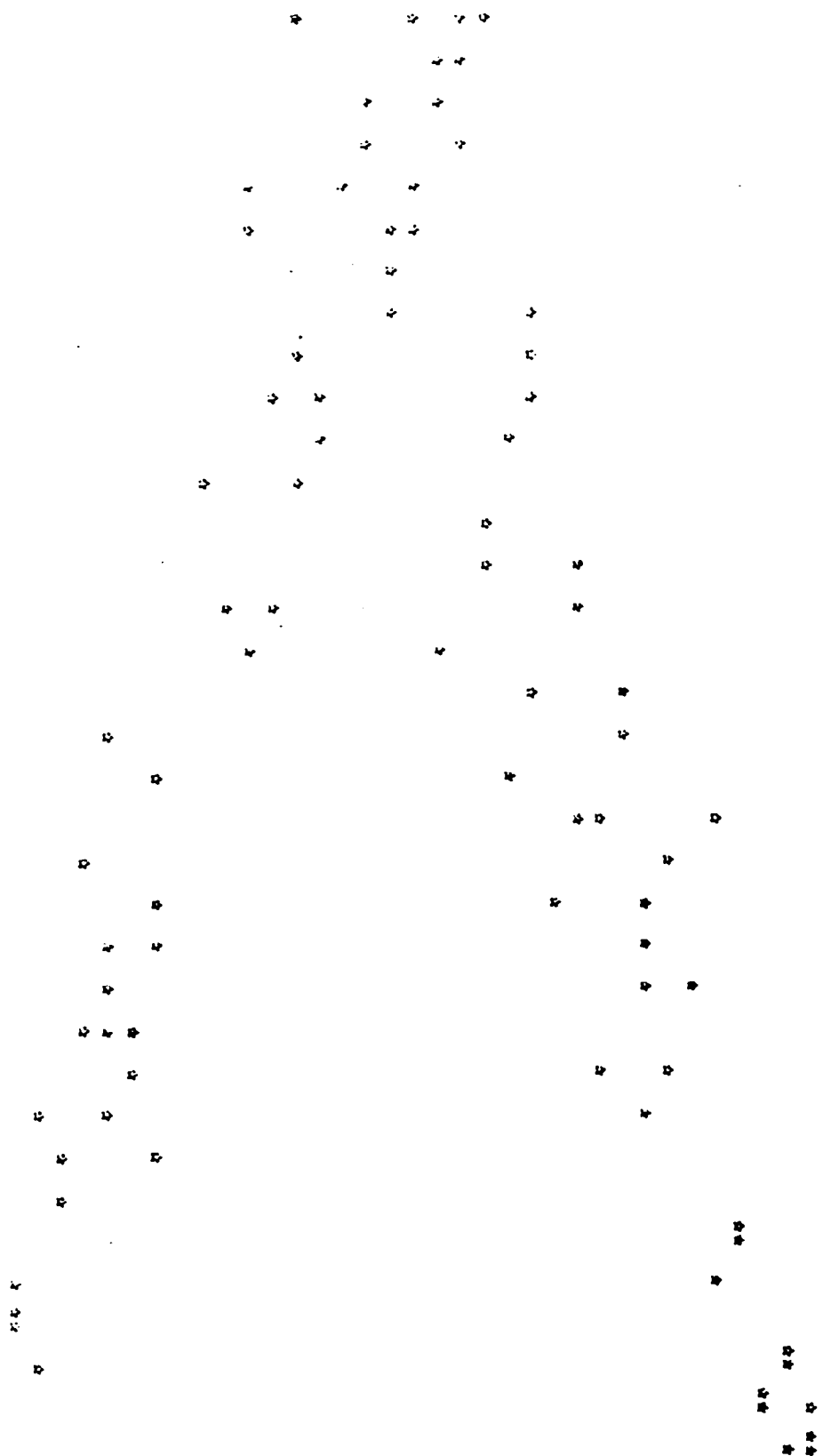


CHI SQUARE = 2872

SQUARE OF CORRELATION COEFFICIENT = 0.1197278613

CAUCHY DISTRIBUTED RANDOM VARIABLE

FIGURE 3 - B



CHIRQUAP = 0174

COEFFICIENT OF CORRELATION COEFFICIENT = 0.729476035

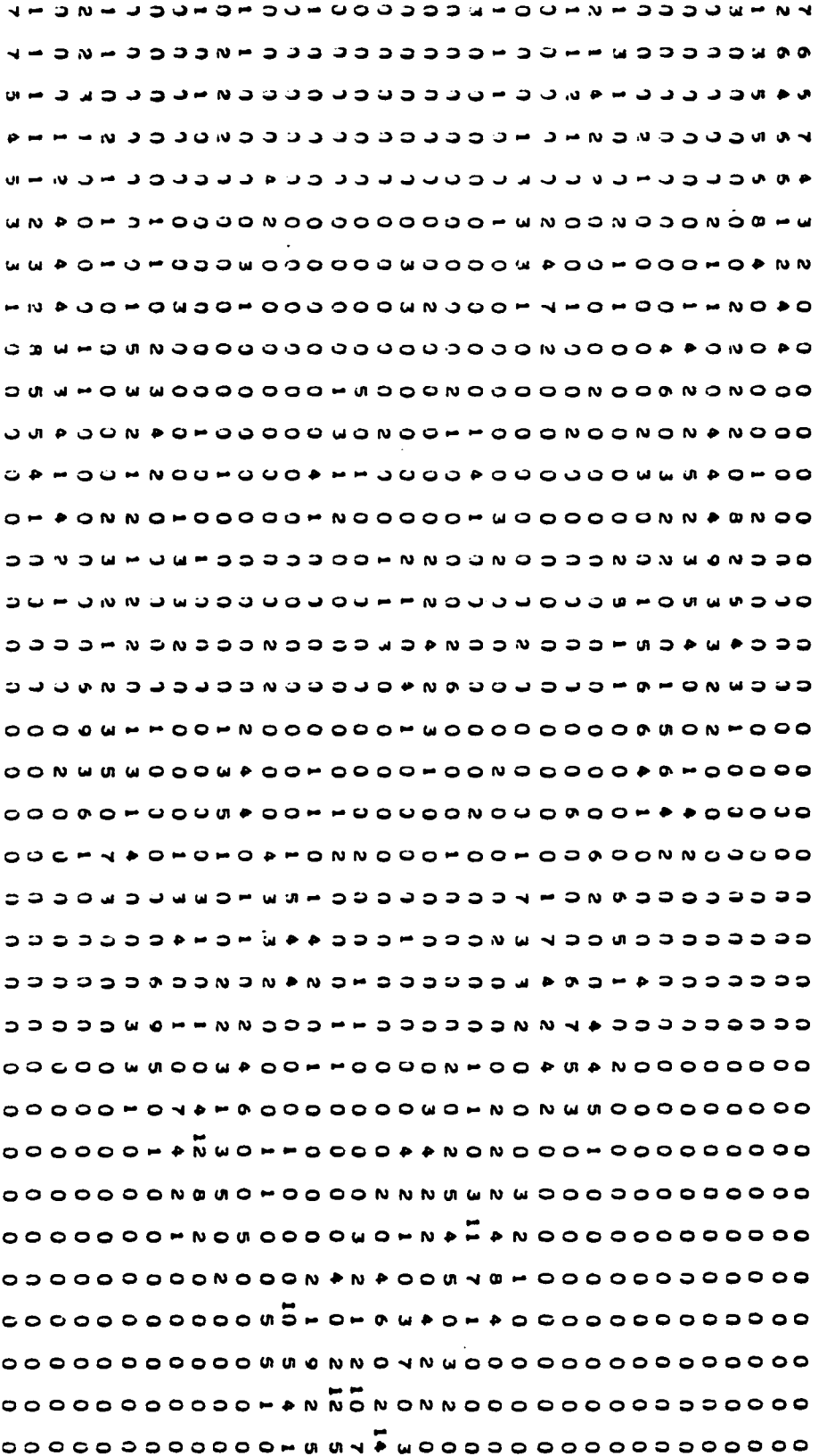
SLASH DISTRIBUTED RANDOM VARIABLE

FIGURE 4 - 3



FIGURE 5 - A

TUKEY(LAMBDA = 1.00) DISTRIBUTED RANDOM VARIABLE



CHISQUARE = 4700

SQUARE OF CORRELATION COEFFICIENT = 0.0109659113

FIGURE 5 - 8
TUKEY(LAMBDA = 1.00) DISTRIBUTED RANDOM VARIABLE

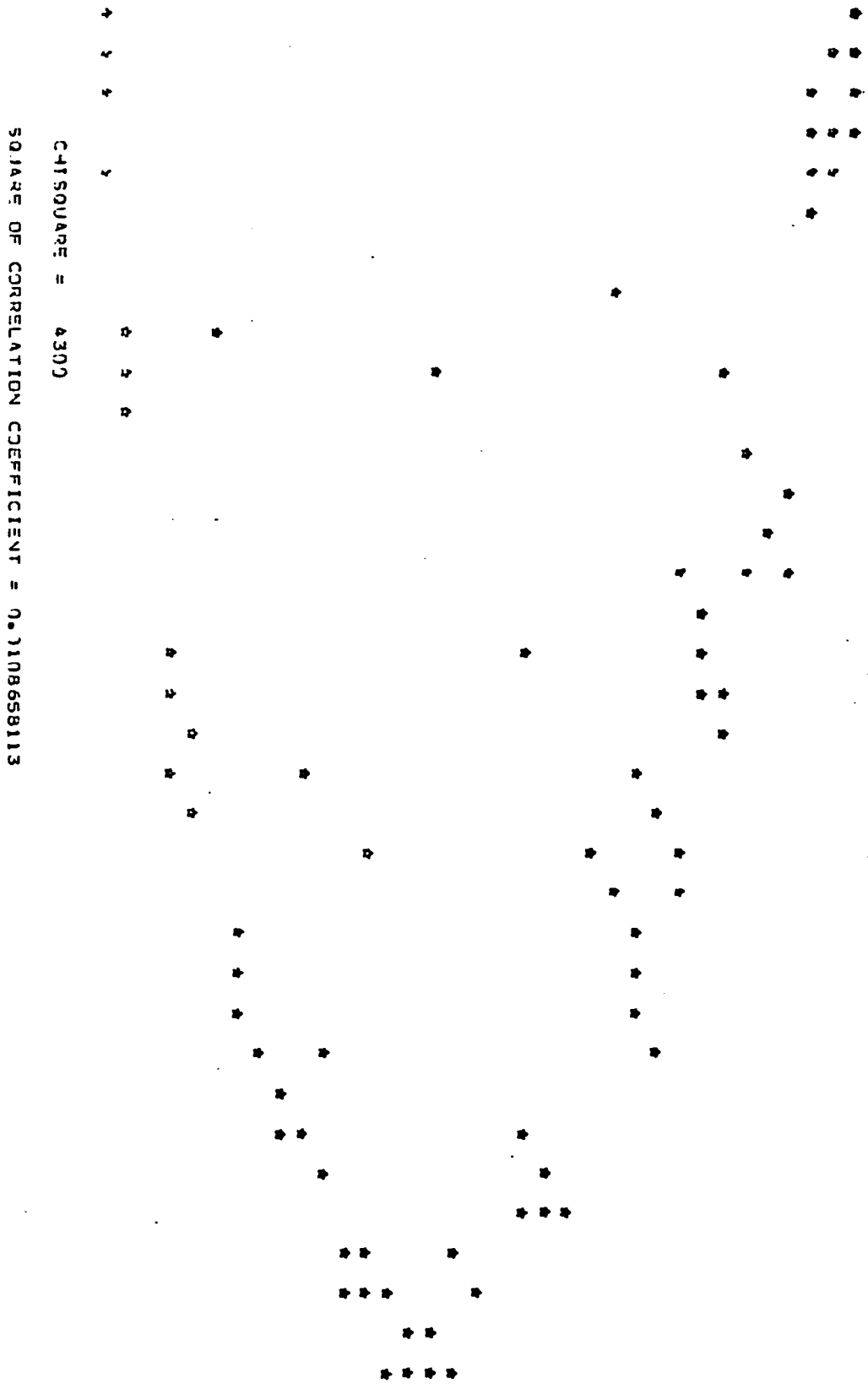
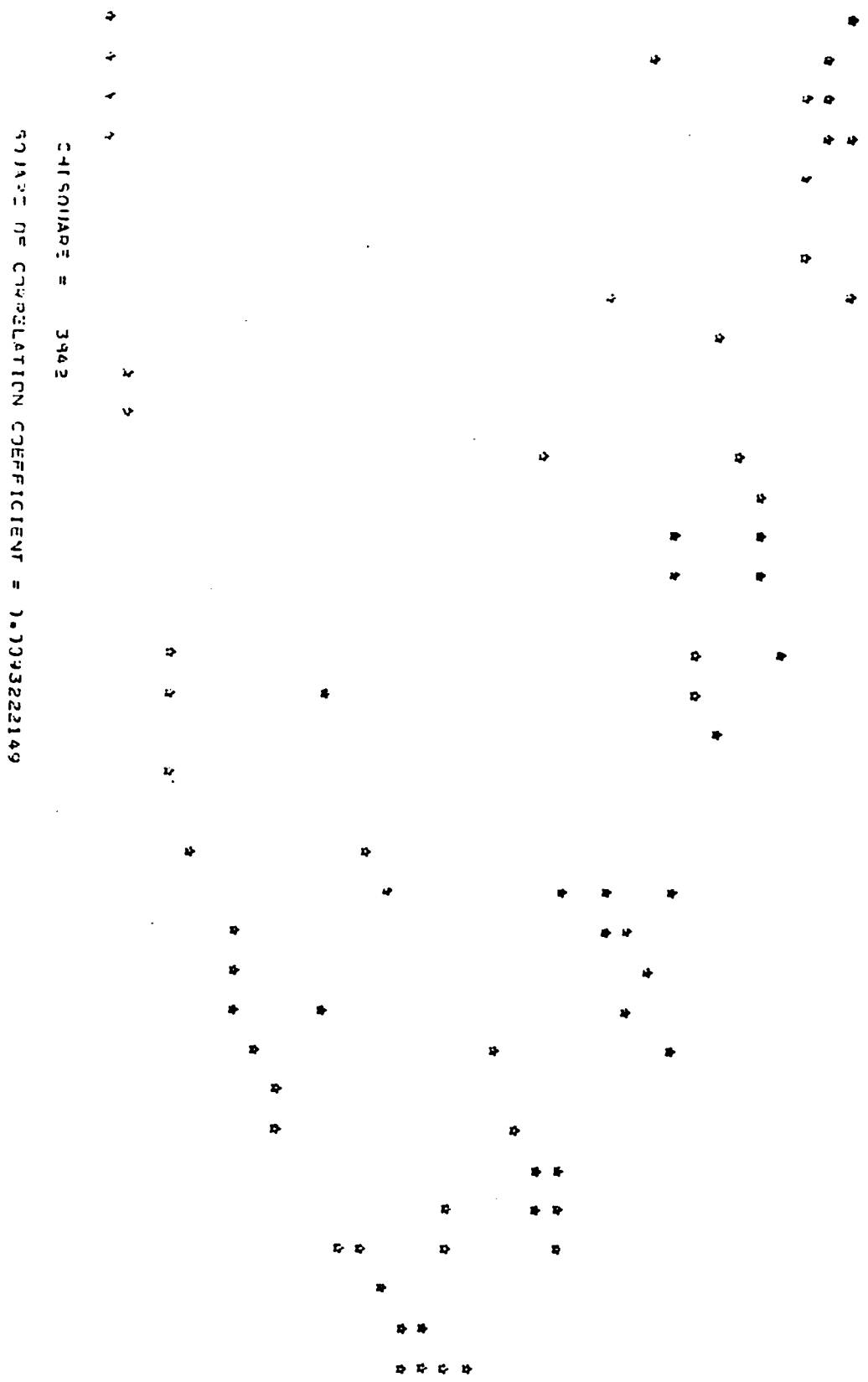


