SCATTER DIAGRAMS FOR DATA ANALYSIS USING A CHARACTERIZATION PROPERTY

MAGDY KHEDR*

SUMMARY

Given a sample from any of a variety of theoretical distributions, the problem of choosing one for the analysis of the given data has long been a major concern to both the theoretical and the applied statistician. The major topic of this paper addresses this topic as applied to classes of distributions. More specifically, given the data, we wish to classify the parent distribution as either normal-tailed, long-tailed, or skewed in a particular direction.

The basis for the classification is a characterization of the normal distribution. Essentially, the indepdence of the sample means and variance is under consideration here. Graphical methods indicate that the relationship between these variables can distinguish the above three classes.

A two-dimensional statistics, (R_1, R_2), is developed from a second order, linear regression model. Empirical power investigations, using the normal, chi-square with two degrees of freedom, and Cauchy distributions as representatives of their respective classes, exhibit the strength of this statistic.

1. REVIEW OF LITERATURE

In the last century, Gauss' theory of errors was extensively used to justify analysis using the normal distribution. Problems began to appear when Karl Pearson, working in biometrices, and Edgeworth, in economics, found the normal distribution inadequate in describing their data. This observation led

* Institute of Statistical Studies and Research, Cairo University.

Pearson to a study of sample moments. Contributions of Pearson and others are reviewed below:

a) Standardized Third and Fourth Moment Tests

The basic idea behind testing with sample moments is quite simple. Shape parameters, β_1 and β_2, which are functions involving the third and fourth population moments, measure the skewness and the kurtosis of the distribution. Graphic methods, usually using $\sqrt{\beta_1}$ and β_2 can be used to discriminate between different distributions. Examples of such methods can be found in Ord (1967), for discrete distributions, and Egon Pearson and Please (1975), for continuous.

Estimators, b_1 and b_2, of the above parameters are used to determine distribution given only sample information. D'Agostino and Egon Pearson (1973) purpose a test by considering transformations of $\sqrt{b_1}$ and b_2. Under the hypothesis of normality, these transformations are each Chi-square with one degree of freedom. The test is completed by looking up a Chi-square distribution with two degrees of freedom. However, since $\sqrt{b_1}$ and b_2 are not independent the test is only approximate. Bowman and Shenton (1975) modify the above work with two new transformations which Monte Carlo studies indicate are very nearly independent. Again, the result is a test based on the Chi-square distribution with two degrees of freedom.

b) Ratios of Spread Estimate and the Shapiro-Wilk Statistic

Another basic approach to testing for normality considers ratios of spread estimates. The first of these was suggested by Geary (1935). His statistic was the ratio of the mean deviation to the standard deviation in a sample. Power studies conducted by D'Agostino and Rosman (1974) indicate the usefulness of this ratio as a quick test for normality. Geary's statistic was simplified by David, Hartley and Pearson (1954) to the ratio of the range to the standard deviation.

Central in any discussion concerning ratio statistics for determining normality is the Shapiro-Wilk (1965) statistic W. Subsequent power studies have established W as an effective statistic; however, for sample of size over 20, W can only be found by approximation. To remedy this problem, several modifications have been suggested by Shapiro and Francia (1972), Weisberg and Bingham (1975), Filliben (1975).

c) Distance Criteria, the Kolmogorov-Smirnov Statistic

Criteria concerning the distance between the expected and the actual observations have long been used to test distributional fit. An example would be the Chi-square test. This test possesses two major difficulties, first, the intervals are arbitrary and second, it is incovenient for small sample sizes.

Other tests based on such criteria usually employ the empirical distribution, EDF for brevity. Anderson and Darling (1954) consider the weighted average of the squared distances between the EDF values and the values under the null distribution, $\{F_n(x)-F_o(x)\}^2$. Here, the null distribution is of any form, but must be fully specified.

Better known is the Kolomogorov-Smirnov statistic, D. Kuiper, see Stephens (1974) suggested modifying D by V. A further modification was suggested by Finkelstein and Schafer (1971) which proves more powerful.

d) Tests Based on Characterizations, the McDonald-Katti Test

Construction of tests based on characterization of the hypothesized distribution present a perplexing dilemma to statisticians. Since characterizations are usually based on some form of independence, testing the characterization poses as large a problem as testing for the distribution itself.

Lock (1976) developed a test based on a characterization of the gamma distribution. A study rather similar to Locke's was undertaken by McDonald and Katti (1974) for the normal distribution. They began by considering the well known characterization of the normal distribution of the independence of sample mean and variance. From this, it has been shown that the sum and the absolute difference of two independent observations also characterizes the normal.

McDonald and Katti found simulated distributions for these tests. As such, only approximate percentage points can be found. An empirical power study shows that for different alternative distributions "best" of the proposed tests changes. The results of this study were favorable when comparing the three tests against Kolmogorov-Smirnov, the $\sqrt{b_1}$, and b_2 tests. However, the Shapiro-Wilk statistic proved at least as good in all but one alternative, the binomial distribution with n=4 and p=0.5.

2. PRELIMINARY INVESTIGATION

Preliminary studies began with several questions concerning the McDonald-Katti paper. Do other long-tailed, symmetric distributions have graphs similar to the Cauchy? Also, can distributions with long-tails, but not quite as long as the Cauchy's, be distinguished from the normal? Finally, can a testing procedure be formed to distinguish between all of these simultaneously, using the graphs as the basis?

Answering such questions is the focus of this paer. To resolve the first two, an empirical examination of Rogers and Tukey's class should prove satisfactory. As for the last, attention is restricted to the normal, Chi-square and Cauchy distributions. Here, if the test statistic indicates independence of sample mean and variance, the data is classifed as normal-tailed. Otherwise, that is, if the test statistic fails to support

such independence, the population is classified either skewed or long-tailed according to test results.

a) The Tukey Ratio System of Distributions

Rogers and Tukey (1972) describe a class of long-tailed, symmetric distributions. This class consists of ratios, X=W/V where W follows the standard normal distribution, V is independent of W and the forms chosen for V generate the various distributions of X. Additional distributions are obtained through mixing as explained below. A few members have special names, while others are referred to by their symbols.

The members of the class of distributions of X under consideration are given in Table 1.

In view of the complexity of the construction of this family, family, it was necessary to verify that the distributions we had obtained through the interpretations were the same as their. Although this was virtually unnecessary for some of the distributions, but it may be imperative for others, for example the Gucumatz. A list of selected percentage points for standarized versions have been provided by Rogers and Tukey. In view of the unknown parameters, Tykey's percentile T is related in this paper to X through T = a + bX. So, we estimate a and b using two percentile points. Additional details are as follows: The C, G and L are well known distributions and provide no problem. Distributions for the probability mixtures are found by merely multiplying the component distributions by their respective probabilities, for example $F_{S/4}(t)=0.25\ F_S(t)+0.75\ F_G(t)$. The Gucumatz distribution is given by

$$F_{GUCU}(t)=\begin{cases} P_1F_c(t) & \text{if } t<-1 \\ P_1F_c(-1)+P_2(F_G(t)-F_G(-1)) & \text{if } -1\leq t<1 \\ P_1F_c(-1+P_2(F_G(1)-F_G(-1))+P_1(F_c(t)-F_c(1)) & \text{if } t\geq 1 \end{cases}$$

Table 1

A List of Tukey Ratio System of Distributions

Symbol	Name	Explanation
C	Cauchy	V is standard normal
G	Standard Normal	V=1 with problability 1
S	Slash	V is uniform on (0,1)
S/4	S/4	A probability mixture of 25% S and 75% G, corresponding to V=1 with probability 0.75 and V uniform on (0,1) with probability 0.25.
Q	Q	V is distributed triangularly on (0,1) with vanishing density at 0.
QS	QS	A 50% - 50% mixture of Q and S, corresponding to V trapezoidally (altitude 0.5 at 0, 1.5 at 1) distributed on (0,1).
3S/4	3S/4	A probability mixture where V=1 with probability 0.75 and V uniform on (0, 1/3) with propability 0.25.
10G/4	10G/4	V=1 with probability 0.75 and V=10 with probability 9.25.
L	Logistic	This distribution is approximated by a member
GUCU	Gucumatz	It is constructed by taking an appropriate function (about 69%) of the standard normal between-1 and +1 and another appropriate fraction (about 105%) of that part of the Cauchy outside these limits.

where P_1 and P_2 are subject to the restriant that F_{GUCU} must be a statistical distribution function.

After F_x has been found, its upper 40^{th} and 20^{th} percentiles denoted by $X_{.4}$ and $X_{.2}$ are computed. These are put into the linear relationship with the corresponding percentiles listed by Rogers and Tukey denoted by $t_{.4}$ and $t_{.2}$, yielding,

$$X_{.4} = a + bt_{.4} \quad \text{and} \quad X_{.2} = a + bt_{.2}$$

From these two equations, the constants a and b are found. Now, setting $T = a + bX$, we have a standardized version of X. Now, given the $F_T(t) = F_X\{(t-a)/b\}$, the remaining percentiles are easily checked.

An example might explain this procedure more clearly. F_S is found using the standard statistical techniques as:

$$F_S(x) = \begin{cases} (\exp(-x^2/2)-1)/x + F_G(x) & \text{for } x \neq 0 \\ .5 & \text{for } x = 0 \end{cases}$$

Using simple iterative techniques, $x_{.4}$ and $x_{.2}$ are found to be .512 and 1.946 respectively. Hence, we have

$$.512 = a+b\ (.256) \text{ and } 1.946 = a+b\ (.973),$$

giving us a=0 and b=2. To check the remaining percentiles, we use

$$F_T(t) = [\exp\{-(t/2)^2/2 - 1\}/(t/2)+F_G(t/2)], \quad t\neq 0$$

Note that, for the GUCU distribution, P_1 and P_2 are found by using the equations

$$F_{GUCU}(t,4) = .6 \quad \text{and} \quad F_{GUCU}(t.2) = .8$$

A comparison of percentage points is given in Table 2. If a difference occurs, Tukey's value is listed below in parentheses. This table also provides the reader with a feel for the various distributions.

