GOODNESS-OF-FIT TESTING

SOME ASPECTS OF GOODNESS-OF-FIT TESTS IN THE PRESENCE OF UNKNOWN PARAMETERS

By
STEVEN CARL GREENSTEIN, B.S.

A Thesis
Submitted to the School of Graduate Studies
in Partial Fulfilment of the Requirements
for the Degree
Master of Science
McMaster University
December 1974

MASTER OF SCIENCE (1974)
(Applied Mathematics)

MCMASTER UNIVERSITY
Hamilton, Ontario

TITLE: Some Aspects of Goodness-of-Fit Tests in the Presence of Unknown Parameters

AUTHOR: Steven Carl Greenstein, B.S. (City College of New York)

SUPERVISOR: Professor M. A. Stephens

NUMBER OF PAGES: vii, 52.

ABSTRACT

Goodness-of-fit testing of simple hypotheses has a long, well known history. When the Null Hypothesis specifies only the form of the Null Cumulative Distribution Function (CDF), with values for one or more parameters unspecified, the problem is not so clear cut. This project examines several methods of testing fit in the presence of unknown parameters. The methods, briefly described below, are all based on the Empirical Distribution Function (EDF).

(1) The unknown parameters are estimated from the sample. Modified EDF statistics using these sample estimates, are computed, and compared to the significance points which have been obtained by computer simulation.

(2) The unknown parameters are estimated from the sample. Transformations are applied to the observations to obtain transformed variates which, under the Null Hypothesis, are distributed as dependent uniform variates. These transforms are tested for uniformity by the EDF statistics.

(3) A series of transformations is applied to the data to obtain transformed variates which under the Null Hypothesis follow a completely specified Distribution Function; the nuisance parameters have been eliminated. This new, simple hypothesis is then tested by the EDF statistics.

(4) For various parameter values throughout the parameter space, the EDF statistics are computed. A region in the parameter space for acceptance of the corresponding simple hypotheses is determined.

ACKNOWLEDGEMENTS

There are many people to whom I must express my appreciation for their kind assistance. As a matter of course, I must thank my advisor Professor M. A. Stephens for his guidance and technical advice. Beyond this, I would like to thank Michael Stephens for his personal support, and confidence in my ability. I must also thank the many people who performed the thankless task of critically reading and constructively criticizing my work. I must thank them for going out of their way to say something encouraging even when there was little to merit praise. I should like to thank those people who combined to make my studies a pleasurable experience, and me a caffein addict. I would like to thank those people who put up with my insufferable periods, and cheered me out of my down moods. Finally, I would to thank my dog, Gregor, who although she couldn't understand why I didn't want to play, accepted my behaviour.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

THE TEST STATISTICS BASED ON THE EDF

1.0 INTRODUCTION

Suppose the sample $x_1, x_2, \ldots x_n$ has been obtained on the random variable X. We wish to test the Null Hypothesis:

$$X \sim F(X, \underset{\sim}{\theta})$$

Differences between the sample CDF and the EDF, defined below, form the basis of several goodness-of-fit statistics.

definition: Given a sample of size n on the random variable X, the EDF, $S_n(x)$, is defined as that proportion of the sample having a value less than or equal to x.

$$S_n(x) = \sum_{i=1}^{n} s(x_i), \text{ where } s(x_i) = 1/n \text{ for } x_i \leq x, \text{ 0 elsewhere.}$$

For a true null hypothesis, the two functions, the CDF and the EDF should be quite close to each other. How this nearness is evaluated determines the different EDF statistics. For all the EDF statistics, the following is assumed:

To each observation x_i, on the random variable X, the Probability Integral Transformation is applied to give a new observation z_i on a random variable Z. The statistics which follow are defined in terms of $\underset{\sim}{X}$, but could as

well be defined in terms of z. The computing formulas are given in terms of the z_i, which are henceforth assumed to be in ascending order. For each EDF statistic, a subscript is often included to indicate sample sizes: where little risk of confusion arises, this has been omitted.

The Probability Integral Transformation (PIT) may be defined as:

$$Z(x) = \int_{-\infty}^{x} f(x)dx, \text{ where } f(x) \text{ is the probability}$$

density function of the random variable X.