FIGURE 6 - B

TUKEY(LAMBDA = 0.75) DISTRIBUTED RANDOM VARIABLE

55



TUKEY(LAMBDA = 0.50) DISTRIBUTED RANDOM VARIABLE

FIGURE 7 - 8

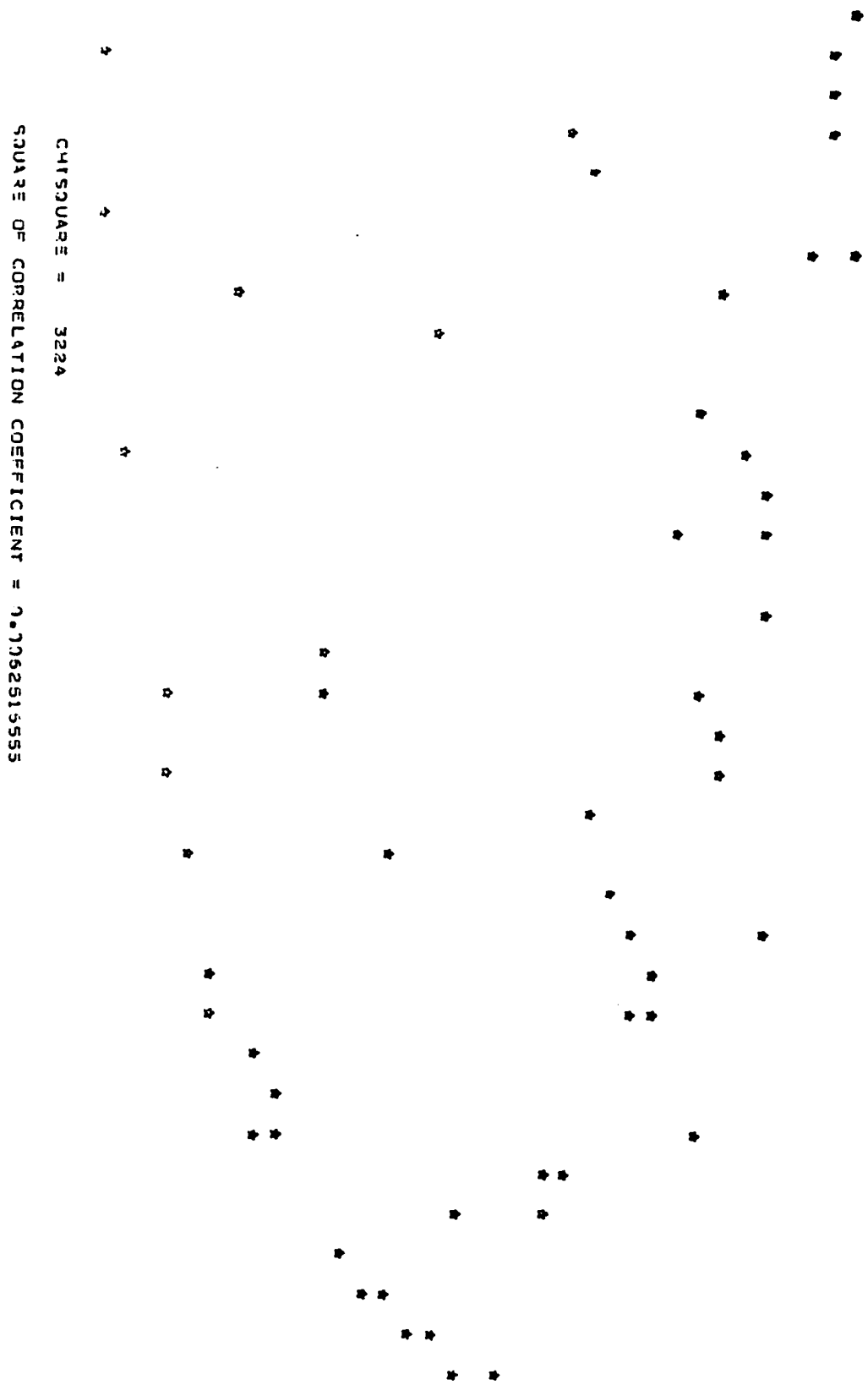
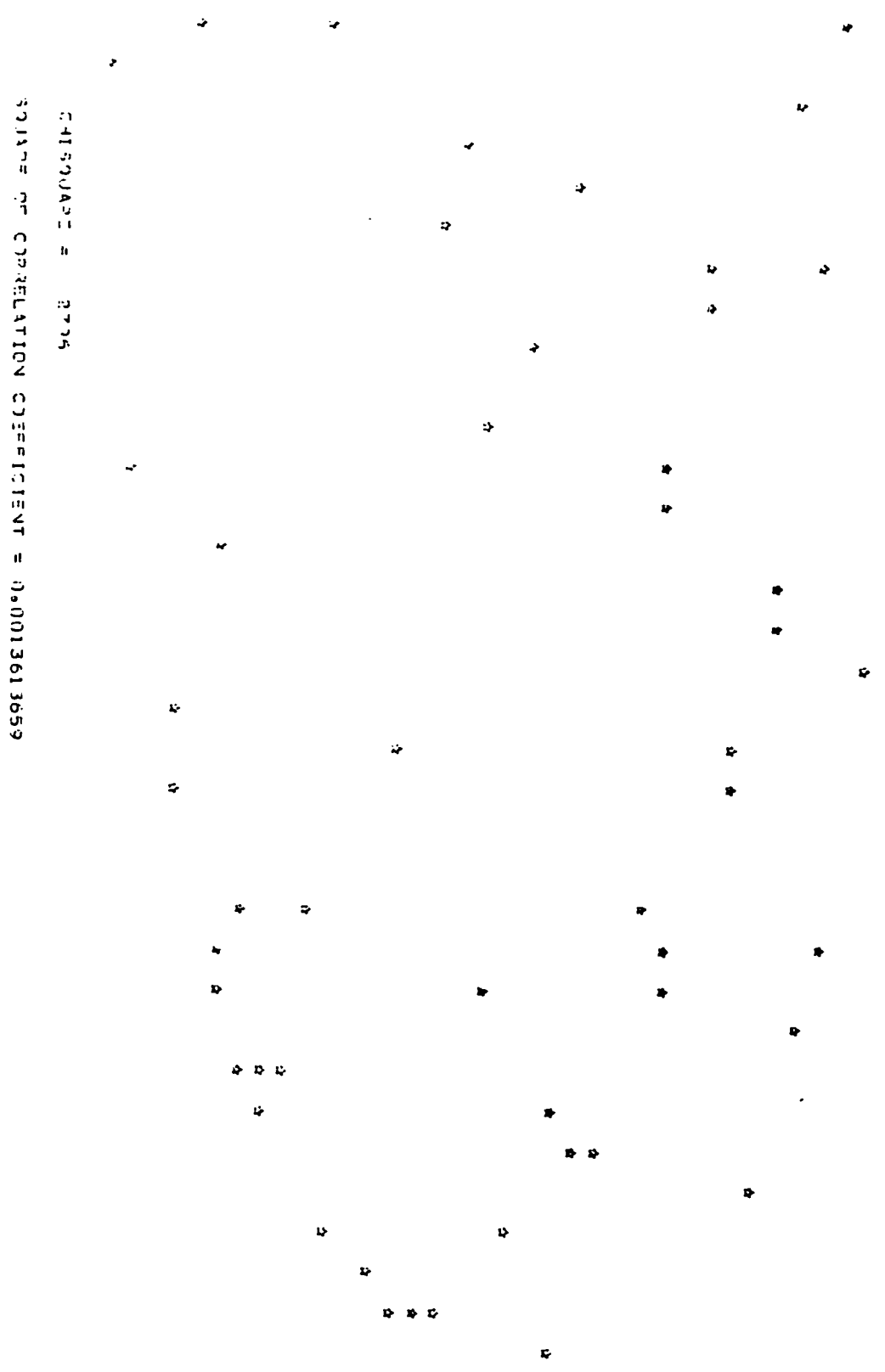
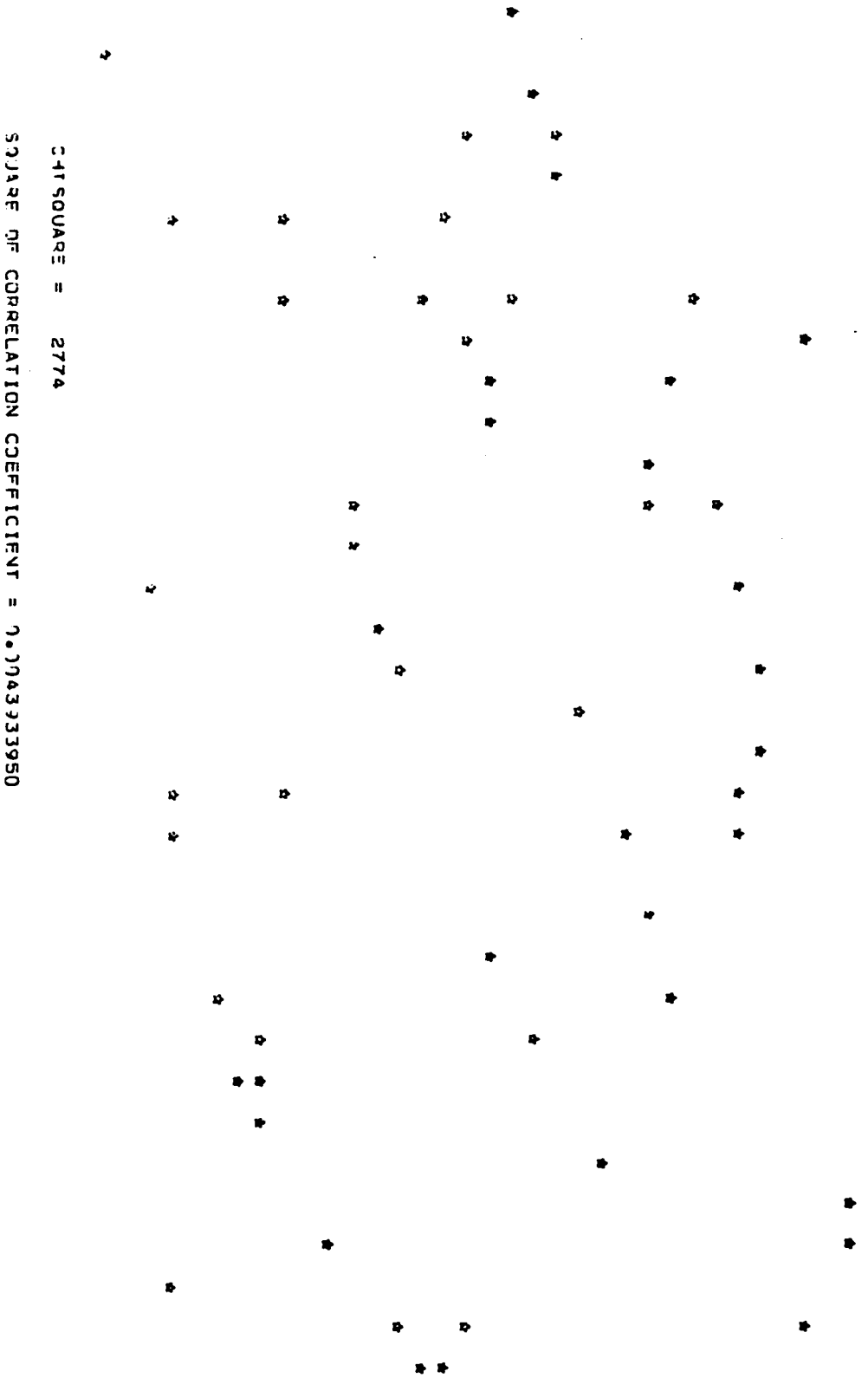