Table 2

Percentage Points for the Distributions

Distribution	40%	20%	10%	5%	2%	1%	0.1%
G	.253	.842	1.28	1.64	2.05	2.3	3.09
L	.254	.869	1.38	1.85 (1.84)	2.44	2.88	4.33
Q	.254	.881	1.45	2.10	2.33	4.71 (4.74)	14.9
10G/4	.255	.927	2.14	6.52	10.9	13.6	20.6 (20.8)
S/4	.254	.867	1.40	2.04	4.36	8.72 (8.73)	87
QS	.255	.920	1.67	2.91	6.48	12.3	117
S	.256	.973	1.99 (2.00)	3.99	9.97	19.9	199
GUCU	.256	1.021	2.26	4.62 (4.61)	11.6	23.3 (23.7)	233 (232)
3S/4	.255	.910	1.75	4.62 (5.20)	11.8	23.7 (23.2)	237
C	.259	1.098	2.04	12.7	12.7	25.4	254

The differences are extremely minore, so we may proceed to the additional studies involving Monte Carlo methods.

generation of random samples now poses little problem. Wherever the X-distribution is obtained through simple forms for W and V, we first generate W and V, and then divide. For others, such as L and GUCU, generating a uniform number Y and obtaining X through $FX^{-1}(Y)$ proved easier.

b) Basis of Sample Classification

Now, given a sample, we need a method to determine from which of the class members it has been drawn. Focussing on the normal as a parent distribution, a well known characterization is the independence of the sample mean and variance. DcDonald and Katti (1974) start their work by investigating this fact. For a sample $x_1, x_2, \ldots, x_n$ of even size n, their preliminary scatter diagrams utilize the independence of the sum and the absolute difference between x_i and x_{i+1} for i=1,3,...,n-1.

The rational for this stems from the fact that a sample, no matter what size, has only one mean and one variance. Hence, if we plot $\bar{x}$ against S^2, we would have only one point. With this sole point, the graph cannot shed much light on the independence of these two variables. Subsequently, an overall sample of size n should be split into m subsamples, giving us m instead of just one point. Of course m should be as large as possible forcing the size of the subsamples to be as small as possible. At least two points are needed to find S^2; therefore, the smallest size for the subsamples is two. For subsamples of size two, using X_i and X_{i+1},

$$\bar{x} = (x_i+x_{i+1})/2 \quad \text{and} \quad S_i^2 = (x_i-x_{i+1})^2/2$$

which are independent if and only if x_i and x_{i+1} are independent identically distributed normal observations. Hence, the variables

$$y_i = x_i + x_{i+1} \quad \text{and} \quad z_i = |x_i-x_{i+1}|, \qquad (1)$$

which are one-to-one function of $\bar{x}_i$ and S_i^2, are also independent.

Since the graph is designed to distinguish distribution, regardless of location or scale parameters, attention must be given to the affect on y_i and z_i made by a linear transformation of X. Let $x_i' = c + dx_i$ and $x_{i+1}' = c + dx_{i+1}$. Define

y_i' and z_i' similar to (1) by $y_i' = x_i' + x_{i+1}'$ and $z_{i+1}' = |x_i' - x_{i+1}'|$. Then, $y_i' = 2c + dy_i$ and $z_i' = d|z_i|$.

In view of this, the origin of the scatter diagram is taken to be (m_y, m_z) where m_y and m_z are respectively the sample medians of the set of y's and z's. The scale of the graph is set so that the difference between the largest and the smallest z is fixed as so many units and the same about the scale on y. With this rule, the graphs focus on trends without regard to location and scale.

c) Preliminary Empirical Results

Using computer techniques, ten random samples, each with n=40, were generated from the distributions mentioned in Section (a). Each sample was then split into twenty pairs using consecutive points. From each set of pairs, a graph was made of y, the pair sums, versus z the absolute pair differences. Thus, we had ten graphs from each distribution.

To set up an order of presentation, we arranged the distributions in the increasing order of their 1% points for the standardized version. The order came out = G, L, Q, 10 G/4, S/4, QS, S, GUCU, 3 S/4, C.

For purposes of comparison, three graphs were chosen from each distribution. One considered "typical", one most like the proceding distribution and one most like the next. To determine which of the ten graphs would be selected for these, a point system based on empirical criteria was employed. Basically, these criteria dealt with outliers, missing points, and the spread of high and low points. For each criterion, a point system awarded -2, -1, 0, 1, or 2 points to a graph depending on the number of graphs from the distribution exhibited said criterion.

For the purposes of presentation, only the typical graphs are supplied as Graphs 1-10. The straight lines are $y=m_y$ and $z=m_z$, circled points depict repeated values, and the number of repetitions is given. Several observations are rather striking.

1) The normal and logistic both lack tails (extreme points). See Graphs 1 and 2.

2) Tails definitely begin to appear in Graph 3 and are prominent in Graphs 5-10.

3) Domination by one tail begins in Graph 4 and is definite in Graphs 5-10.

4) Repeated points is an obvious characteristic in Graphs 6-9, but in much less apparent in Graph 10.

These observations are made to show trends in the graphs going from the normal-tailed to the long-tailed distributions.

3. A STATISTIC TO TEST NORMAL-VS-CAUCHY-VS-CHI-SQUARES

The graphs presented in the last section were provided to give a feel for how the relationship between the variable y and z changes for the various distributions. A quick test can now be performed. First, given a sample $x_1,x_2,\ldots,x_n$ of size n, to find the Tukey distribution that best approximates the parent population, split the sample into (n/2) consecutive pairs, (x_i,x_{i+1}), $i=1,3,\ldots,n-1$. Find (y_i,z_i) using (1) in the last section part (b) and plot the (n/2)y - z pairs. The final decision as to which distribution appears to fit the data the best may be based on some point system or on intuitive understanding of the graphs. The suggestion would be that, if one is interested in such a test, many graphs should be made from samples from known distributions to gain insight in their intrinsic differences.

Being a quick test, there are obvious difficulties which this procedure. However, there are also notable strengths. First, the graphs do show obvious trends. Second, the statistician's personal judgement and experience are utilized. This judgement should be augmented by a thorough understanding of the y-z graph by constructing these graphs for distributions of particular interest.

So, for this large class, there is a general eye-ball scheme. Restricting the attention to only three alternative, normat-tailed, long-tailed, or skewed to the right, we have more organized study which shows much power. This indicates that the quick test can have mcuh value.

The test proposed in this paper not only distinguishes between normality and non-normality, but also indicates whether a non-normal distribution is either long-tailed or skewed. Thus, we are really providing for a test of three hypotheses:

H_1: the data is distributed normally

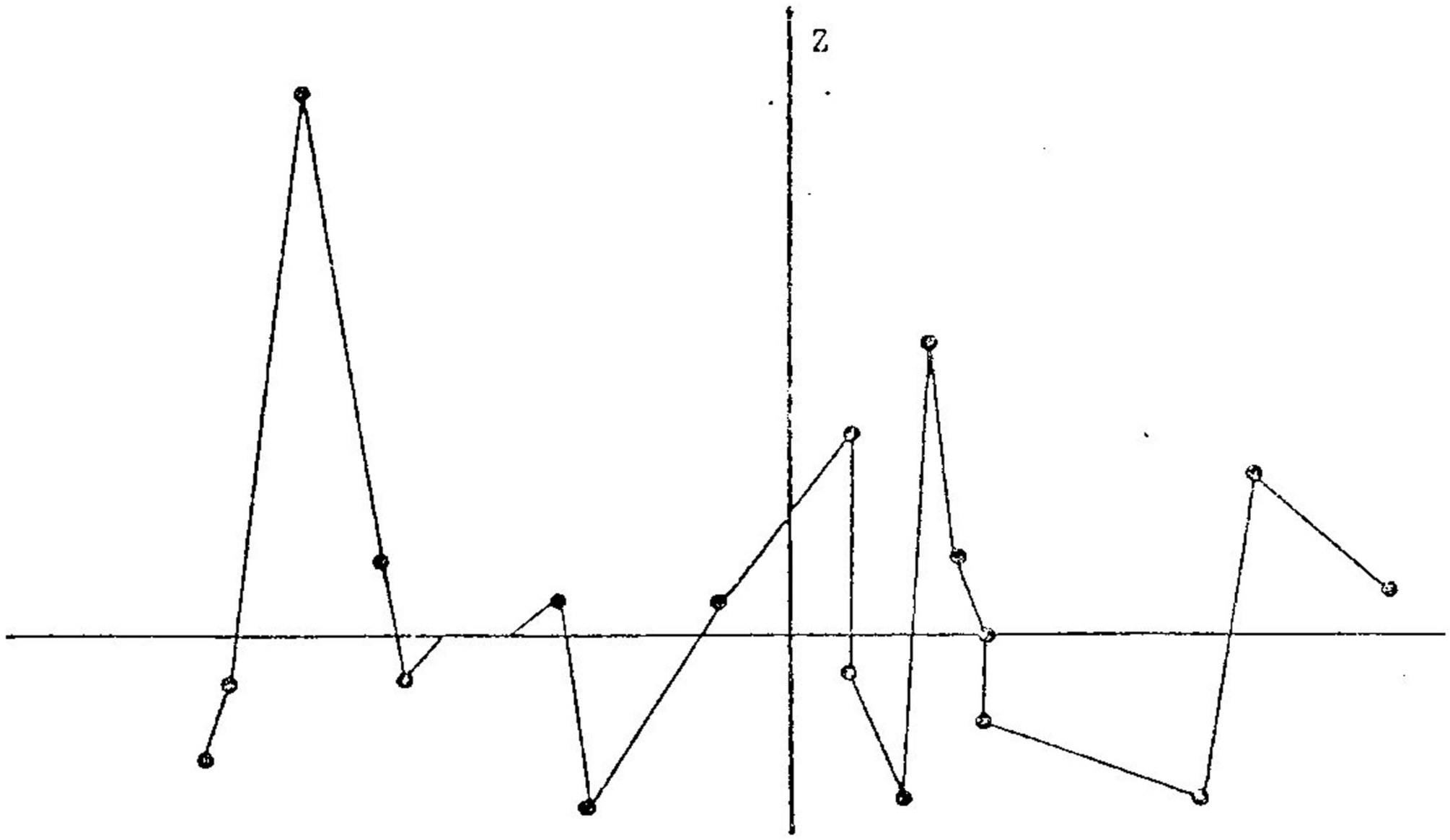
H_2: the data is distributed as a chi-square with 2 degrees of freedom

H_3: the data is distributed as Cauchy.