Sometimes the PIT will be used with some parameters estimated from the data. In what follows, estimates of mean μ and variance σ of a population will be given by $\bar{x} = \Sigma x_i/n$ and $S^2 = \Sigma(x_i-\bar{x})^2/(n-1)$, unless otherwise stated.

1.1 The Kolmogorov-Smirnov D Statistic.

Kolmogorov (1933) introduced the following statistic for testing fit:

$$D = \sup_x |S_n(x)-F(x)|$$

The statistic D measures the greatest absolute difference between the CDF and EDF, and has the following properties:

(1) The probability distribution of D depends only on the sample size n, and not on the distribution being tested.

(2) The asymptotic distribution is known

$$\Pr(D < c/\sqrt{n}) = 1-2 \sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2c^2}$$

(3) The pre-asymptotic significance points have been computed by, among others, Birnbaum (1952).

The significance points may be used to construct a confidence bound around the sample CDF, viz $F(x) \mp D_n(\alpha)$ where $D_n(\alpha)$ is the value of the statistic at level of significance α. If, for any value in the sample space, the EDF falls outside this bound, the Null Hypothesis is rejected.

The statistic may be computed from the following form:

$D = \max (D^+, D^-)$ where

$D^+ = \max_i (i/n - z_i)$

$D^- = \max_i (z_i - (i-1)/n)$

1.2 Kuiper's V Statistic

Kuiper (1960) proposed a modification to the D statistic of Kolmogorov, defined as follows:

$$V = \sup_x[S_n(x)-F(x)]-\inf_x[S_n(x)-F(x)\ ]$$

This statistic is found to be origin invariant with respect to the value x_0 at which cumulation begins, and may be used not only for tests of fit on the line, but on the circle. The statistic is found to have the following properties:

(1) It is dependent only on sample size and not on the Null Distribution being tested.

(2) Its asymptotic distribution, and a reasonable approximation to the pre-asymptotic behavior is given by:

$$\Pr(V < c/\sqrt{n}) = 1 - \sum_{j=1}^{\infty} 2(4j^2c^2-1)e^{-2j^2c^2} + \frac{8c}{3\sqrt{n}} \sum_{j=1}^{\infty} j^2(4j^2c^2-3)e^{-2j^2c^2}$$

A table of percentage points for the statistic may be found in Stephens (1965). The statistic V may be described as the absolute sum of largest positive and smallest negative differences between the sample CDF and the EDF, and may be computed from:

$V = D^+ + D^-$, where D^+ and D^- are defined as before.

1.3 The Cramer-Von Mises Statistic W^2

The Cramer-Von Mises test is based on the statistic W^2 defined by:

$$W^2 = n \int_{-\infty}^{\infty} [S_n(x) - F(x)]^2 \, dF(x)$$

The limiting distribution of the W^2 statistic is given by:

$$\lim_{n \to \infty} \Pr[W^2 \leq t] = \prod_{i=1}^{n} \frac{(2it)^2}{\sin[(2it)^{1/2}]}^{1/2}$$

This asymptotic distribution has been tabulated by Anderson and Darling (1952). The pre-asymptotic significance points have been found by Pearson and Stephens (1962) using Monte-Carlo methods for n = 10, and by curve fitting for n = 5,10.

For calculation, the simpler computing form of

$$W^2 = \sum_{i=1}^{n} (z_i - \frac{2i-1}{2n})^2 + 1/(12n)$$

may be obtained. This may be demonstrated by dividing the interval into the n+1 subintervals defined by the sample observations. The integration is then performed, and the results are summed to obtain the value of the integral through the entire sample space.

1.4 The Watson U^2 Statistic

Watson (1961) suggests the following statistic, a modification of the W^2 statistic, for a test of fit:

$$U^2 = n \int_{-\infty}^{\infty} \{S_n(x) - F(x) - \int_{-\infty}^{\infty} S_n(y) - F(y)\, dF(y)\}^2 dF(x)$$

Watson shows the statistic to be identical to the Cramer-Von Mises statistic with respect to that origin x_0 which minimizes W^2, ie $U^2 = \min_{x_0} W^2(x_0)$, where cumulation is initiated at x_0. Beyond the origin invariance of this statistic, which enables it to test for uniformity of direction on the circle, has the favorable property of its distributions rapid convergence to its asymptote, given by computing

$$Pr(U^2 > c) = \sum_{j=1}^{\infty} (-1)^{j-1} 2e^{-2j^2\pi^2 c} .$$

A table of significance points for both the asymptote, and for some pre-asymptotic sample sizes, may be found in Stephens (1963).