FIGURE 8 - B
TUKEY(LAMBDA = 0.25) DISTRIBUTED RANDOM VARIABLE



TUKEY(LAMBDA = 0.00) DISTRIBUTED RANDOM VARIABLE

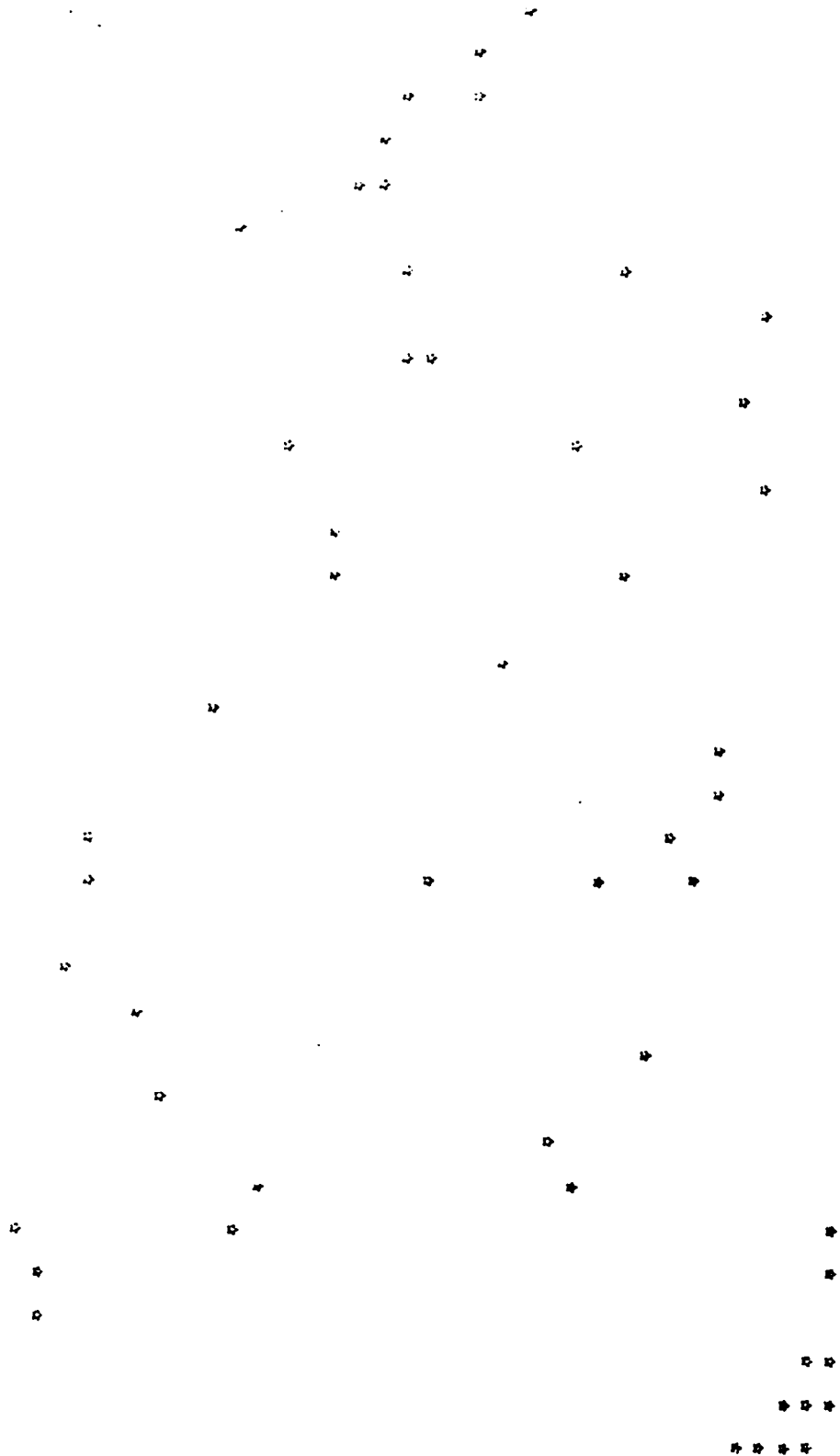
FIGURE 9 - 8



TUKEY(LAMBDA = -0.25) DISTRIBUTED RANDOM VARIABLE

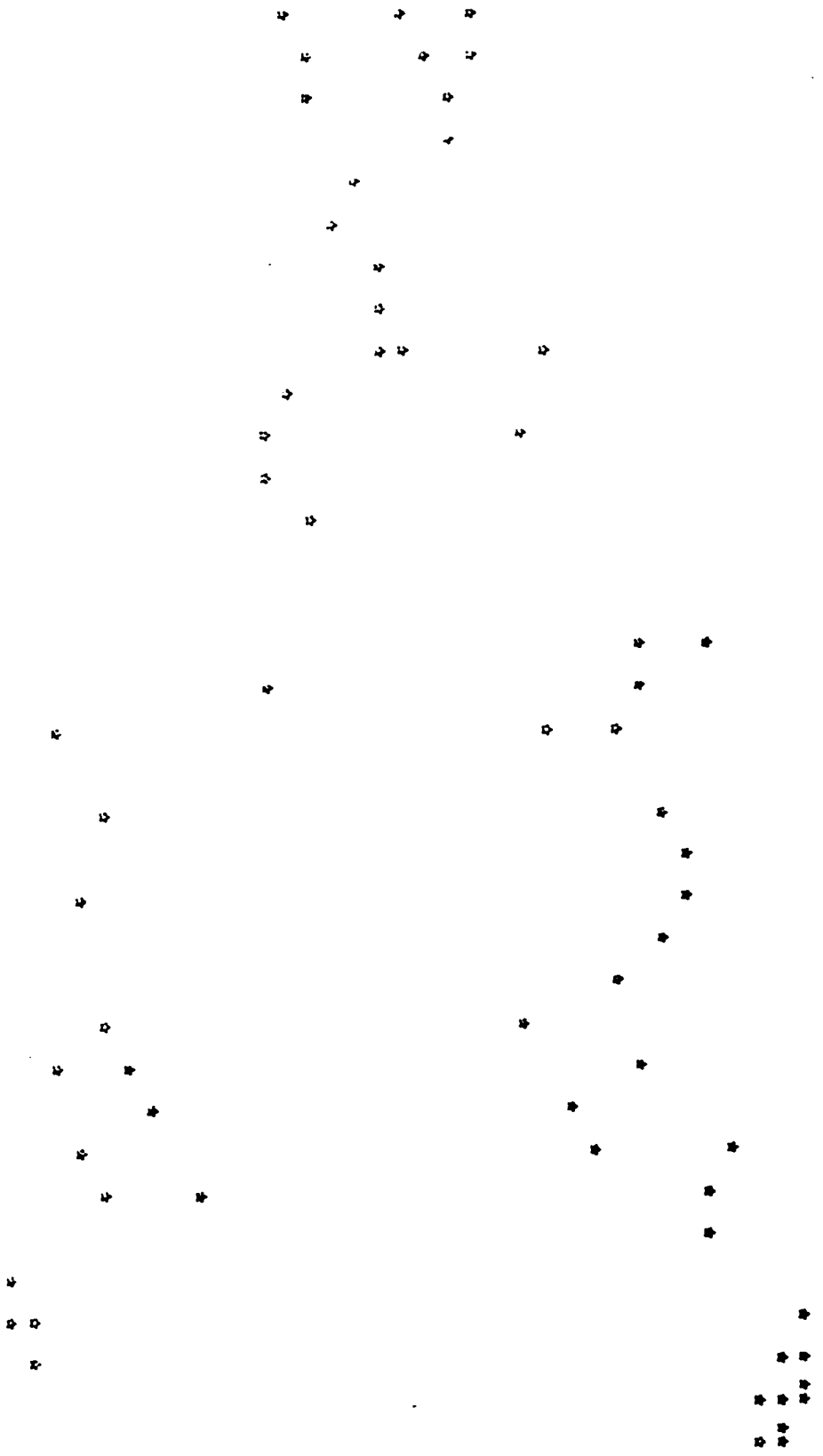
FIGURE 10 - 3

CHI SQUARE = 3174
SQUARE OF CORRELATION COEFFICIENT = 0.7739091732



TUKEY(LAMBDA = -0.50) DISTRIBUTED RANDOM VARIABLE

65

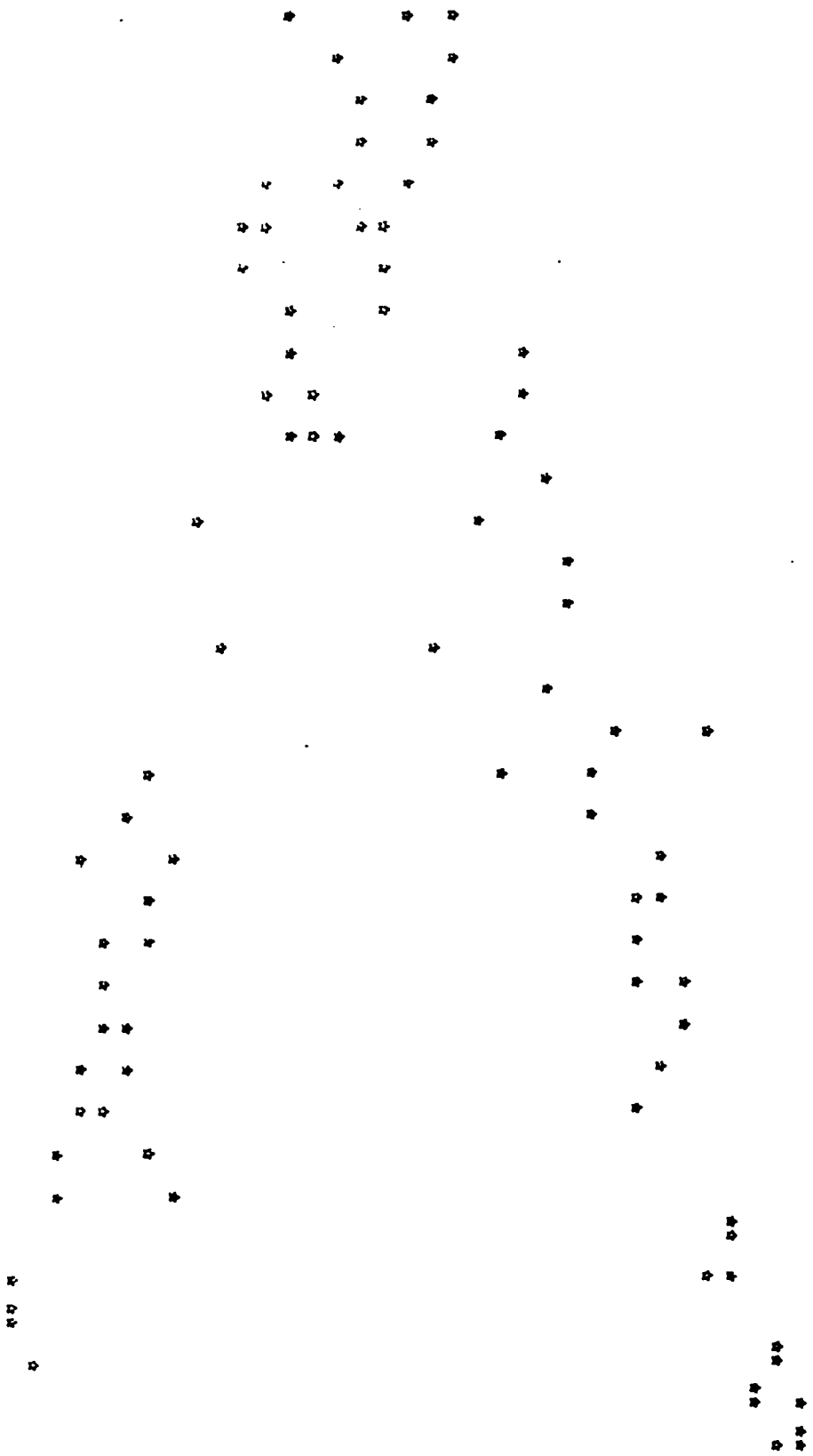


CHI SQUARE = 3903