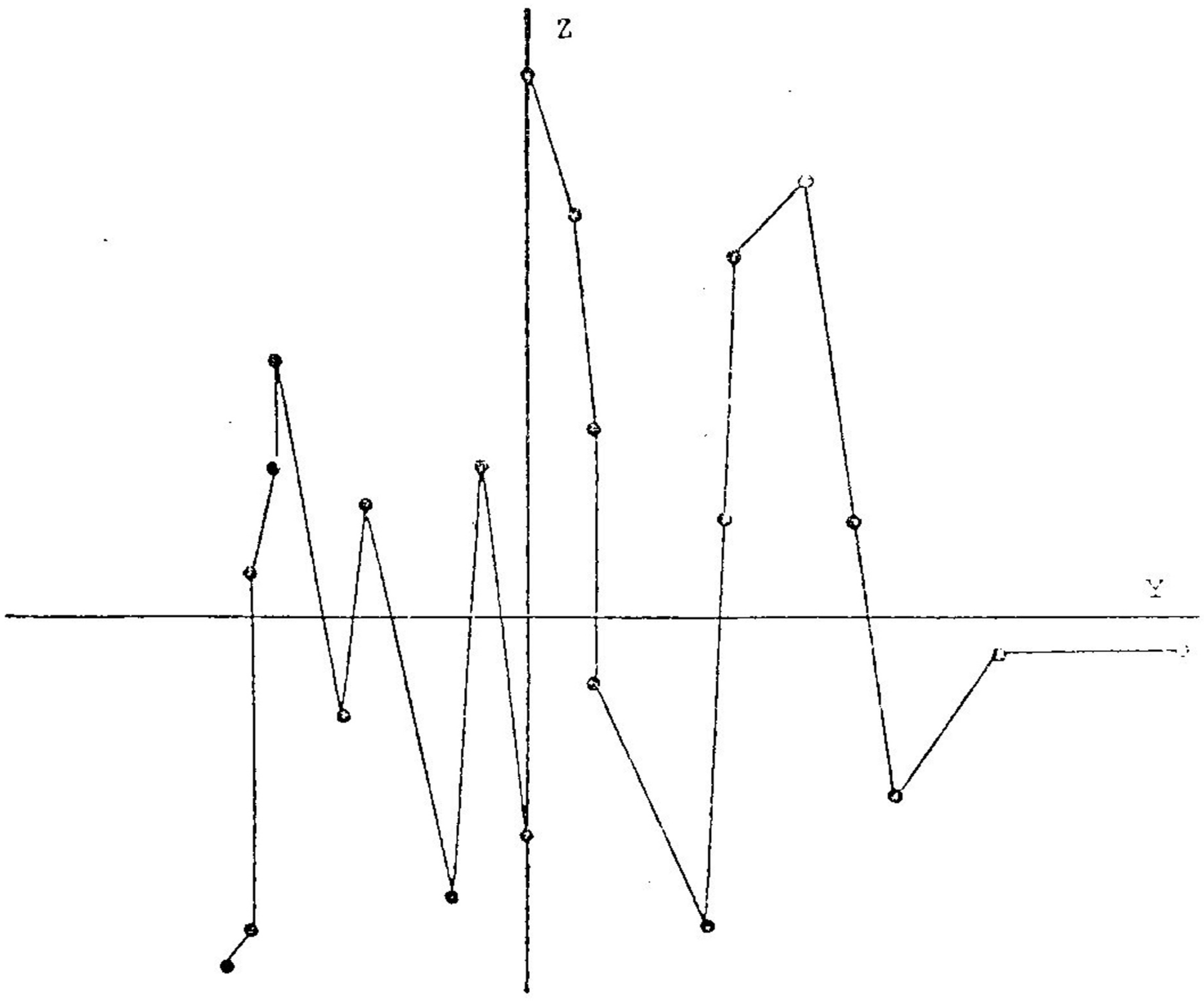
For this 3-way test, the suggestion is to consider a second order, linear regression model:

$$Z = \alpha + \beta Y + \gamma Y2 + \epsilon \qquad (2)$$

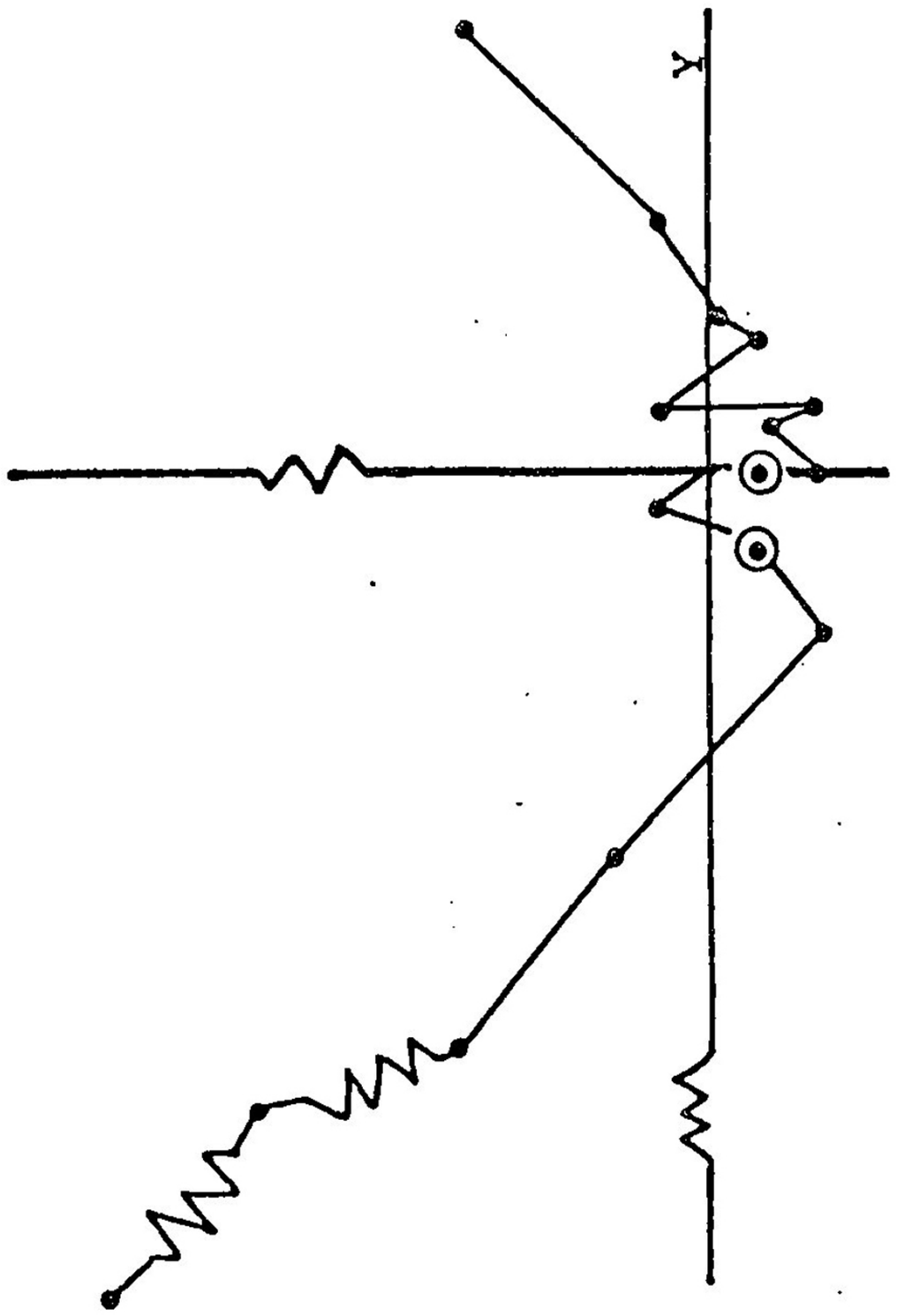
If Y and Z are independent, then the regression coefficients β and are both equal to zero. An F-like ratio, that is, dividing the regression sum of squares due to the individual coefficients by the error sum of squares, is a natural statistic for testing that both coefficients are equal to zero. Since the basic assumptions of normal regression analysis are not met, our ratios cannot be compared to the percentage points of F-distribution. In this