By a method of analogous to that of W^2, the following computing form may be obtained:

$$U^2 = \sum_{i=1}^{n} (z_i - \frac{2i-1}{2n} - \bar{z} + 1/2)^2 + 1/(12n)$$

where $\bar{z} = \sum_{i=1}^{n} z_i/n$.

1.5 The Anderson-Darling Statistic A-D

Anderson and Darling (1952) suggested another means of measuring distance between the EDF and sample CDF, namely:

$$A\text{-}D = n \int_{-\infty}^{\infty} [S_n(x)-F(x)]^2 \psi[F(x)]dF(x)$$

where $\psi(t)$, $0<t<1$, is a preassigned weighting function. The statistic is seen to be a modification to the Cramer-Von Mises statistic, reducing to W^2 when $\psi(t) \equiv 1$, $0<t<1$. The authors obtain a computing form as:

$$A\text{-}D = 2 \sum_{i=1}^{n} \{\phi_2(z_i) - \frac{2i-1}{2n} \phi_1(z_i)\} + n \int_0^1 (1-t)^2 \psi(t)dt$$

where $\phi_1(t) = \int_0^t \psi(s)ds$

$$\phi_2(t) = \int_0^t s\cdot\psi(s)ds.$$

Throughout this paper, when reference is made to the A-D statistic, it may be assumed that $\psi(t) = 1/[t(1-t)]$. This weighting function has been chosen so as to provide a heavier weighting to the differences between the CDF and EDF which occur in the tails of the sample space. We note

that the measure of difference is based on $Y(z) = \sqrt{n}\,[S_n(x)-z]$ where z is the value of the sample CDF at $X = x$. The variance of $Y(z)$ is $z(1-z)$, and thus this weighting function scales the weights according to the variability of the difference on which it is based. When this weighting function is used, the computing form reduces to:

$$A\text{-}D = -\frac{1}{n}\sum_{i=1}^{n}(2i-1)\{\ln z_i + \ln(1-z_{n-i+1})\} - n.$$

CHAPTER II

UNIFORMITY AND SUPER-UNIFORMITY: SOME COMPARISONS

2.0 The Probability Integral Transformation when Parameters are Estimated.

When the Null Hypothesis is simple, the PIT transforms the original observations x_i to new observations z_i which are, before ordering, independently and uniformly distributed Uniform (0,1). Let us now consider the situation in which parameters are replaced by their sample estimates. If the PIT is applied, the transformed observations no longer have these distributional properties under the Null Hypothesis. David and Johnson (1947) have shown that for a Distribution Function specified by an estimated scale parameter D(x) and/or an estimated location parameter, m(x) the density of a single transformed observation is of the form:

$$p(z_i) = f(x_i^*)^{-1} \cdot p(x_i^*) \text{ where } x_i^* = \frac{x_i - m(x)}{D(x)}$$

$f(x_i^*)$ is the value of the density of the random variable X at this particular standardized variate, and $p(x_i^*)$ is the density function of the function of the observations x_i^*. Note that in general the values of $f(x_i^*)$ and $p(x_i^*)$ will differ, and thus the distribution of a transformed

observation will be other than uniform. In order to obtain the density as given, the estimators must fulfill the following requirements:

(1) $m(x_1 + a, x_2 + a, \ldots, x_n + a) = m(x_1, x_2, \ldots, x_n) + a$

(2) $m(a, a, \ldots, a) = a$

(3) $D(x_1 + a, x_2 + a, \ldots, x_n + a) = D(x_1, x_2, \ldots, x_n)$

(4) $D(a, a, \ldots, a) = 0$

(5) $D(kx_1, kx_2, \ldots, kx_n) = |k| D(x_1, x_2, \ldots, x_n)$

In figure 1(a), a comparison of the densities of a transformed observation from the Normal Distribution, parameters μ and σ estimated by $\bar{x}$ and S^2, respectively, is presented for sample size n = 7,16 and 25. While for the smaller sized samples there is great difference in shape between this density and a uniform one, these differences lessen as sample size increases. Even for samples as small as n = 25, the density of a z_1 is similar to Uniform throughout the central 95% of the [0,1] interval. The marked differences in shape occur only at the extreme tails of the interval. In figure 1(b), the density of a single transformed observation, from the Exponential Distribution, for sample size n = 7,16, and 25 has been presented. It may be seen that there are qualitive similarities between these two density functions (Figures 1(a) and 1(b)). In both instances, the estimation of parameters leads to a

Figure 1a

The Density Function of a Single Transformed Observation from the Normal Distribution, Parameters Estimated by $\bar{x}$ and s.