SQ. OF CORR. COEFFICIENT = 0.294260035

TUKEY(LAMBDA = -1.00) DISTRIBUTED RANDOM VARIABLE

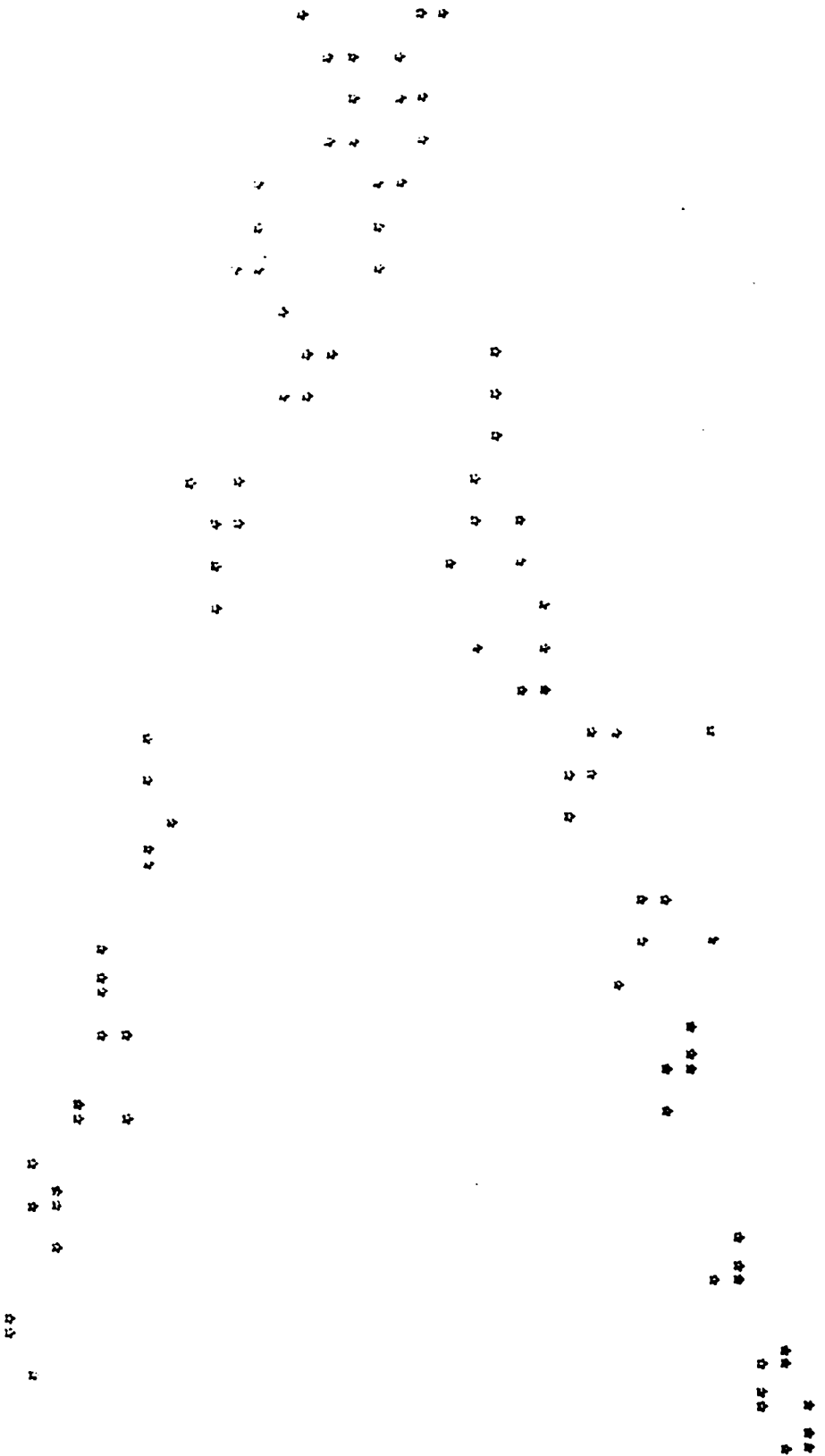
FIGURE 12 - 9



CHISQUARE = 5510

SQUARE OF CORRELATION COEFFICIENT = 0.73882005

TUKEY(LAMBDA = -1.50) DISTRIBUTED RANDOM VARIABLE

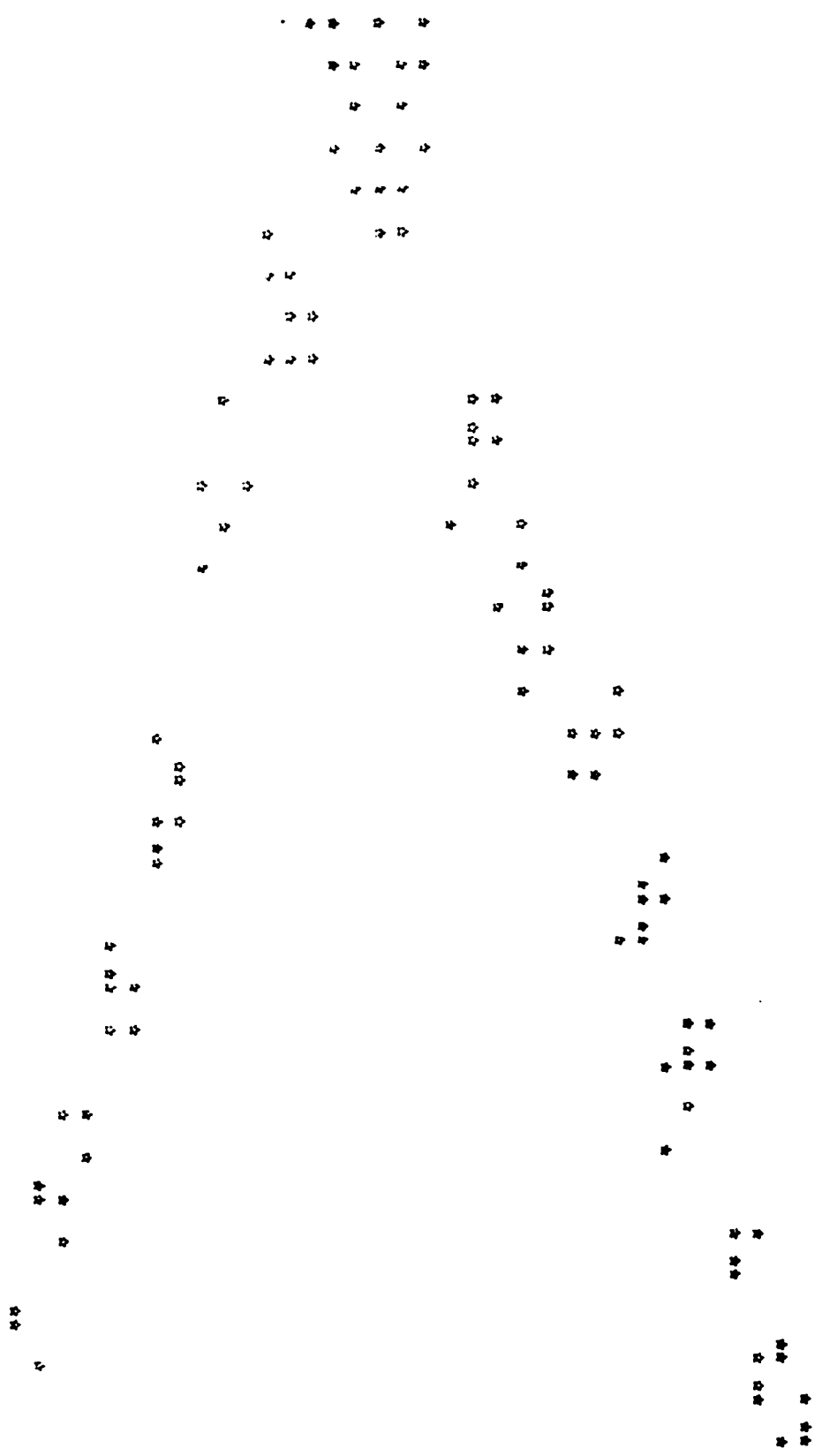


CHI SQUARE = 877A

SQUARE WIF CORRELATION COEFFICIENT = 1.0383901345

FIGURE 14 - 8

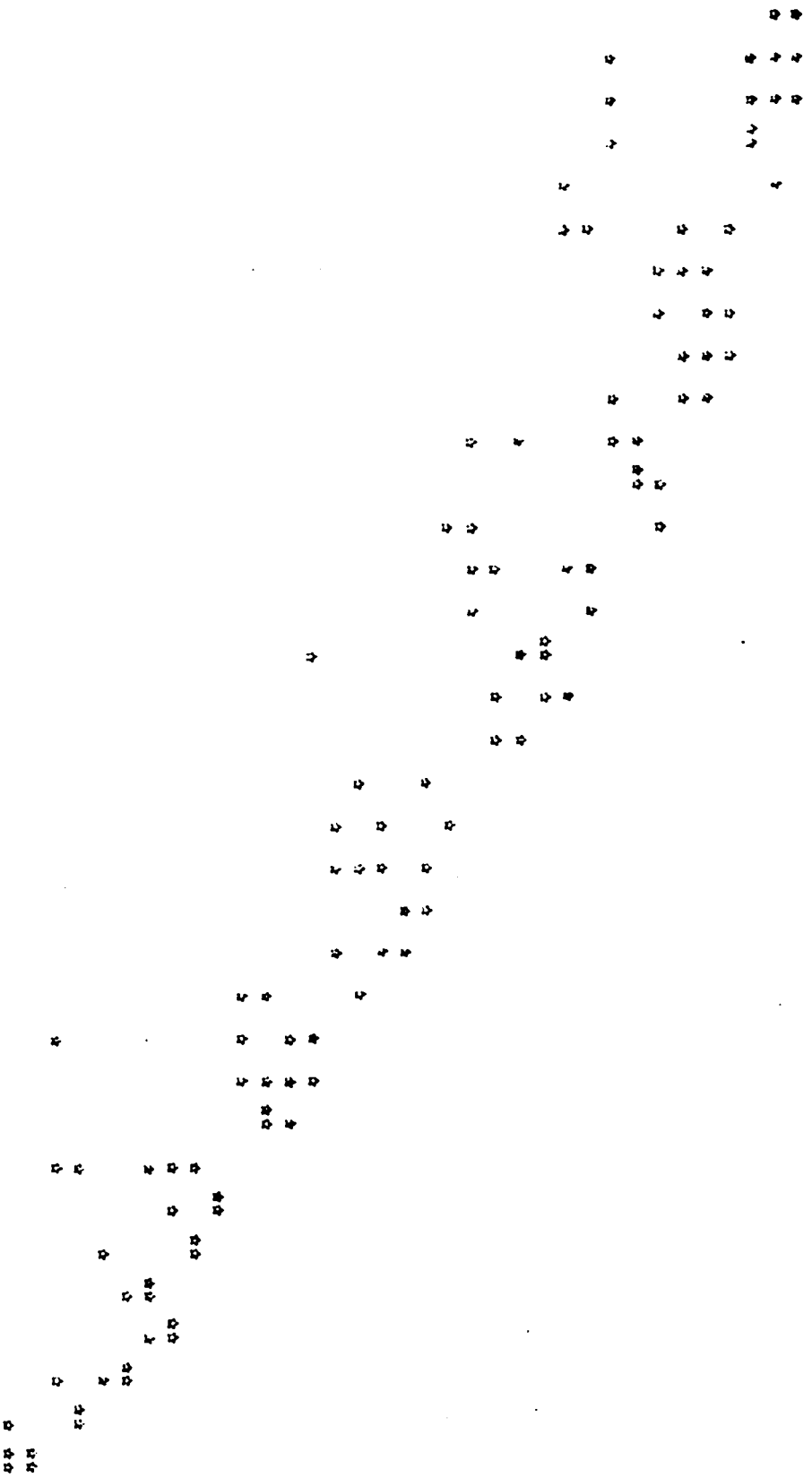
TUKEY(LAMBDA = -2.00) DISTRIBUTED RANDOM VARIABLE