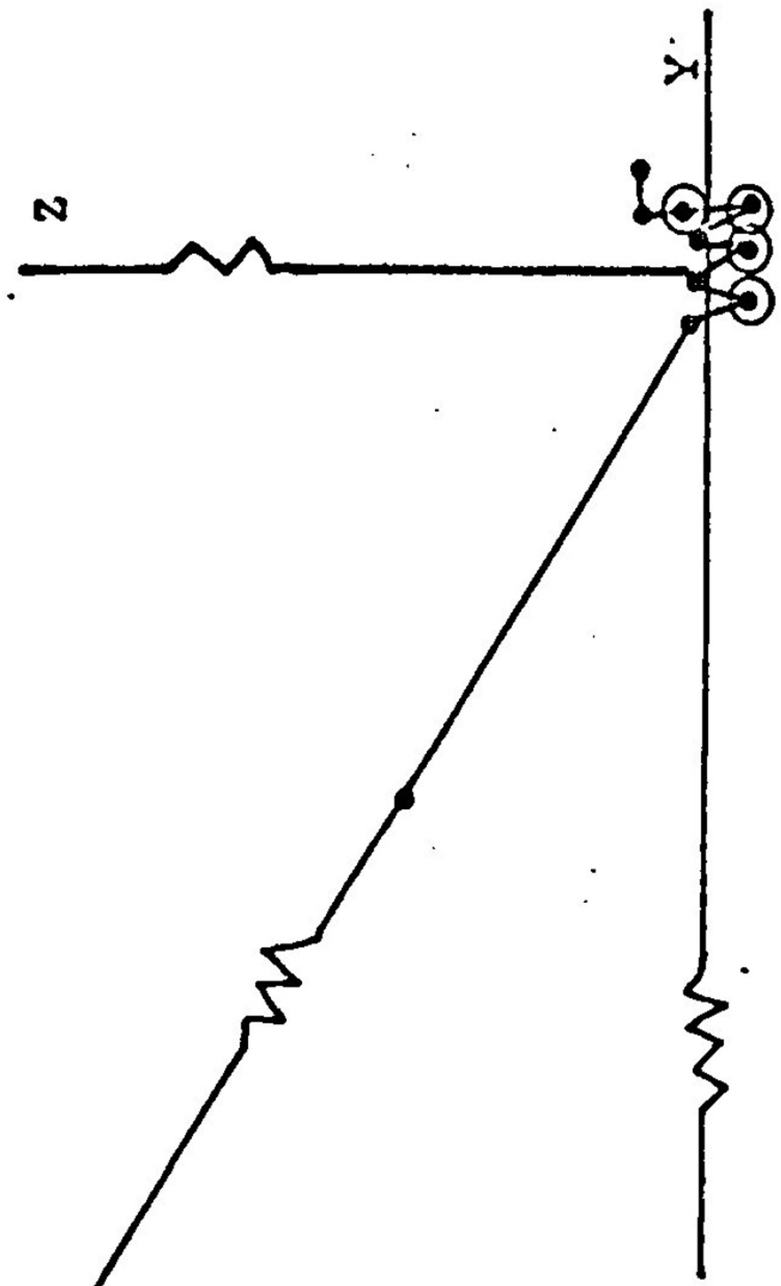
Graph 1: Plot of Y vs Z for the Normal Distribution



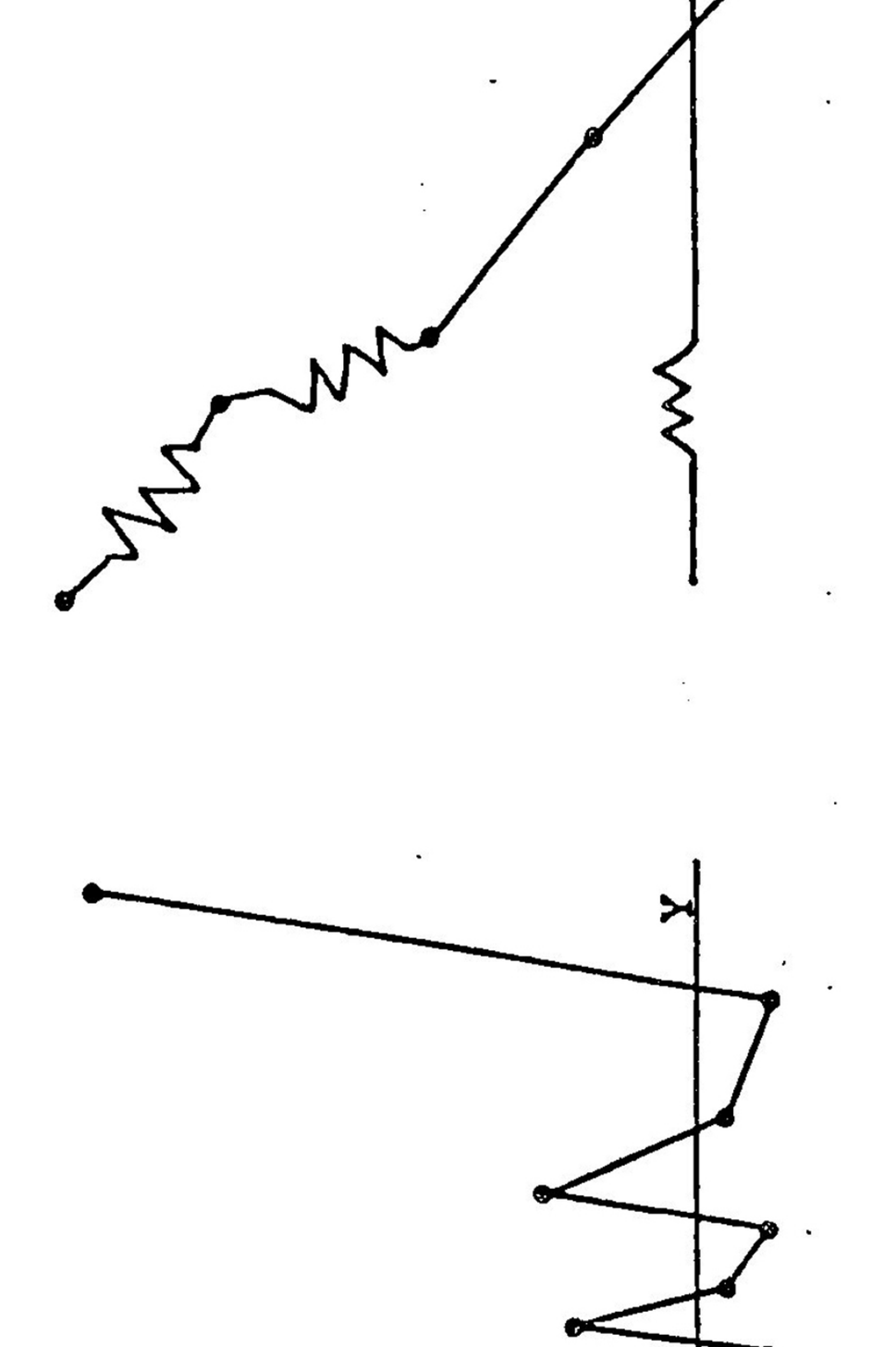
Graph 2: Plot of Y vs Z for the Logistic Distribution



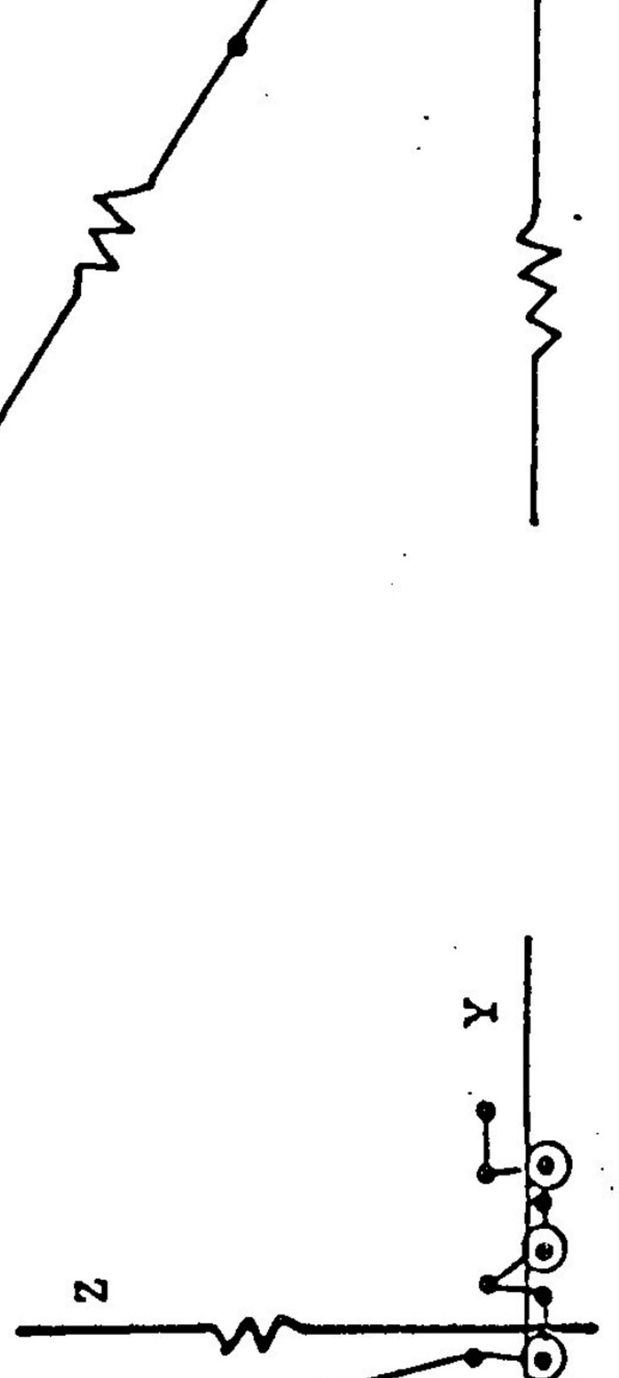
Graph 4: Plot of Y Vs Z for the 10G/4



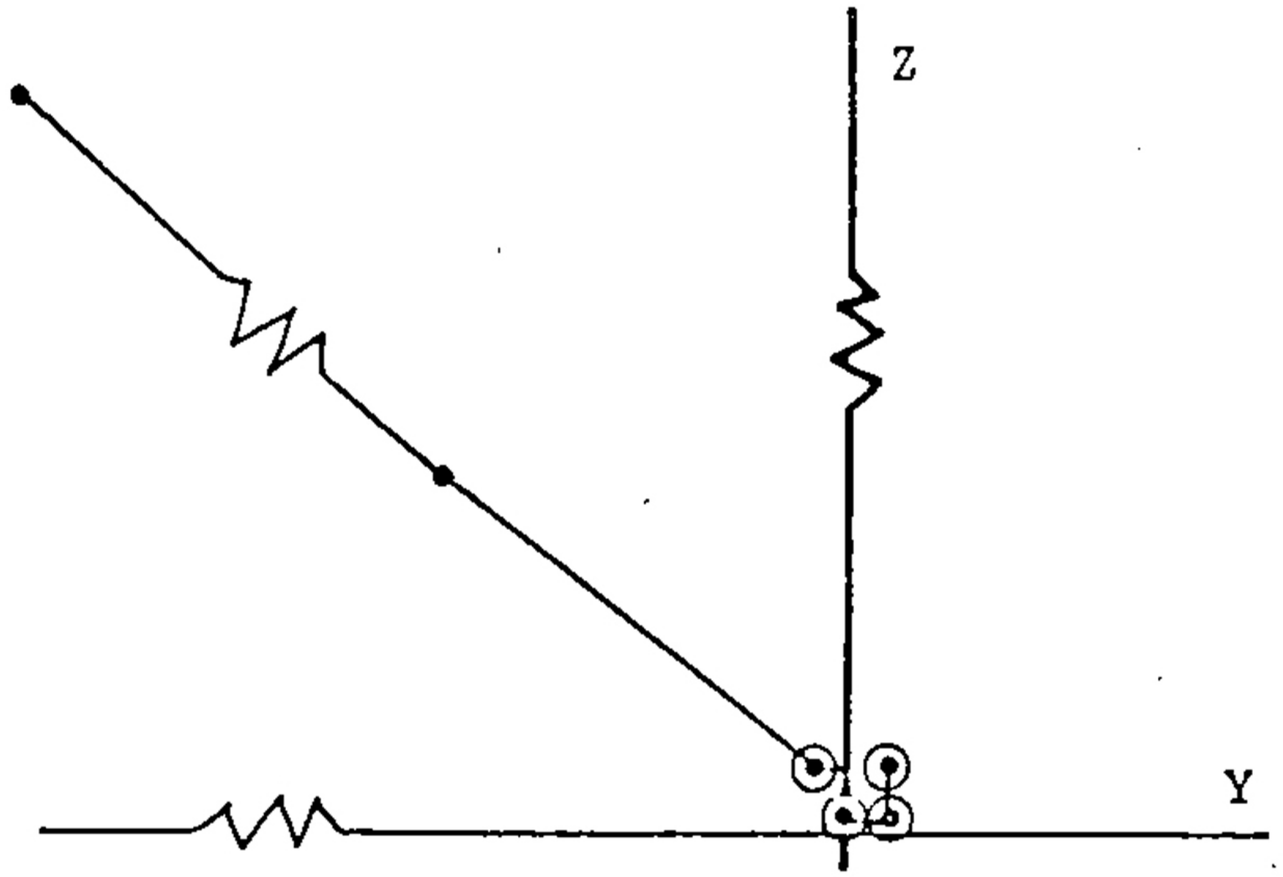
Graph 6: Plot of Y Vs Z for the QS



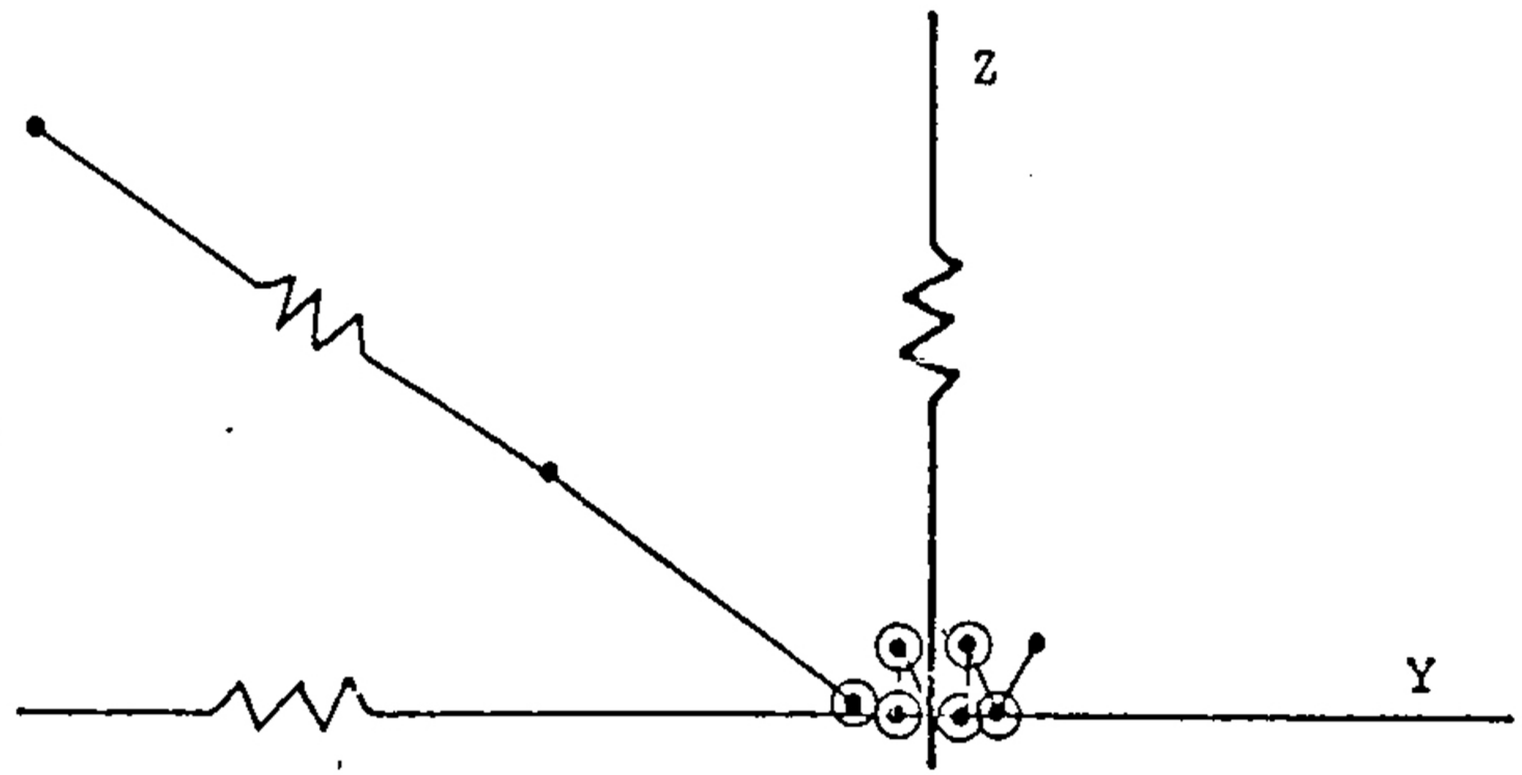
Graph 3: Plot of Y Vs Z for the Q



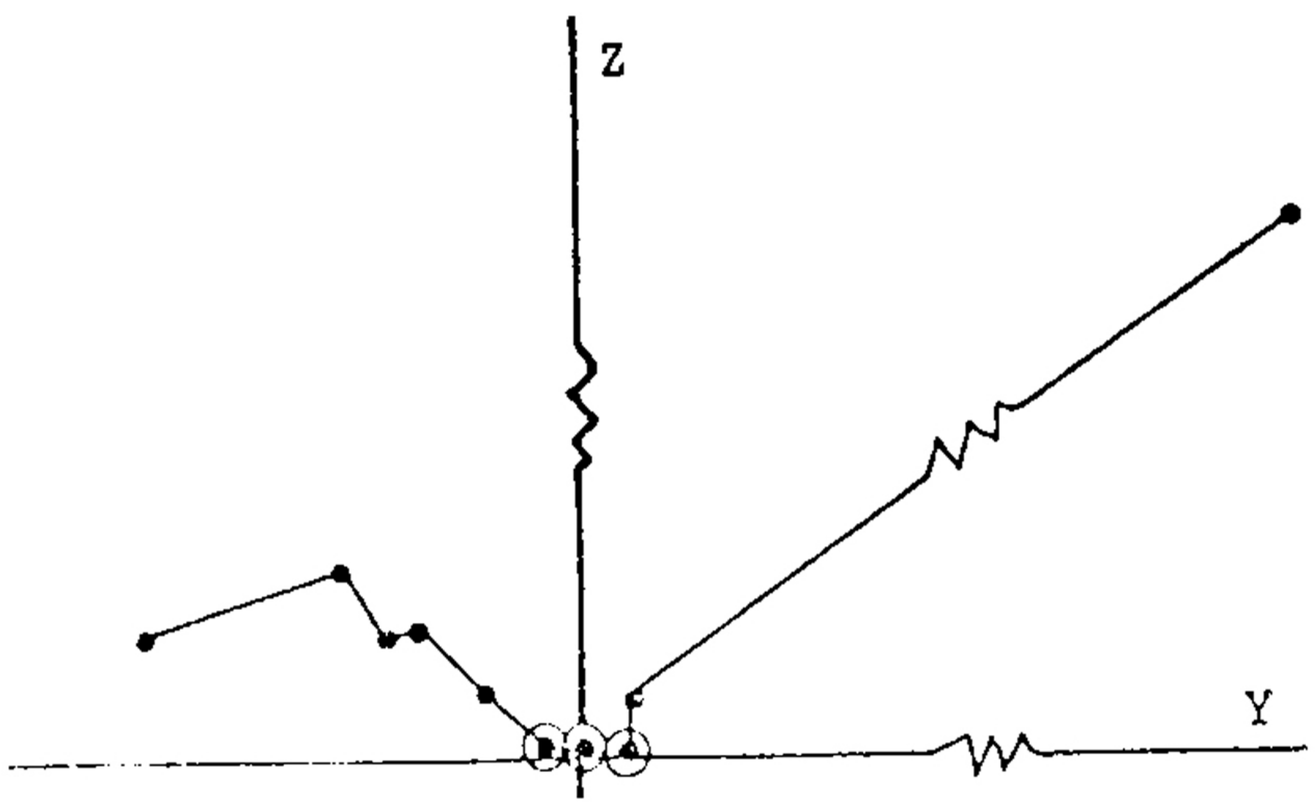
Graph 5: Plot of Y Vs Z for the S/4



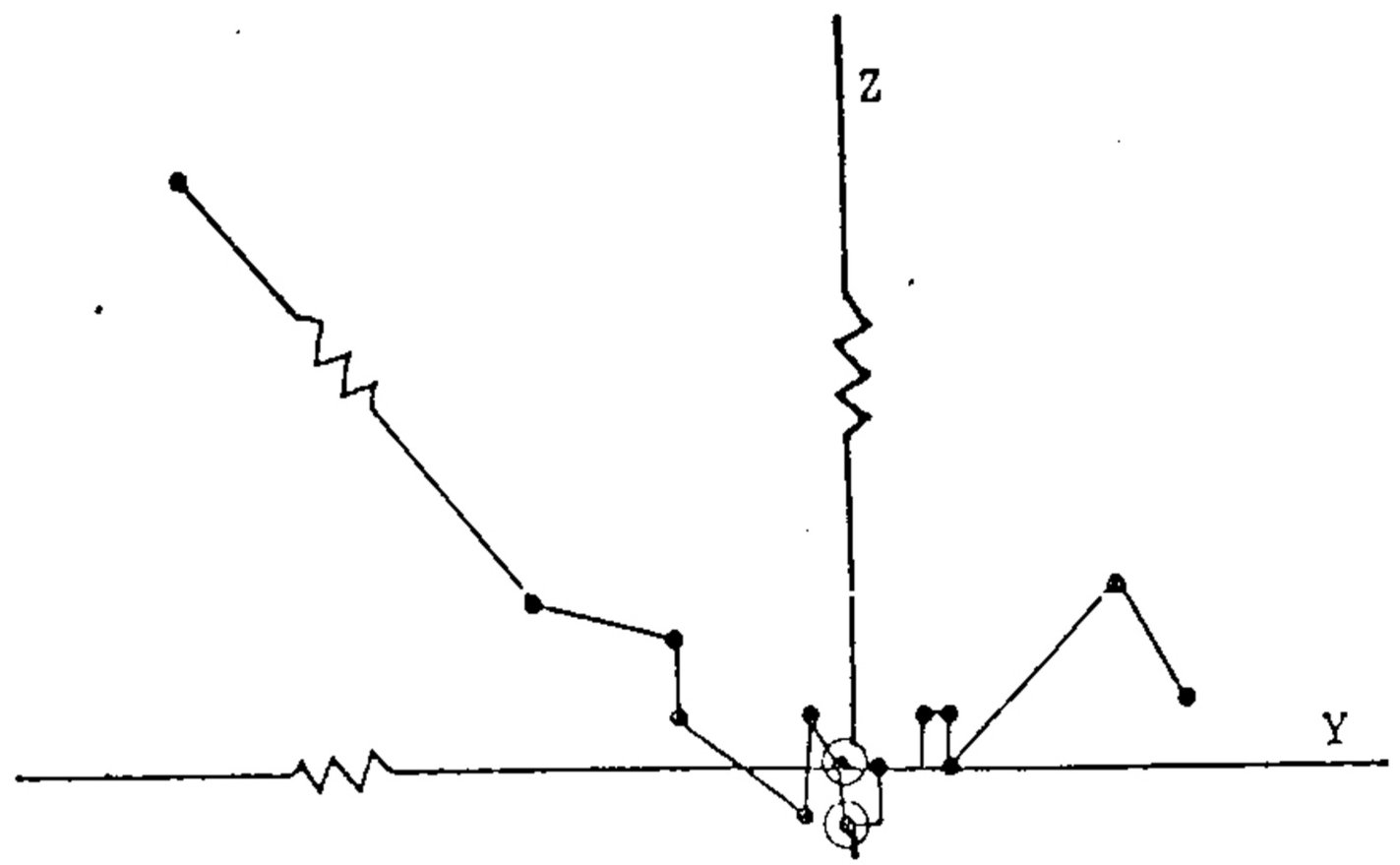
Graph 7: Plot of Y Vs Z for the S



Graph 8: Plot of Y Vs Z for the GUCU



Graph 9: Plot of Y Vs Z for the 3S/4



Graph 10: Plot of Y Vs Z for the Cauchy

section, an empirical study provides regions of acceptance for the various hypothesis yielding power of roughly 80%.

a) Procedure for the 3-Way Test

Given a sample $x_1, x_2, \ldots, x_n$ of size n, we would like to find a two dimensional vector test statistic (R_1, R_2). The basic idea is to rewrite equation (2) with the second degree orthogonal polynomial

$$z = a + b\,\phi_1(y) + c\,\phi_2(y)$$

where

$$\phi_1(y) = y+d \qquad \text{and } \phi_2(y) = y^2+e_1y + e_2$$

are the orthogonal polynomials and a, b, and c are found using the least squares criterion.

A fundamental alternation from the quick test is that the (y,z) pairs are now found using all combinations of data points, not merely the consecutive pairs. This is done to extract maximal information from the sample. For this reason, equation (1) are changed to

$$y_{ij} = x_i + x_j \qquad \text{and} \quad z_{ij} = |x_i - x_j| \qquad (3)$$

where $1 \leq i < j \leq n$.

To find the constants a,b,c,d, e_1, and e_2 requires summations overall the $\binom{n}{2}$ possible (y,z) combinations. Extensive simplification of thse sums can be found when the data is listed in descending order, say $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. A list of simplifications needed to find (R_1, R_2) as given below:

a) $\Sigma y_{ij} = (n-1)\,\Sigma\, x_{(i)}$

b) $\Sigma\, y_{ij}^2 = (n-2)\,\Sigma\, x_{(i)}^2 + (\Sigma\, x_{(i)})^2$

c) $\Sigma\, y_{ij}^3 = (n-4)\,\Sigma\, x_{(i)}^3 + (\Sigma\, x_{(i)}^2)\ (\Sigma x_{(i)})$

d) $\Sigma y_{ij}^4 = (n-8)\ \ x_{(i)}^4 + (\Sigma x_{(i)}^3\ \ (\Sigma x_{(i)} + 3\ (\Sigma x_{(i)}^2)^2$

e) $\Sigma z_{ij} = \Sigma (n-2i+1)\ x_{(i)}$

f) $\Sigma z_{ij}^2 = n \Sigma x_{(i)}^2 - (\Sigma x_{(i)})^2$.

g) $\Sigma z_{ij}\ y_{ij} = \Sigma (n-2i+1)\ x_{(i)}^2$

h) $\Sigma z_{ij}\ y_{ij}^2 = \Sigma (n-2i+1)\ x_{(i)}^3 + \Sigma_{i<j} \Sigma\ x_{(i)}\ x_{(j)}\ (x_{(i)}-x_{(j)})$.

Now, the steps in finding the test statistic (R_1, R_2) are:

1) Take a sample of size n and list its members in decending order.