CHI SQUARE = 13452

COEFF OF CORRELATION COEFFICIENT = 0.914735079

PARETO DISTRIBUTED RANDOM VARIABLE

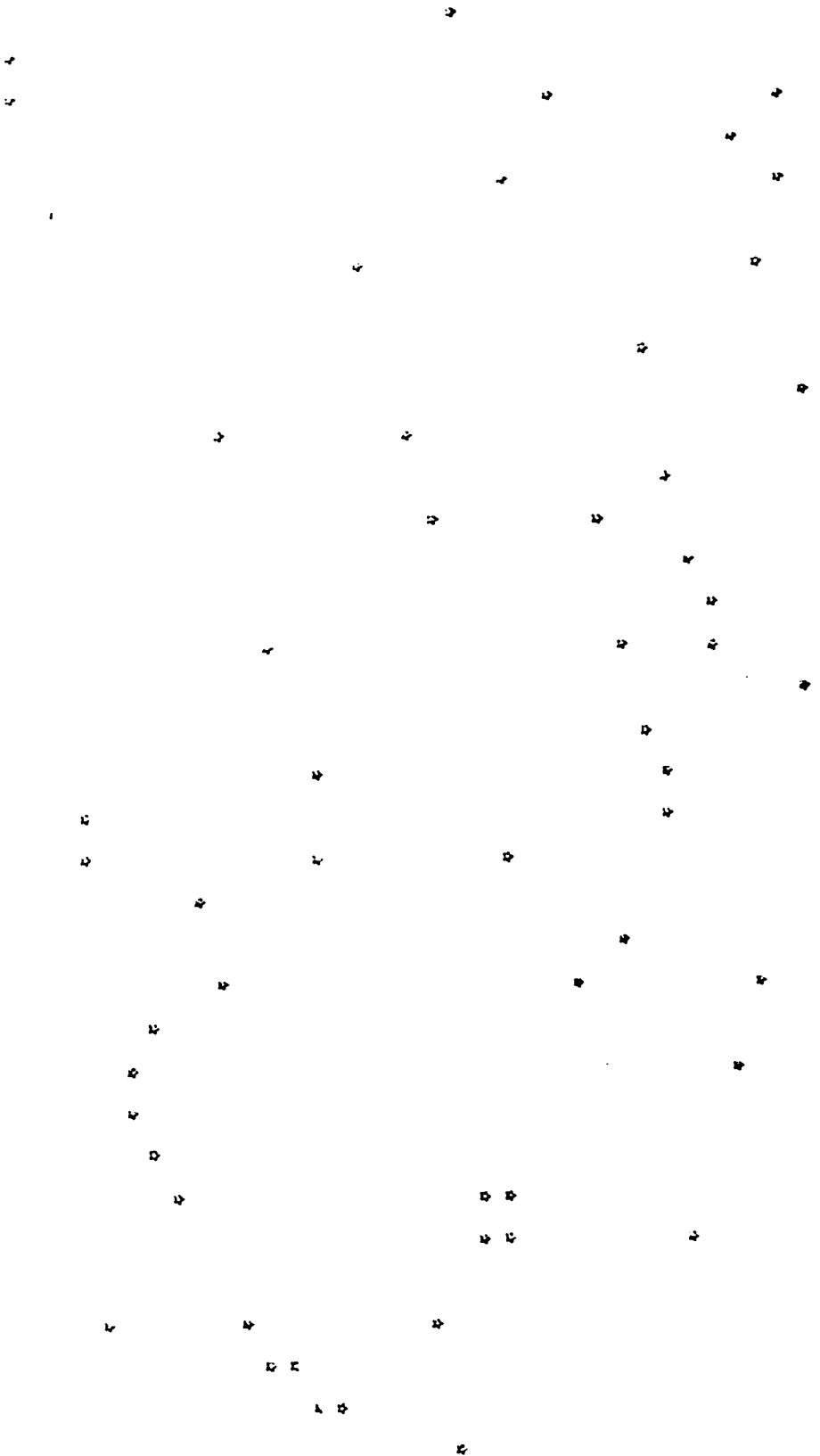


CATISQUARE = 8649

SQUARE OF CORRELATION COEFFICIENT = 0.574683402

FIGURE 16 - 3

WEIBULL DISTRIBUTED RANDOM VARIABLE WITH SHAPE PARAMETER C = 3.0



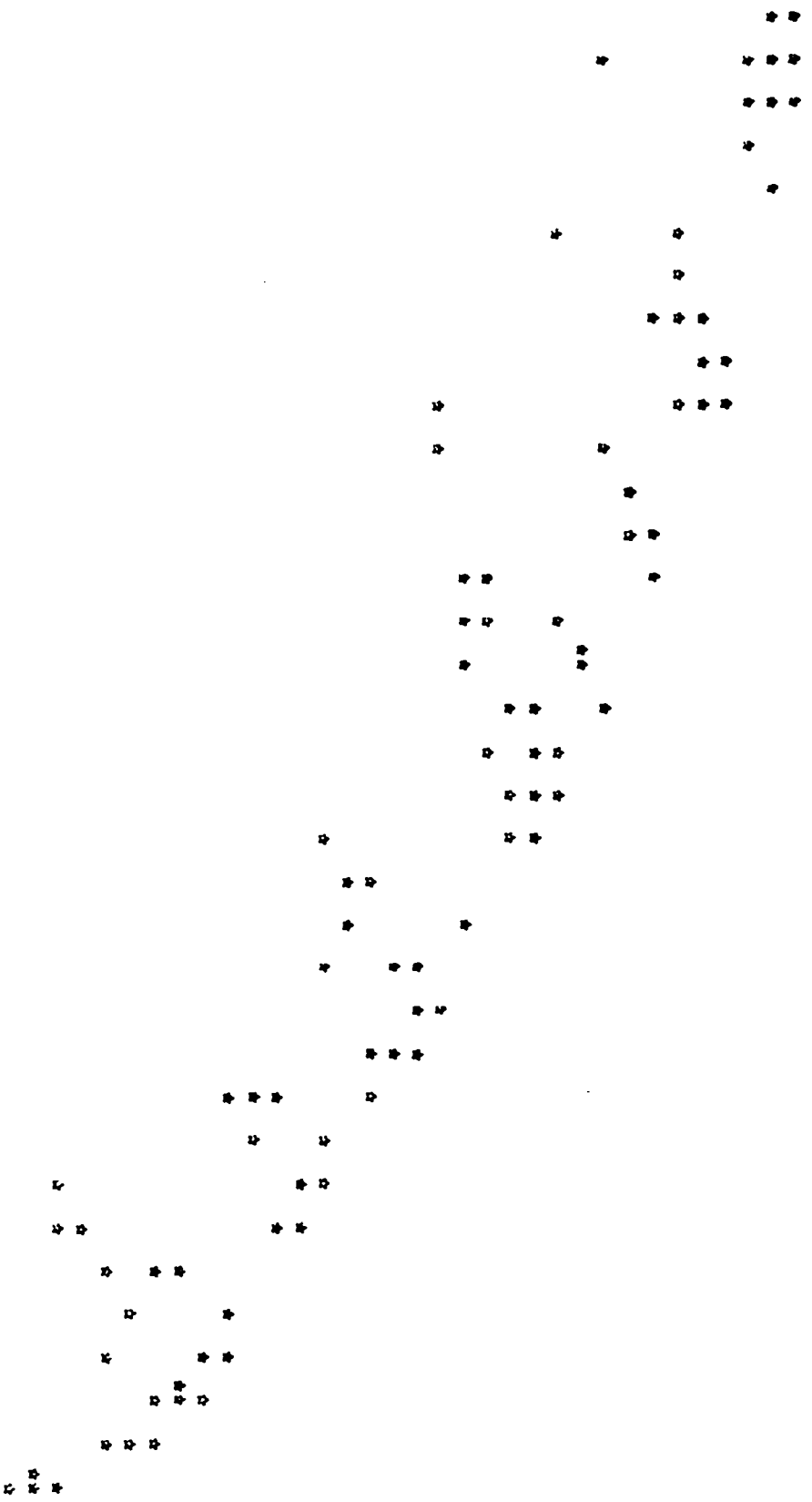
CHISQUARE = 2914

SQUARE OF CORRELATION COEFFICIENT = 0.0091623920

FIGURE 17 - 8

WEIBULL DISTRIBUTED RANDOM VARIABLE WITH SHAPE PARAMETER C = 1.0

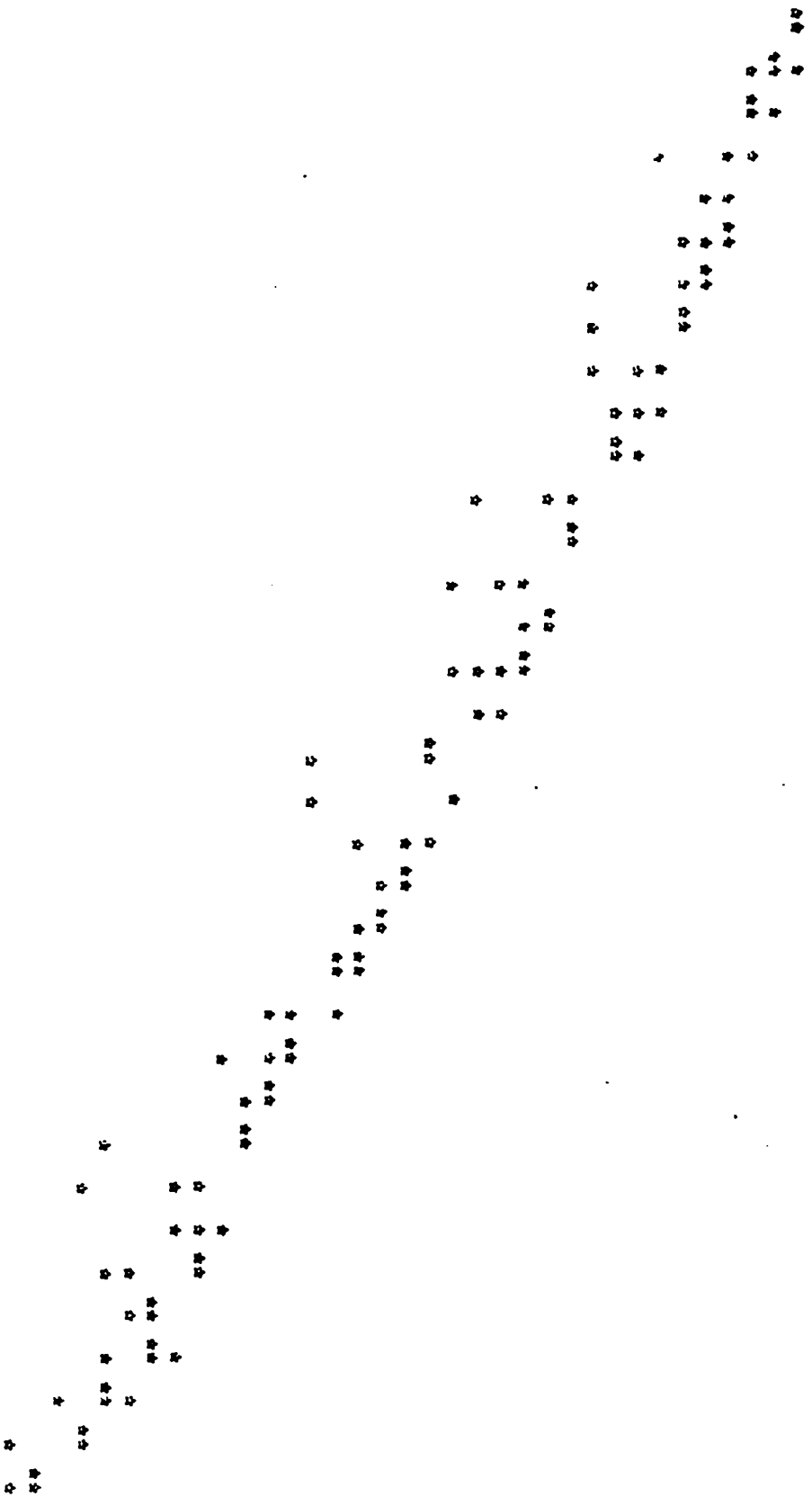
77



CHISQUARE = 5234

SQUARE OF CORRELATION COEFFICIENT = 0.348837980

WEIBULL DISTRIBUTED RANDOM VARIABLE WITH SHAPE PARAMETER C = 0.5

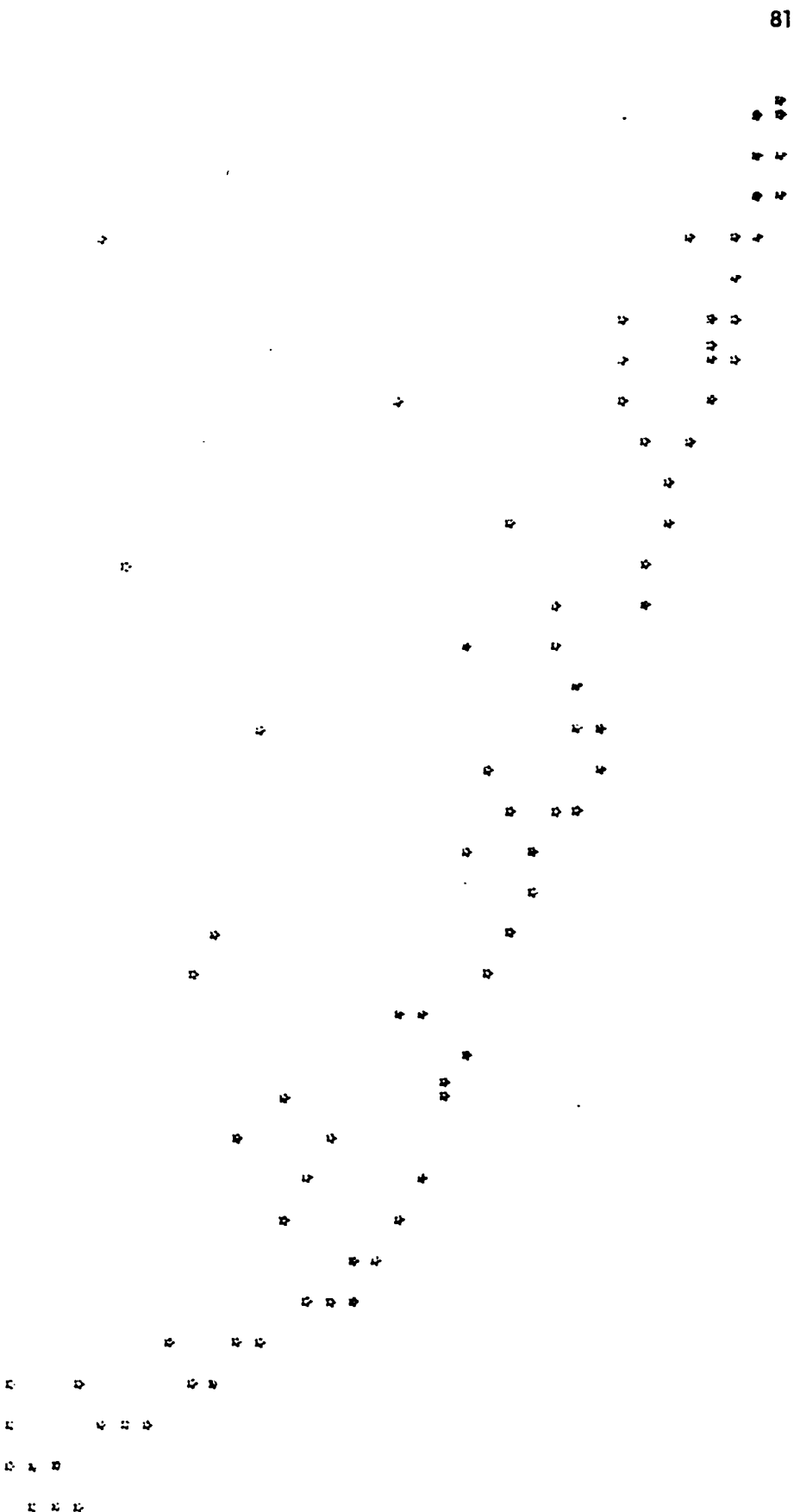


CHISQUARE = 121.76

SQUARE OF CORRELATION COEFFICIENT = 0.593332385

FIGURE 19 - 8

RANDOM VARIABLE X = Z + 2Z WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE



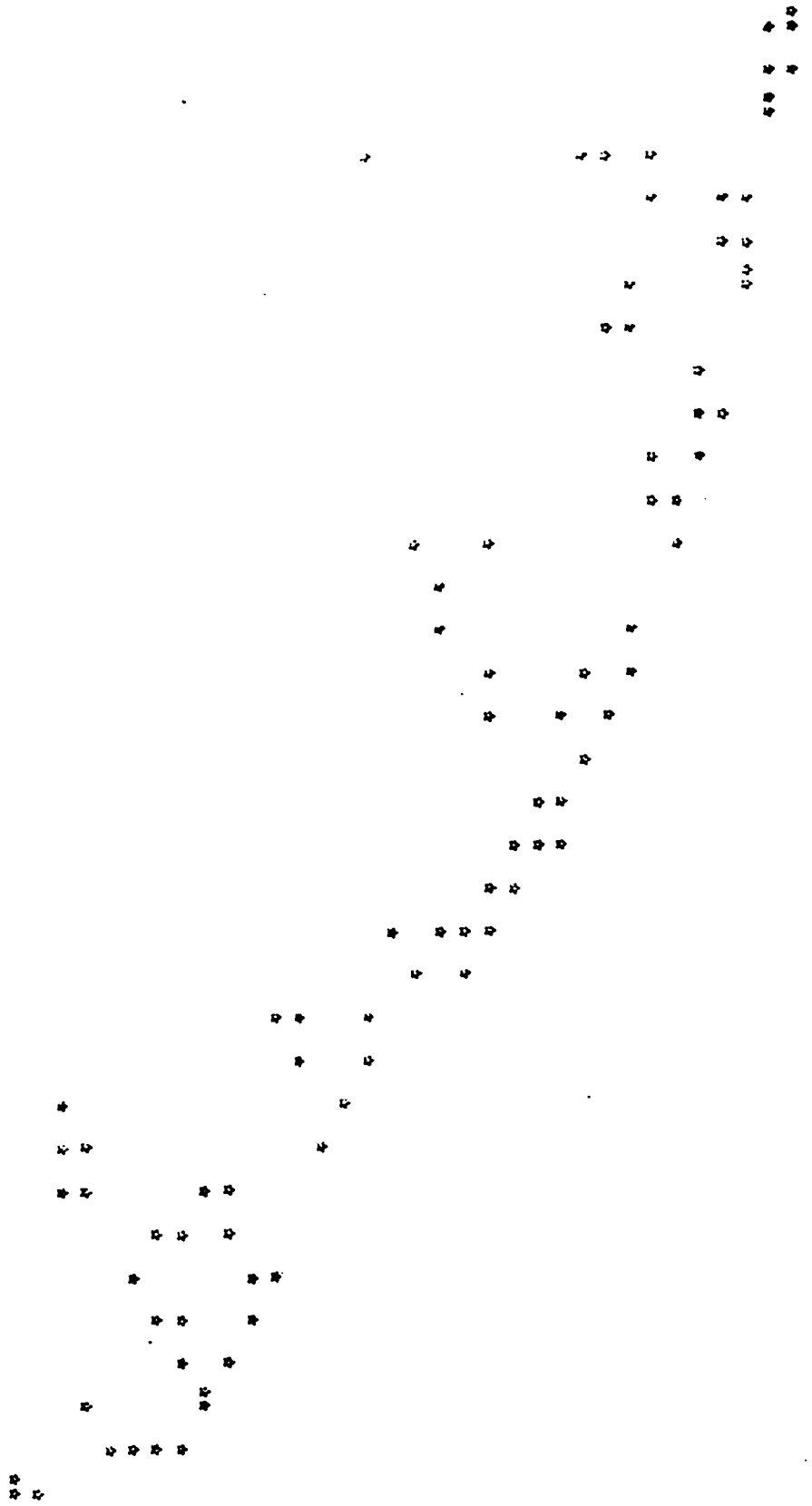
CHI SQUARE = 6309

SQUARE OF CORRELATION COEFFICIENT = 0.18540497

FIGURE 20 - 8

RANDOM VARIABLE $X = 0.5Z + 0.5Z^2$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