2) Find d, e_1 and e_2 using the properties of orthogonal polynomials, resulting with

$d = -\Sigma y/nc_2$, $e_1 = \{(\Sigma y^2)(\Sigma y)/nc_2 - \Sigma y^3)/(\Sigma y^2 + d\Sigma y\}$

and $e_2 = d\ e_1 - \Sigma\ y^2/nc_2$, where $nc_2 = \binom{n}{2}$.

3) Find a,b, and c using ordinary least squares.

$a = \Sigma\ z/nc_2$, $b = (\Sigma zy + d\Sigma z)/(\Sigma y^2 + d\Sigma y)$, and

$c = (\Sigma zy^2 + e_1 \Sigma zy + e_2 \Sigma z)/(\Sigma y^4 + e_1 \Sigma y^3 + e_2 \Sigma y^2)$.

4) Compute the sum of squares due to the coefficients and error through the use of the formulae

$SS_a = a\ \Sigma z$, $SS_b = b(\Sigma zy + d\Sigma z)$,

$SS_c = c(\Sigma zy^2 + e_1 \Sigma zy + e_2 \Sigma z$, and

$SS_e = \Sigma z^2 - (SS_a + SS_b + SS_c)$.

5) The test statistic (R_1, R_2) is found by

$R_1 = \Delta(b)\ (SS_b) / SS_c$ and $R_2 = \Delta(c)\ (SS_c)/SS_e$,

where $\Delta(w)$ is defined by

$$\Delta(w) = \begin{cases} -1 & \text{if } w < 0 \\ 0 & \text{if } w = 0 \\ 1 & \text{if } w > 0 \end{cases}$$

The final decision as to which of the three hypotheses to choose is based on empirical, Monte Carlo results presented in the coming section.

b) Invariance of (R_1, R_2)

One last consideration must be noted. To deal with the composite hypothesis of normality with unspecified mean and variance, our statistic should be linearly invariant. To check if it is so, consider the linear transformation of the original data

$x'_{ij} = m(x_i + t/2)$, with $m \neq 0$. Now, let

$y'_{ij} = m(y_{ij} + t)$ and $z_{ij}' = |m|\ z_{ij}$, where

y_{ij} and z_{ij} are defined in equation (3).

Looking at R_1, the ratio of the sums squares, SS_b/SS_e, is known to be linearly invariant due to its similarity to an F statistic. Therefore, it suffices to consider $\Delta(b')$, where b' is b's equivalent when working with the transformed data X'. Throughout the discussion, let the symbol' indicates the equivalent constant or variable when working with the transformed data.

It is easily seen that $d' = m(d-t)$. Now writing

$b = \Sigma z(y+d)/\ \Sigma y\ (y+d)$, we see that

$b' = |m|\ m\ \Sigma z(y+d)/m^2\ \Sigma\ (y+t)(y+d) = |m|\ b/m$,

since $\Sigma(y+d) = 0$

Hence,

$$\Delta(b') = \begin{cases} -\Delta(b) & \text{if } m < 0 \\ \Delta(b) & \text{if } m > 0. \end{cases}$$

So R_1 is only invariant up to its sign. Such dependence only causes problems with skewed distributions. Hnece the skewed alternative must indicate the direction of the taill. Here, it is assumed that the tail was to the right.