83

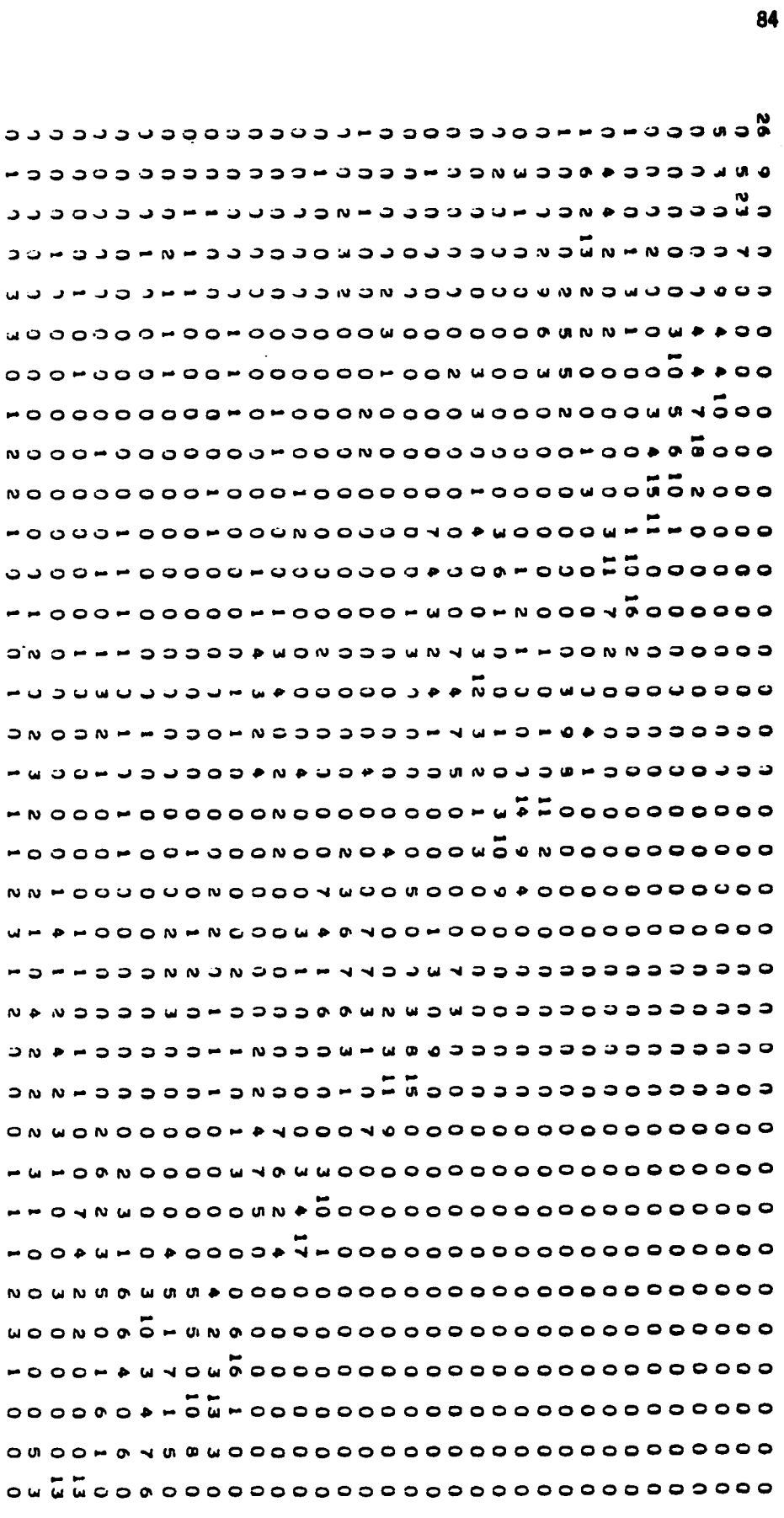


CHI SQUARE = 6808

COEFF OF CORRELATION COEFFICIENT = 0.380413055

FIGURE 21 - A

RANDOM VARIABLE X = $0.75Z + Z^2$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

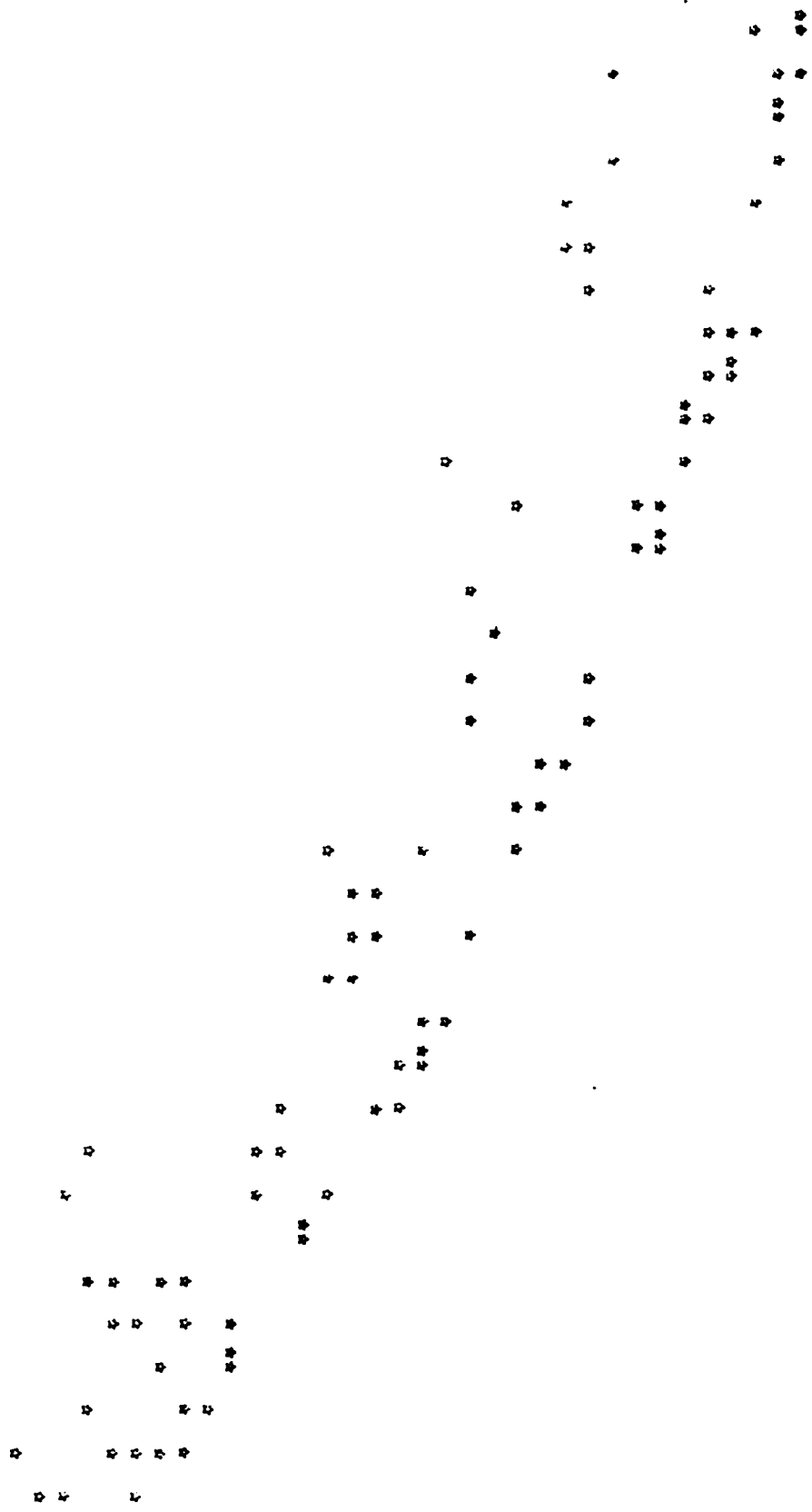


CHI-SQUARE = 7716

SQUARE OF CORRELATION COEFFICIENT = N.323101401

FIGURE 21 - 8

RANDOM VARIABLE $X = 0.75Z + 0.25Z^2$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE



85

CHISQUARE = 7716

SQUARE OF CORRELATION COEFFICIENT = 0.323101401

FIGURE 22 - B

RANDOM VARIABLE $X = 0.94Z + 0.12Z^2$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE



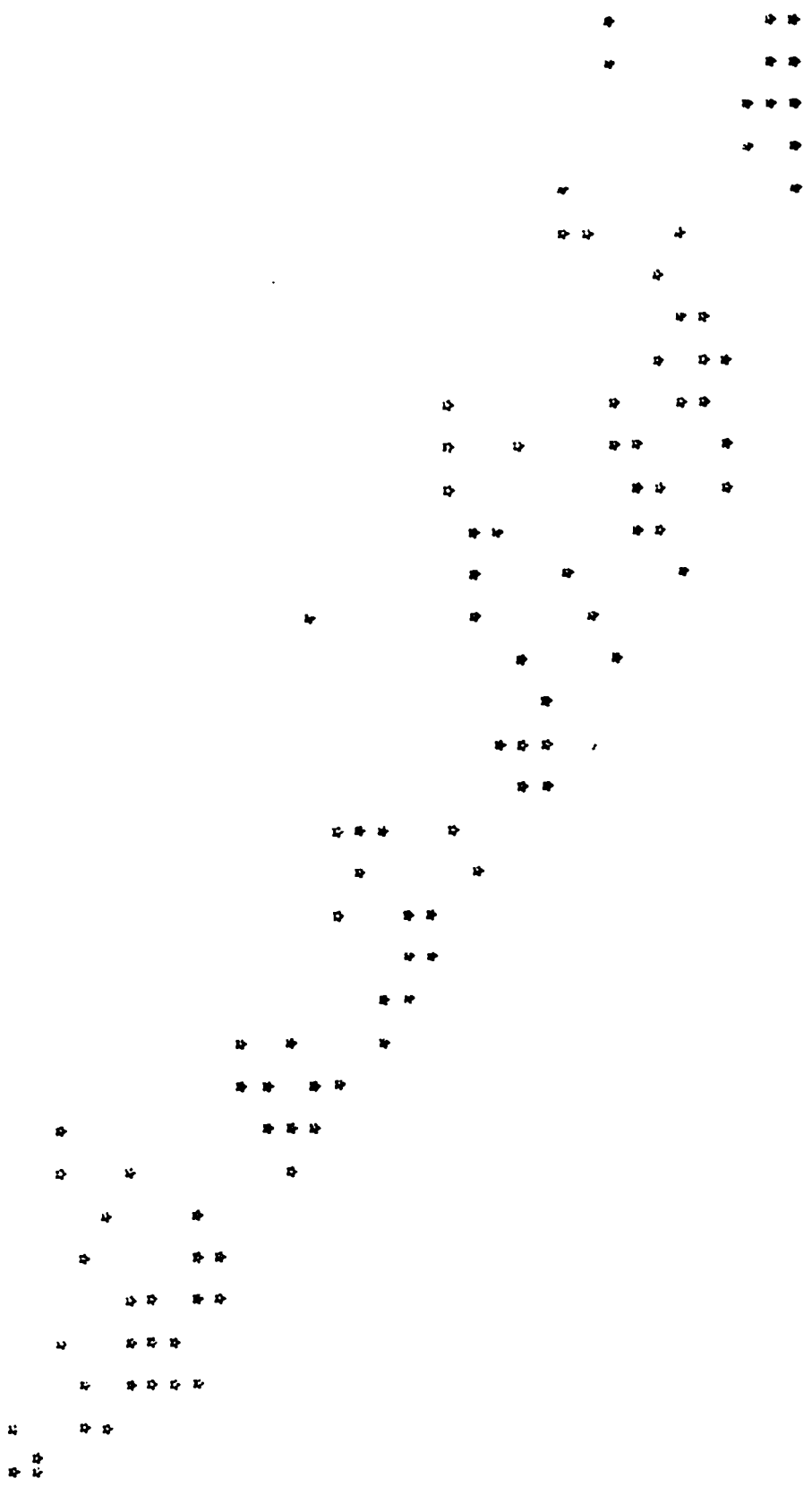
CHISQUARE = 7114

SQUARE OF CORRELATION COEFFICIENT = 0.7673080643

FIGURE 23 - 9

RANDOM VARIABLE X = EXP(Z) WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

89



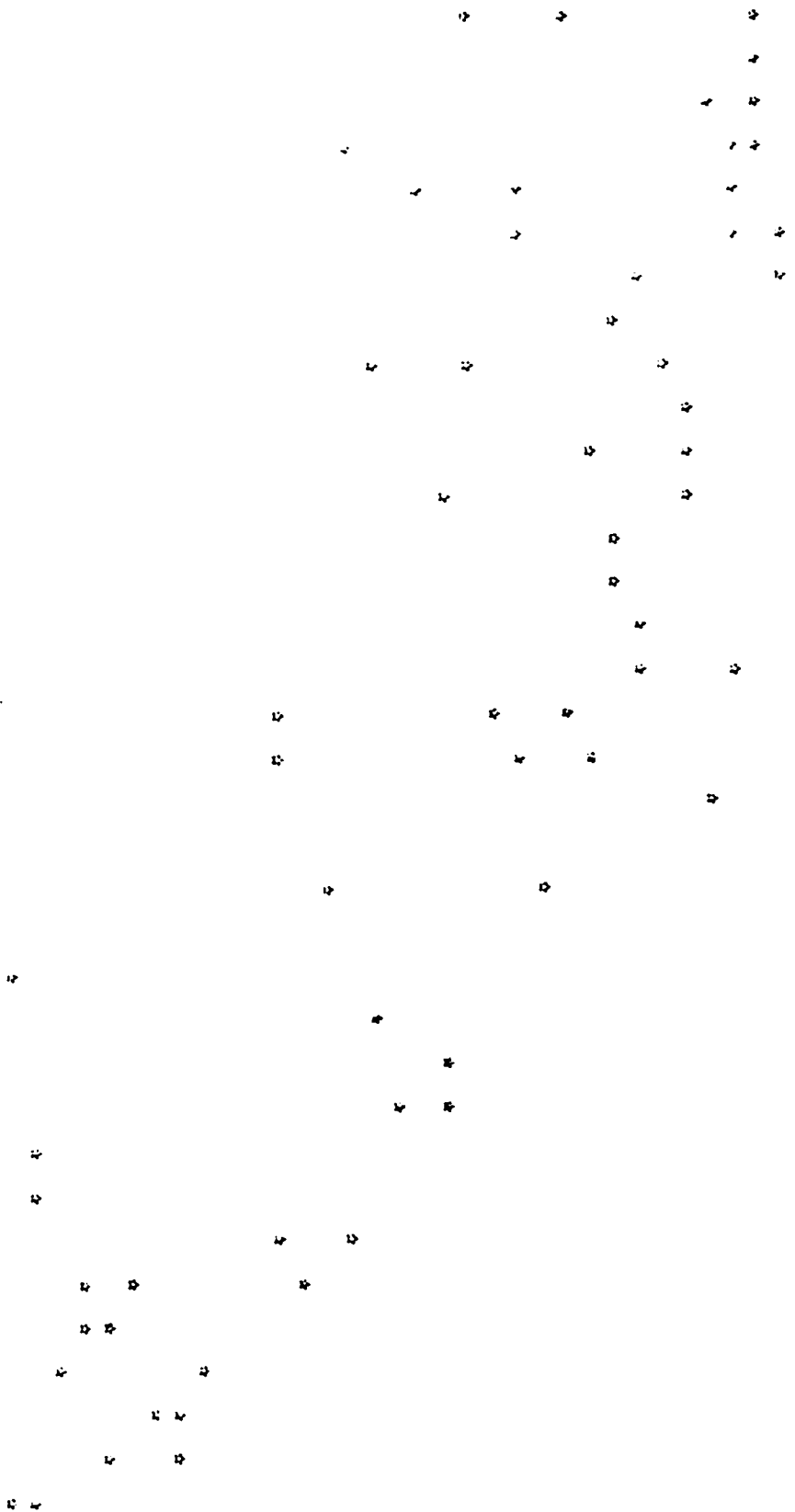
R-SQUARE = 6056

SQUARE OF CORRELATION COEFFICIENT = 0.486233592

FIGURE 24 - B

RANDOM VARIABLE X = EXP(0.592 Z) WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

91



CHI SQUARE = 3.74

SOURCE OF CORRELATION COEFFICIENT = 0.213813047

FIGURE 25 - 3

RANDOM VARIABLE $X = \text{EXP}(0.25Z)$ WHERE Z IS A NORMAL (0.1) RANDOM VARIABLE

93

CHISQUARE = 2772
SQUARE OF CORRELATION COEFFICIENT = 0.0777051449

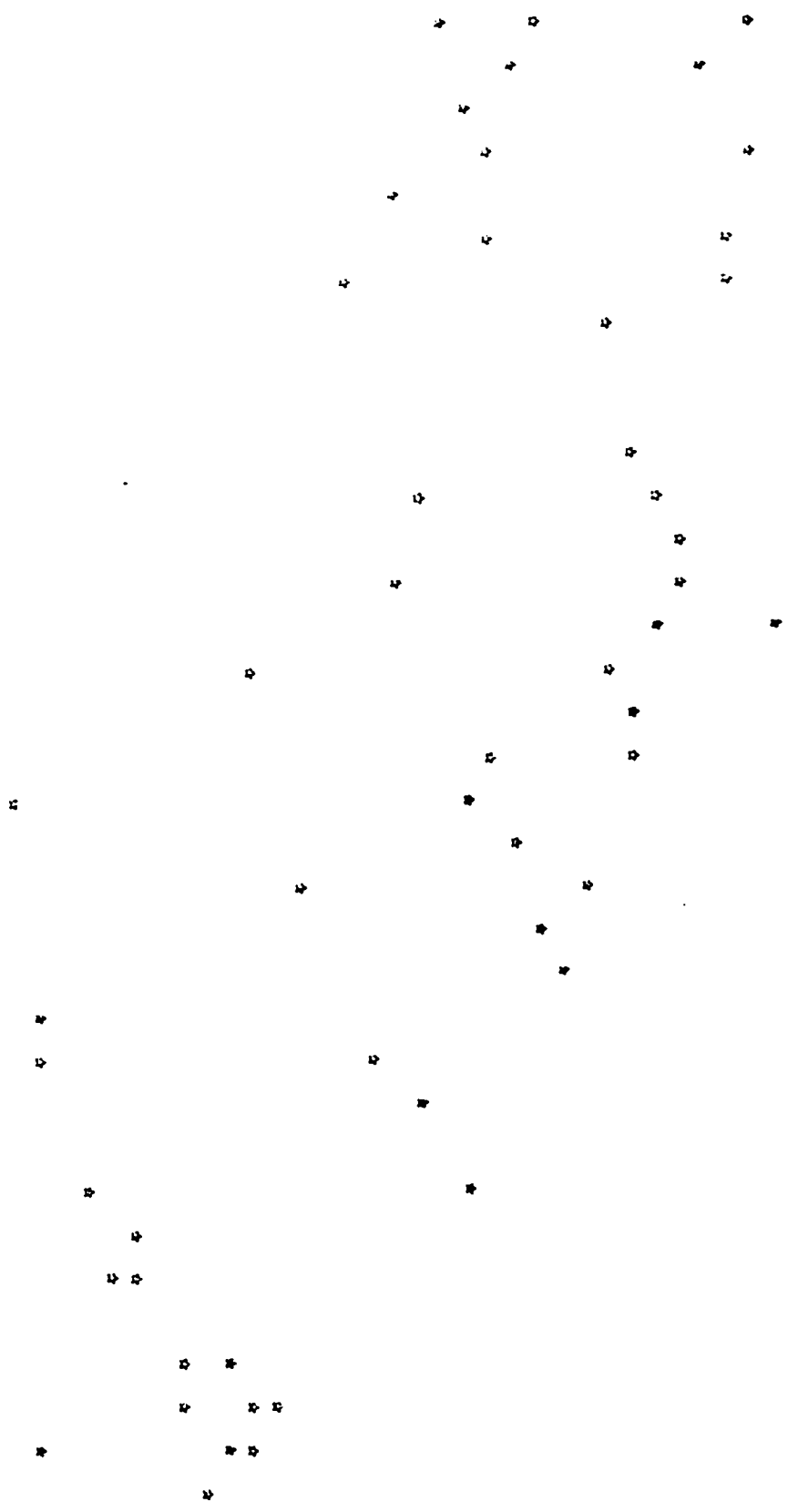
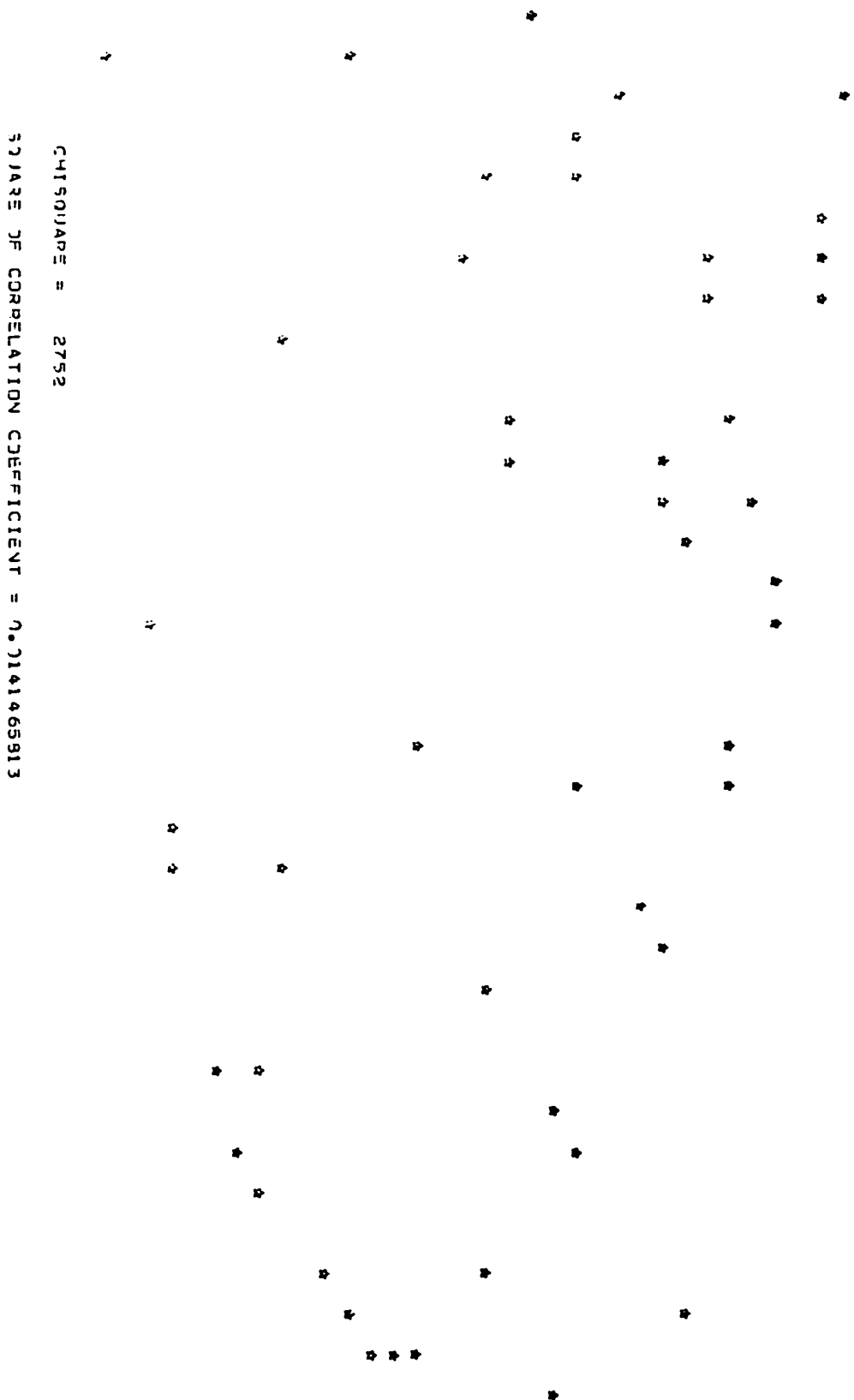


FIGURE 26 - B

RANDOM VARIABLE $X = \text{EXP}(0.14Z)$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