Turning to R_2, again it suffices to consider only $\Delta(c')$. After a little algebra, it is seen that

$$e_1' = m(e_1 - 2t) \quad \text{and} \quad e_2' = m^2 (e_2 - e_1 t + t^2).$$

Now, writing

$$c = \Sigma z(y^2+e_1y+e_2) \,/\, \Sigma y^2 (y^2+e_1y+e_2), \text{ we see that}$$

$$c' = |m| \; m^2 \; \Sigma z(y^2+e_1y+e_2) \,/\, m^4 \; \Sigma (y^2+2ty+t^2)(y^2+e_1y+e_2)$$

$$= |m| \; c/m^2,$$

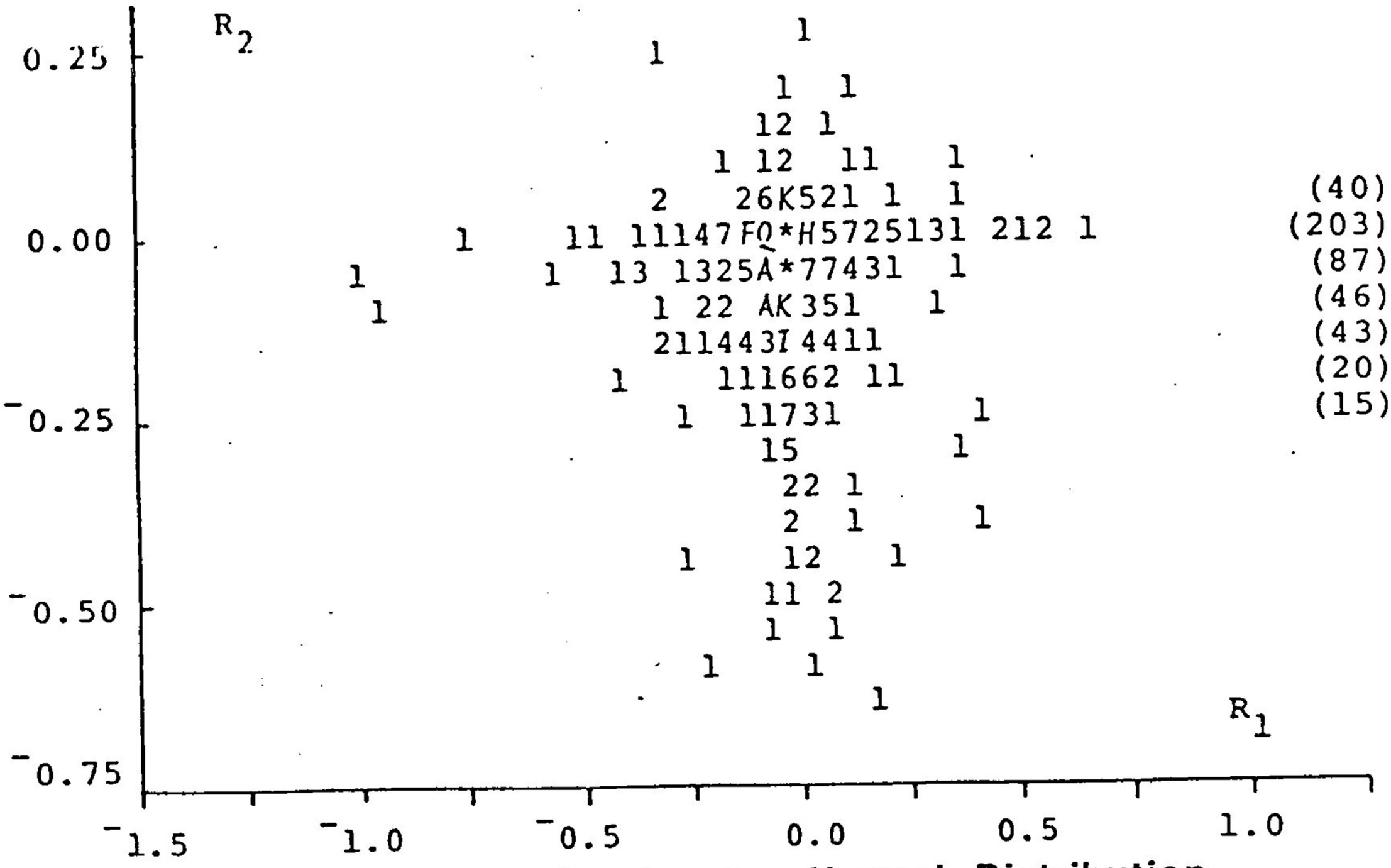
Since $\Sigma(y^2+e_1y+e_2) = \Sigma y(y^2+e_1y+e_2) = 0$. Hence,

$\Delta(c') = \Delta(c)$ and R_2 is found to be linearly invariant.

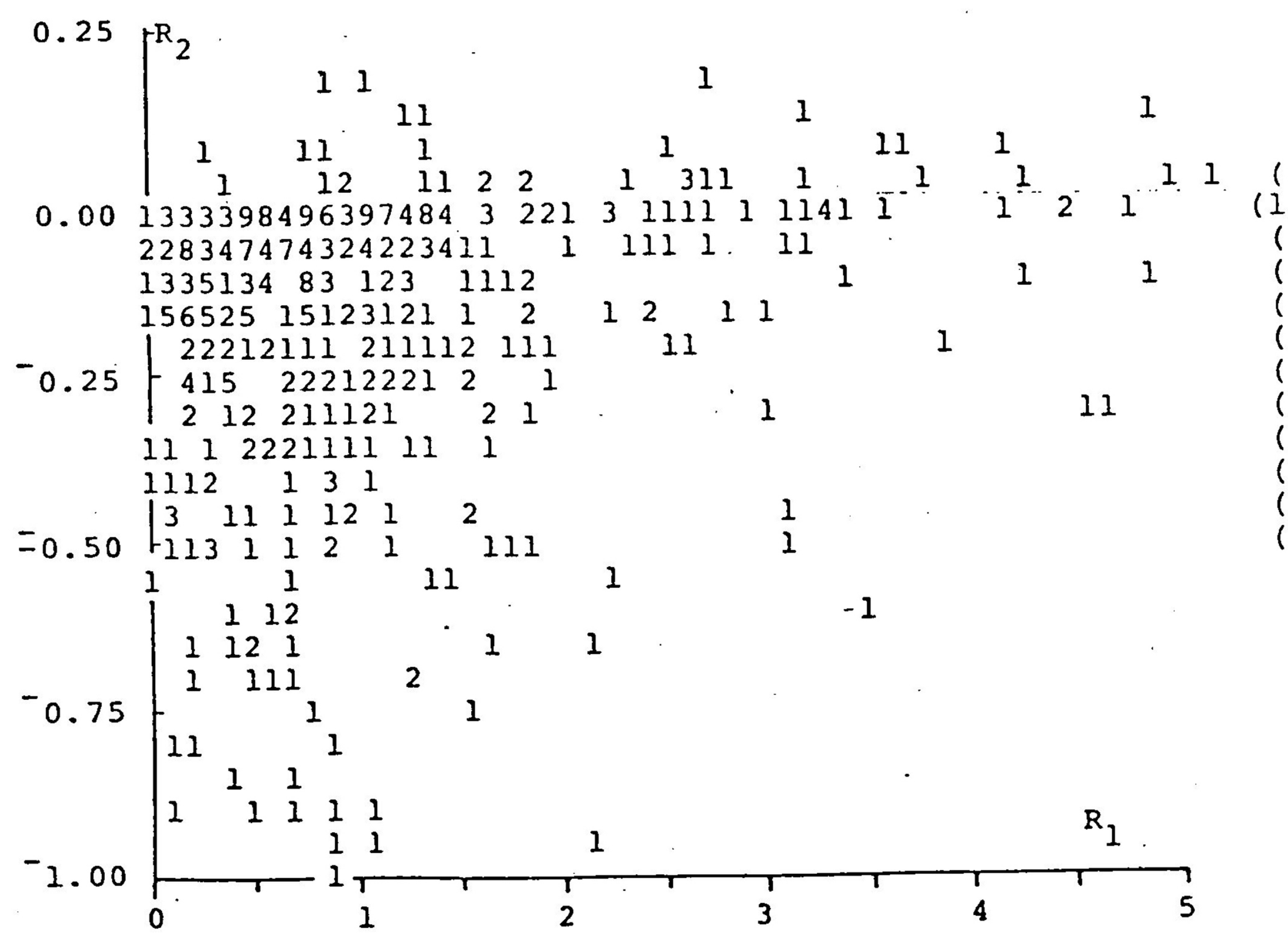
c) Empirical Results Using (R_1, R_2)

To determine the various acceptance regions, a 500 samples of size n=20 were generated from each of the parent distributions normal, chi-square with two degrees of freedom, and Cauchy. For each distribution, the 500 (r_1, r_2) pairs were found and graphed. Graphs 11-13 are the overall graphs for the three distributions. To indicate the number of repeated points at a position, an intricate system was employed. A point which repeated 10-35 times was indicated by the letters A-Z with A indicating 10, B11, and so on up to Z indicating 35 repetitions. Points repeating 36 to 135 times were represented by an asterisk and those repeating more than 135 times by "+". Since "*" and "+" do not specify the exact number of points, the horizontal row sums are provided in parentheses in the right margin for partial guidance.