95



RANDOM VARIABLE X = Z*EXP (Z) WHERE Z IS A NORMAL (0.1) RANDOM VARIABLE

FIGURE 27 - B

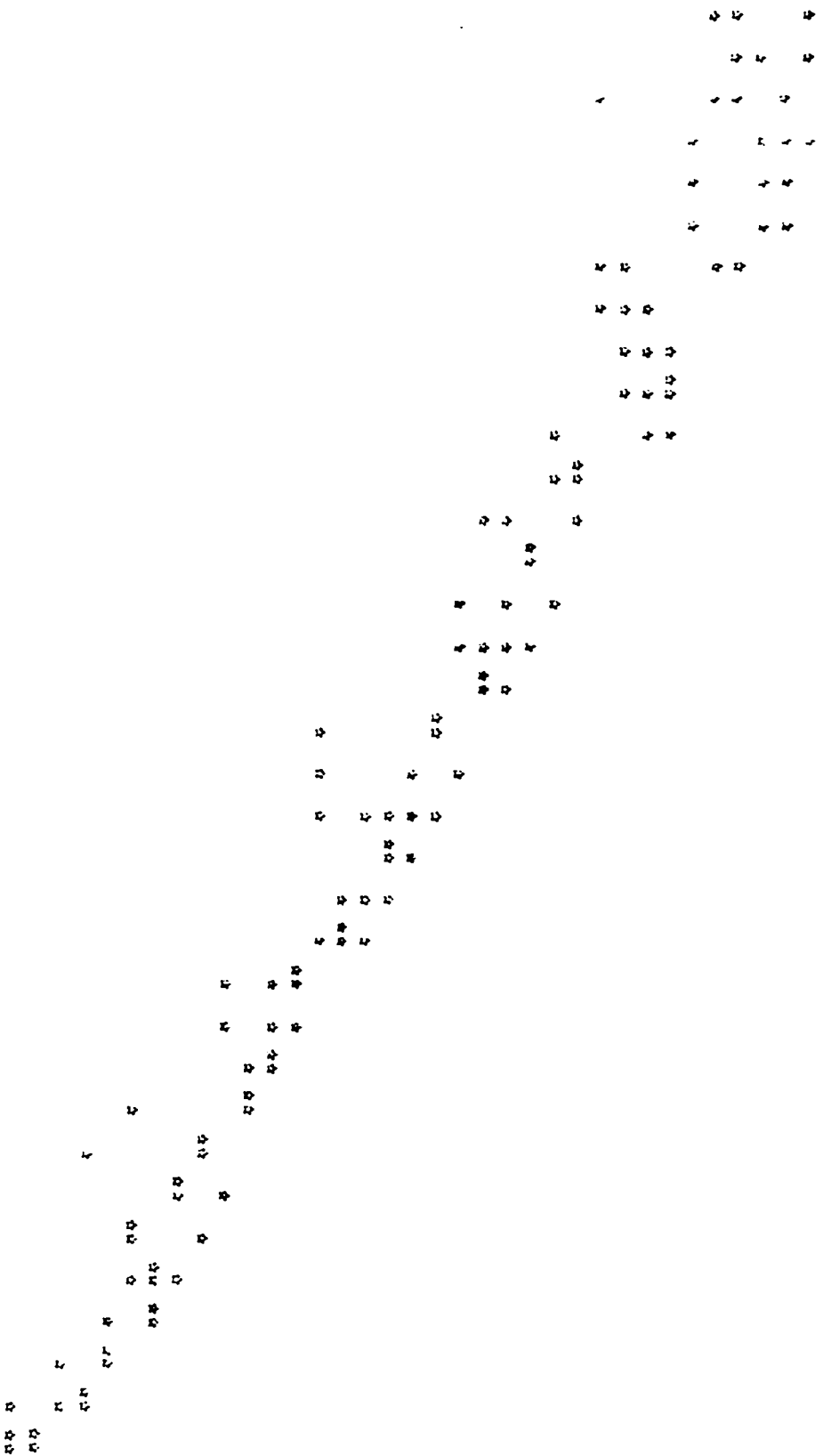
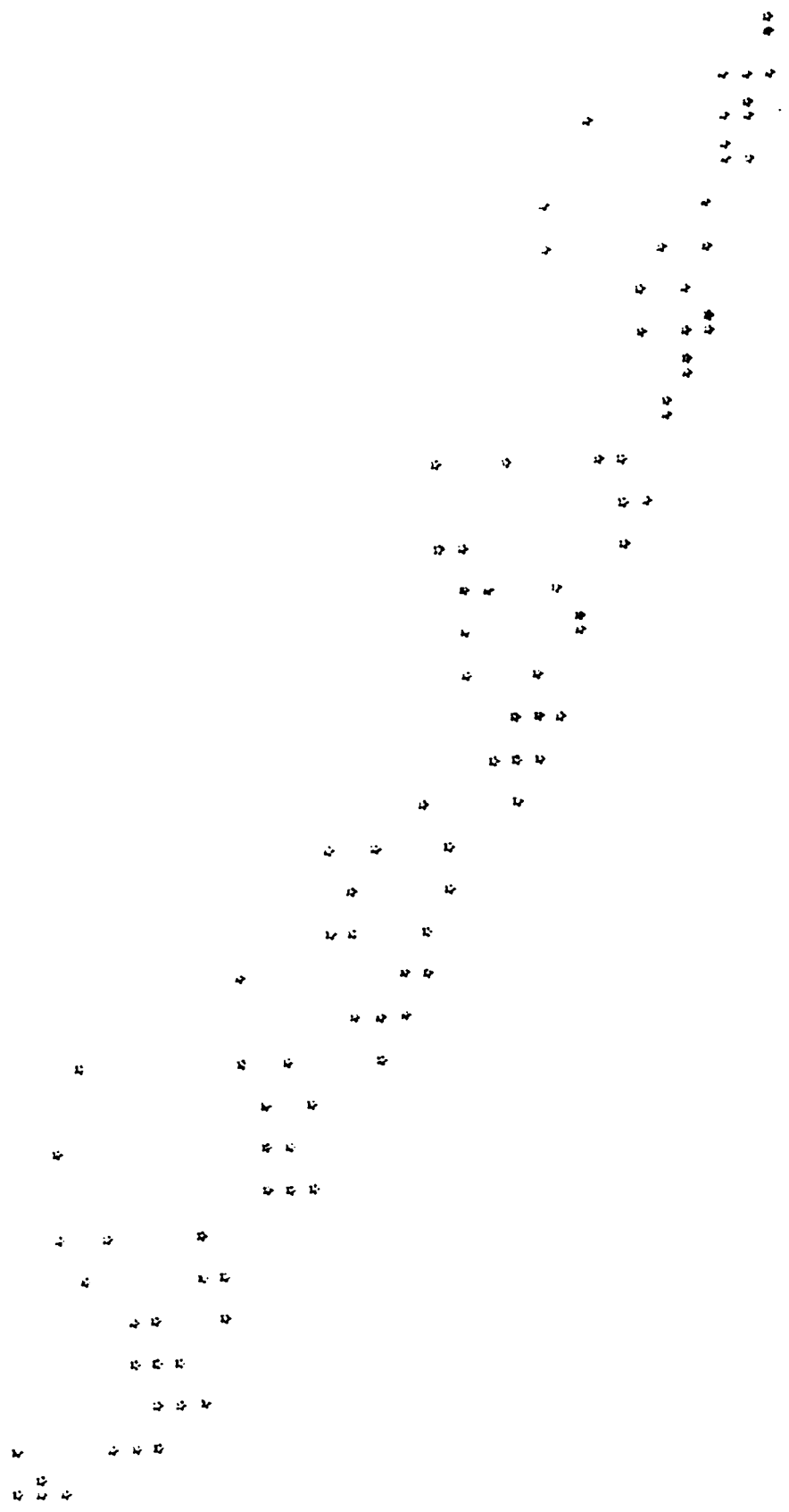


FIGURE 29 - B

RANDOM VARIABLE $X = Z \exp(0.5Z^2)$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

99

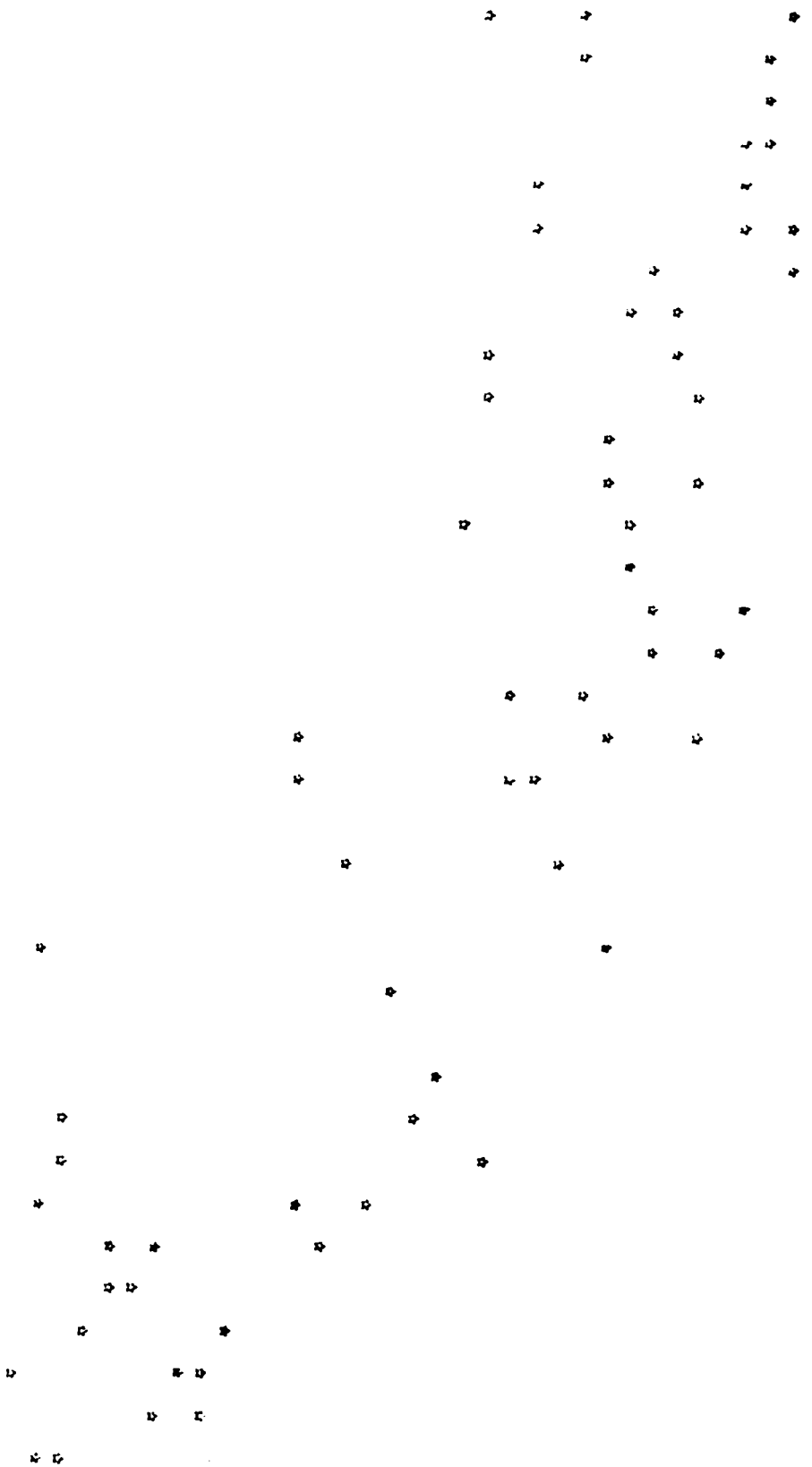


CHI-SQUARE = 92.21

SQUARE OF CORRELATION COEFFICIENT = 0.355606150

RANDOM VARIABLE = $X = Z \text{EXP}(0.25Z)$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

FIGURE 29 - 8

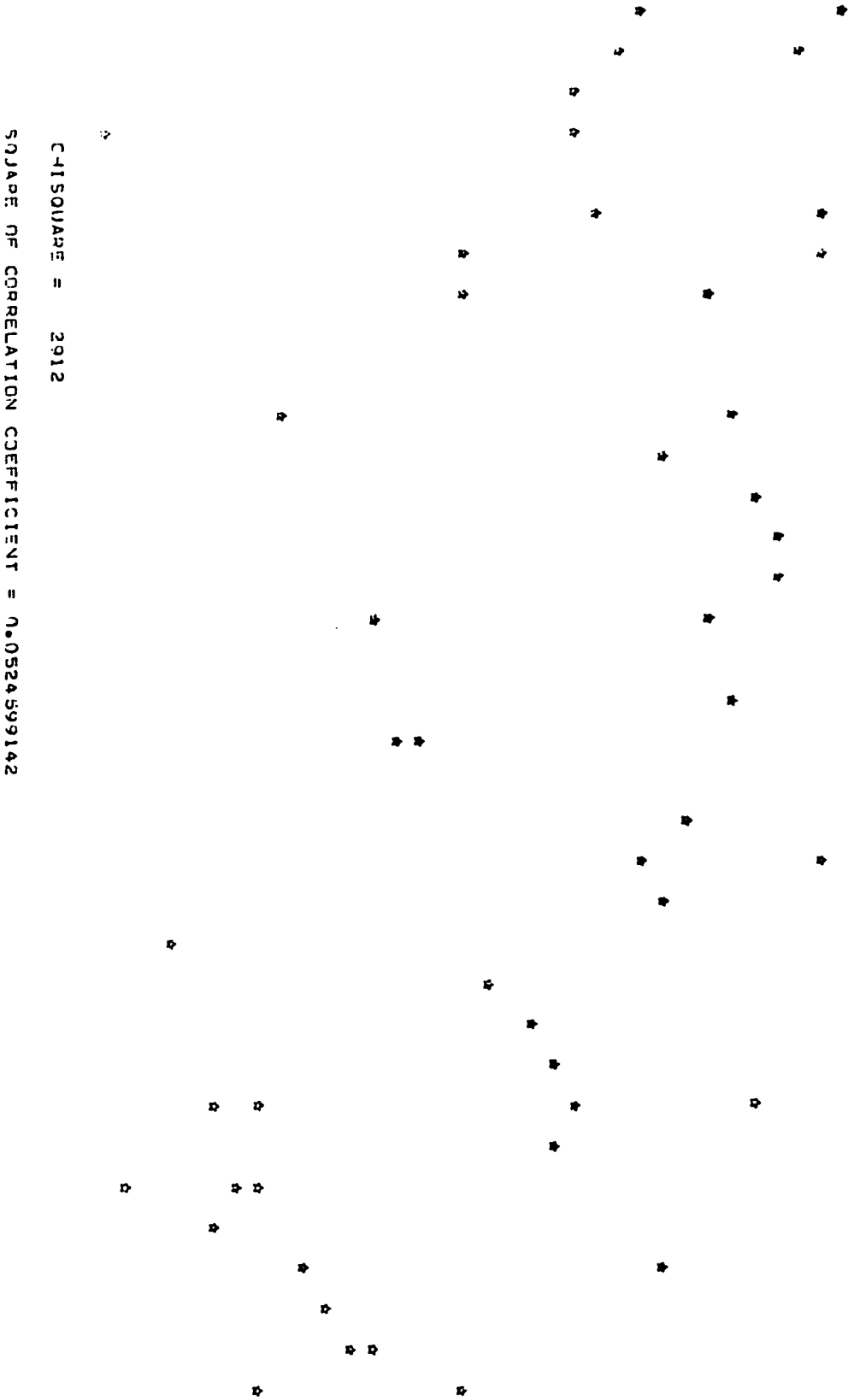


MEAN SQUARE = 3774

SQUARE OF CORRELATION COEFFICIENT = 0.22399239

RANDOM VARIABLE $X = Z \exp(0.1 + Z)$ WHERE Z IS A NORMAL (0,1) RANDOM VARIABLE

FIGURE 30 - B



REFERENCES

1. Abramowitz, M. and Stegun, I.A., Handbook of Mathematical Functions, Dover, New York (1970).
2. Andrews, D.F.; Bickel, P.J.; Hampel, F.R.; Huber, P.J.; Rogers, W.H. and Tukey, J.W., Robust Estimates of Location; Survey and Advances, Princeton University Press (1972).
3. Churchill, E., "Information Given By Odd Central Moments", Ann. Math. Stat. 17 (1946), 244-246.
4. Conover, W.J., Practical Nonparametric Statistics, John Wiley, New York (1970).
5. Finch, S.J., "A Robust Univariate Test of Symmetry", JASA 72 (1977), 387-392.
6. Hastings, N.A.J. and Peacock, J.B., Statistical Distributions, Butterworth, London (1975).
7. Hollander, M. and Wolfe, D.A., Nonparametric Statistical Methods, John Wiley, New York (1973).
8. Johnson, N.L. and Kotz, S., Distributions in Statistics: Continuous Univariate Distributions - 1, Houghton Mifflin, Boston (1970).
9. Johnson, N.L. and Kotz, S., Distributions in Statistics: Continuous Univariate Distributions - 2, Houghton Mifflin, Boston (1970).
10. Kagan, A.M.; Linnik, Yu. V., and Rao, C.R., Characterization Problems in Mathematical Statistics, John Wiley, New York (1973).

11. Mood, A.M.; Graybill, F.A. and Boes, D.C., Introduction to the Theory of Statistics, McGraw-Hill Kogakusha, Tokyo (1974).
12. Shapiro, S.S.; Wilk, M.B. and Chen, H.J., "A Comparative Study of Various Tests for Normality", JASA, 63 (1968), 1343-1372.