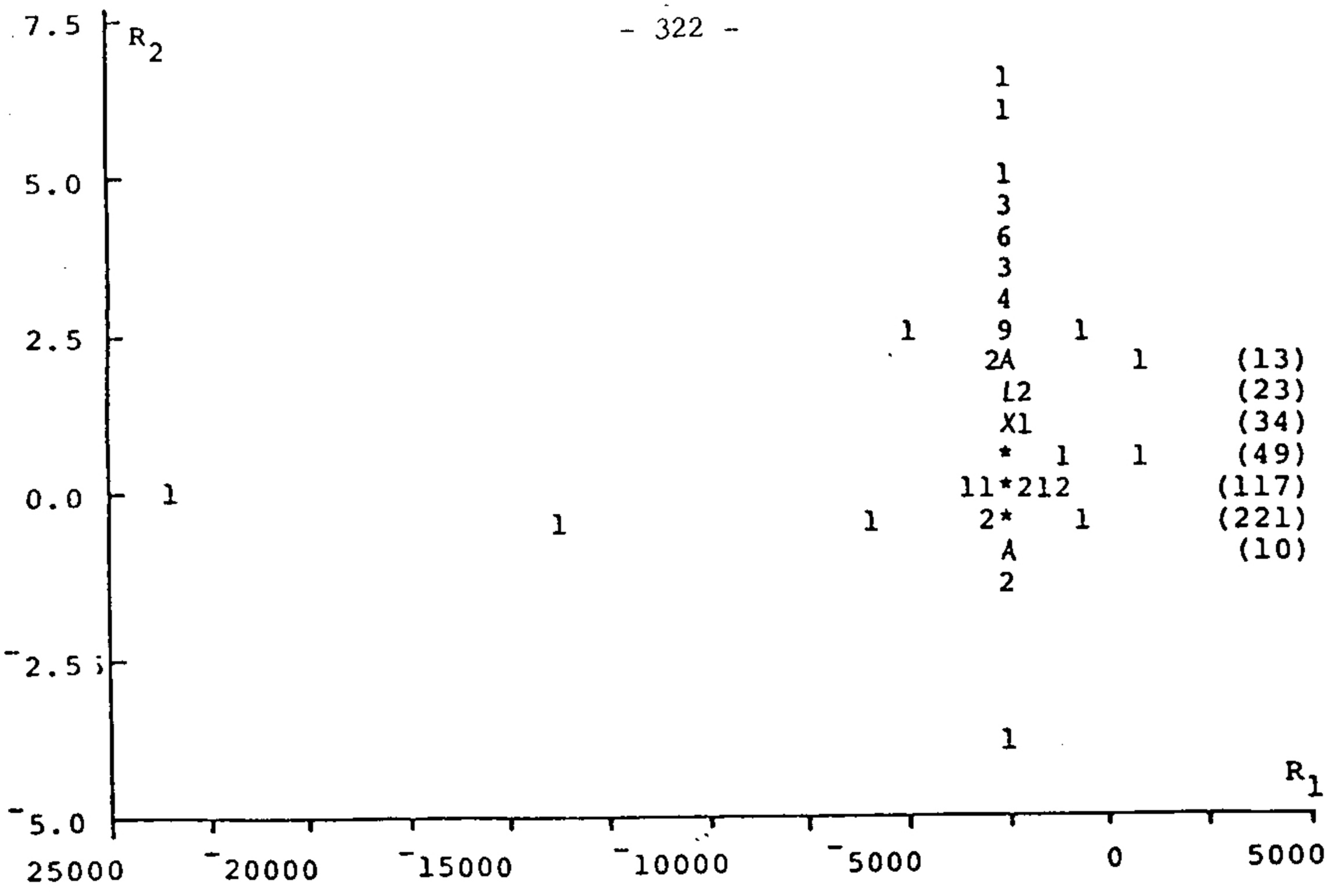
The (R_1, R_2) space was divided into three regions so as to get a power of approximately 80% for each of the three alternatives. To draw these regions, graphs 11, 12, and 13 were redrawn with a wide variety of magnifications of different regions. Graph 14 gives the final acceptance regions for the three alternatives.



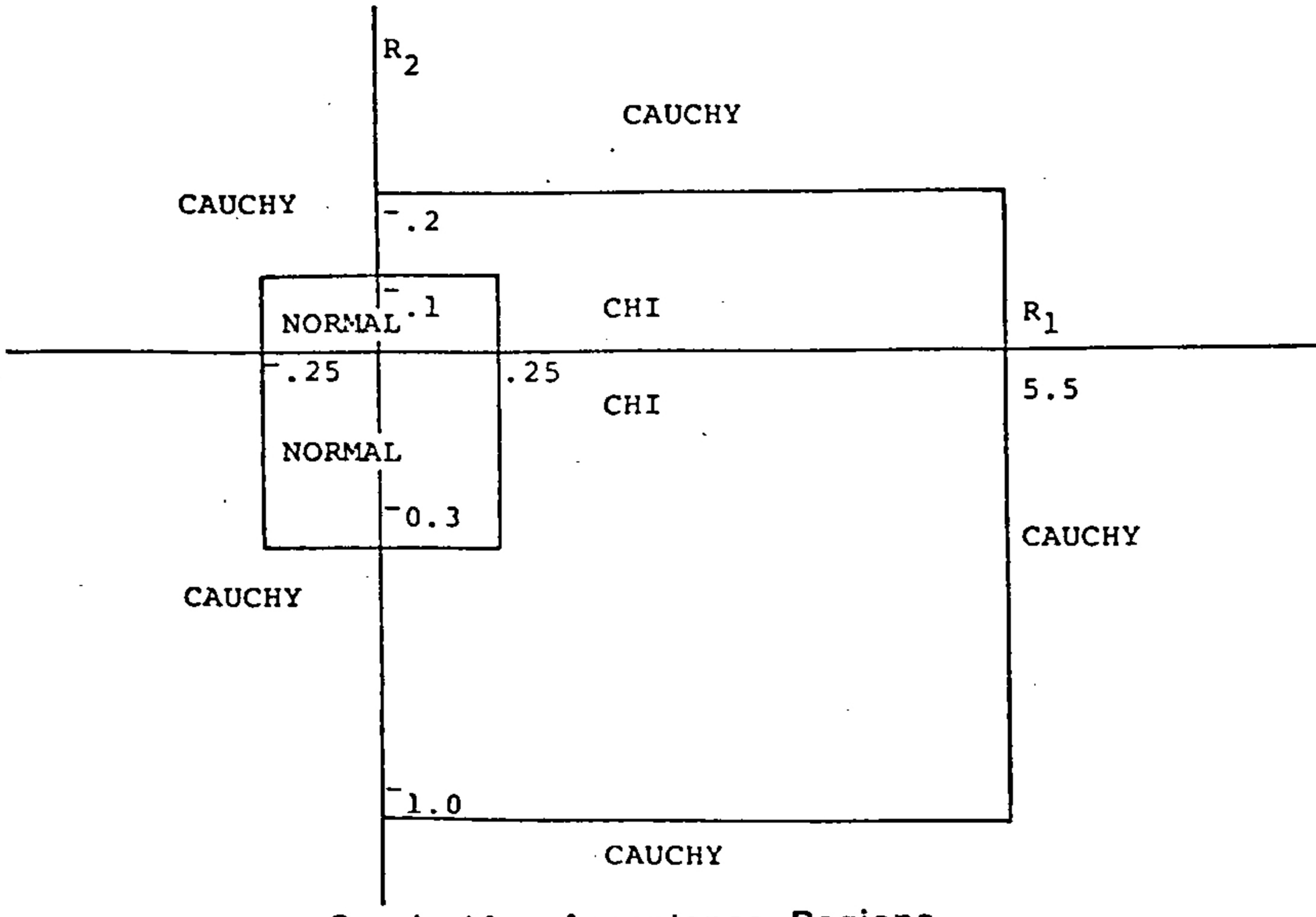
Graph 11: R_1 Vs R_2 for the Normal Distribution



Graph 12: R_1 Vs R_2 for the Chi-Square Distribution with two degrees of freedom



Graph 13: R_1 Vs R_2 for the Couchy Distribution



Graph 14: Acceptance Regions

Next, 10,000 new samples with n=20 were generated from each parent distribution and classified according to graph 14. Table 3 records the results of this power study.

In this table, the strenght of the test statistic is clearly exhibited. For samples with n=20, it indicates that roughly 83% of those samples taken from any normal distribution will be classified as normal, 88% those taken from a Chi-square with 2 degrees of freedom will be correctly classified, and the proportion of correct classifications for samples from the Cauchy distribution is roughly 78%. Hence, (R_1,R_2), and therefore its basis (y,z), is shown to be an effective discriminator of distribution.

Table 3

Monte Carlo Power Study

		Distribution		
		Normal	Chi-Square	Cauchy
Proportion of the time accepted	Normal	0.8344	0.0753	0.0875
	Chi-Square	0.0752	0.8869	0.1237
	Cauchy	0.0904	0.0378	0.7888

4. CONCLUSION

Given a sample from any of a variety of the theoretical distributions, the problem of choosing one for the analysis of the given data has long been a major concern to both the theoretical and applied statisticians.

A two-dimensional statistic, (R_1,R_2) is developed from a second order, linear regression model to classify the parent distribution. Empirical power show that using the normal, Chi-square with two degrees of freedom, and Cauchy distributions as representatives of their respective classes, exhibit the strength

of this statistic. This statistic is shown to be an effective discriminator of distribution.

REFERENCES

[1] Anderson and Darling (1954). A Test of Goodness of Fit. Journal of the American Statistical Association, 49, 765-769.

[2] Dowman and Shenton (1975). Omnibus Test Controus for Departures from Normality Based on √b1 and b2. Biometrika, 62, 243-250.

[3] D'Agostino and Pearson (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b_2 and $\sqrt{b_1}$. Biometrika, 60, 613-622.

[4] D'Agostino and Rosman (1974). The Power of Geary's Test of Normality. Biometrika, 61, 181-184.

[5] David, Hartley, and Pearson (1954). The Distribution of the Ratio, in a Single Normal Sample, of Range to Standard Deviation. Biometrika, 41, 482-493.

[6] Filliben (1975. The Probability Plot Correlation Coefficient Test for Normality. Technometrics, 17,111-117.

[7] Geary (1935). The Ratio of the Mean Deviation to the Standard Deviation as a Test of Normality. Biometrika, 27, 310-332.

[8] Locke(1976). A test for the Composity Hypotnesis that a Population has a Gamma Distribution. Commun. Stat, A5, 351-364.

[9] McDonalad and Katti (1974). Test for Normality using a Characterization of the Normal Distribution. Proceeding of the American Statistical Association.

[10] Ord (1967). On a System of Discrete Distributions. Biometriak, 54, 649-656.

[11] Pearson and Please (1975). Relation Between the Shape of the Population Distribution and the Robustness of Four Simple Test Statistics. Biometrika, 62, 223-241.

[12] Rogers and Tukey (1972). Understanding Some Long Tailed Symmetrical Distributions. Statistica Neerlandica, 26, 211-226.

[13] Shaprio and Frncia (1972). An Approximate Anslysis of Variance Test for Normality. Journal of the American Statistical Association, 67, 215-216.

[14] Shapiro and Wilk (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52, 591-611.

[15] Stephens (1974). EDF Statistics for Goodness of Fit and Some Comparisons. Journal of the American Statistical Association, 69, 730-737.

[16] Weisberg and Bingham (1975). An Approximate Analysis of Variance Test for Non-normality Suitable for Machine Calculation. Technometrics, 17, 133-134.