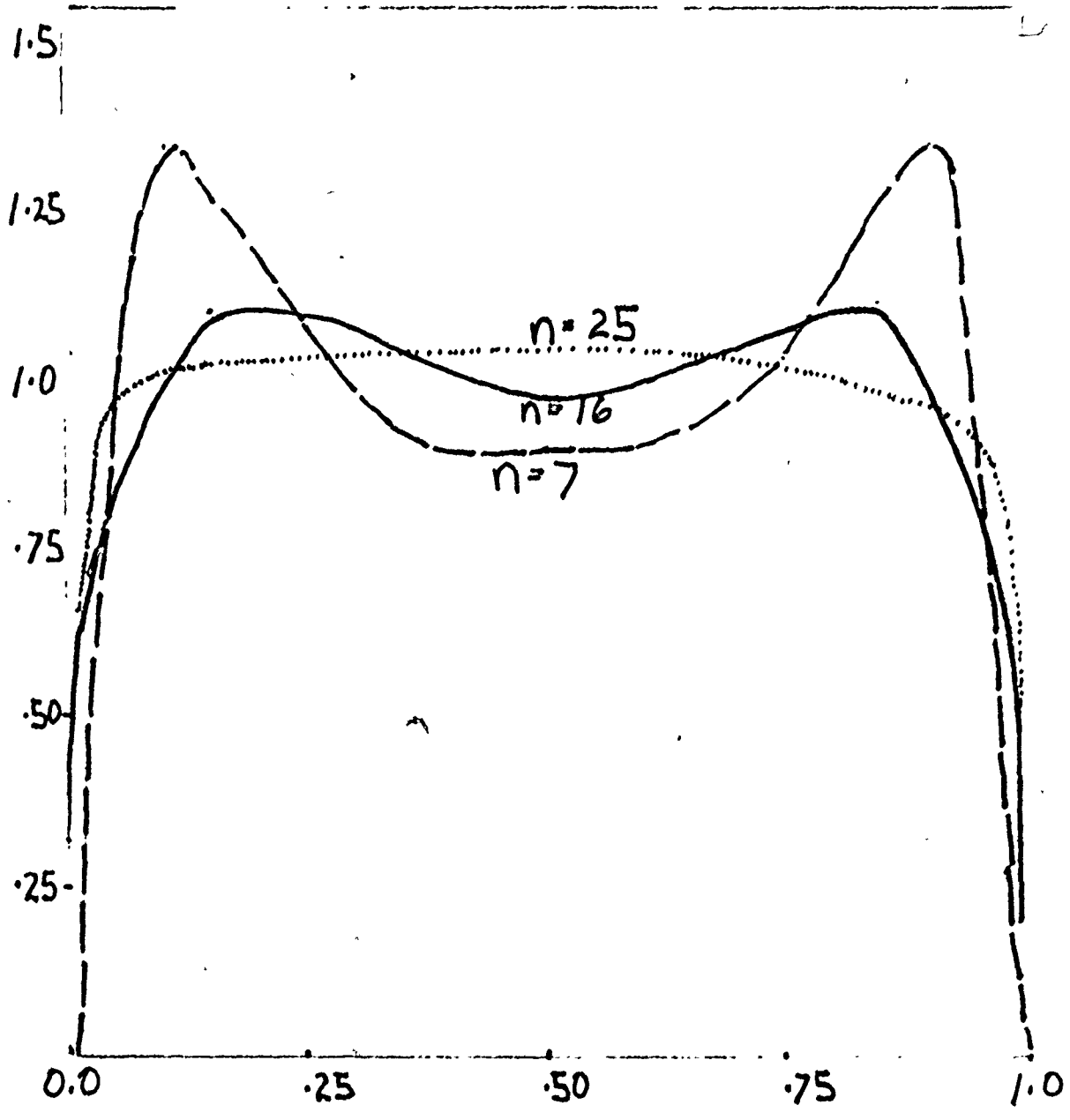
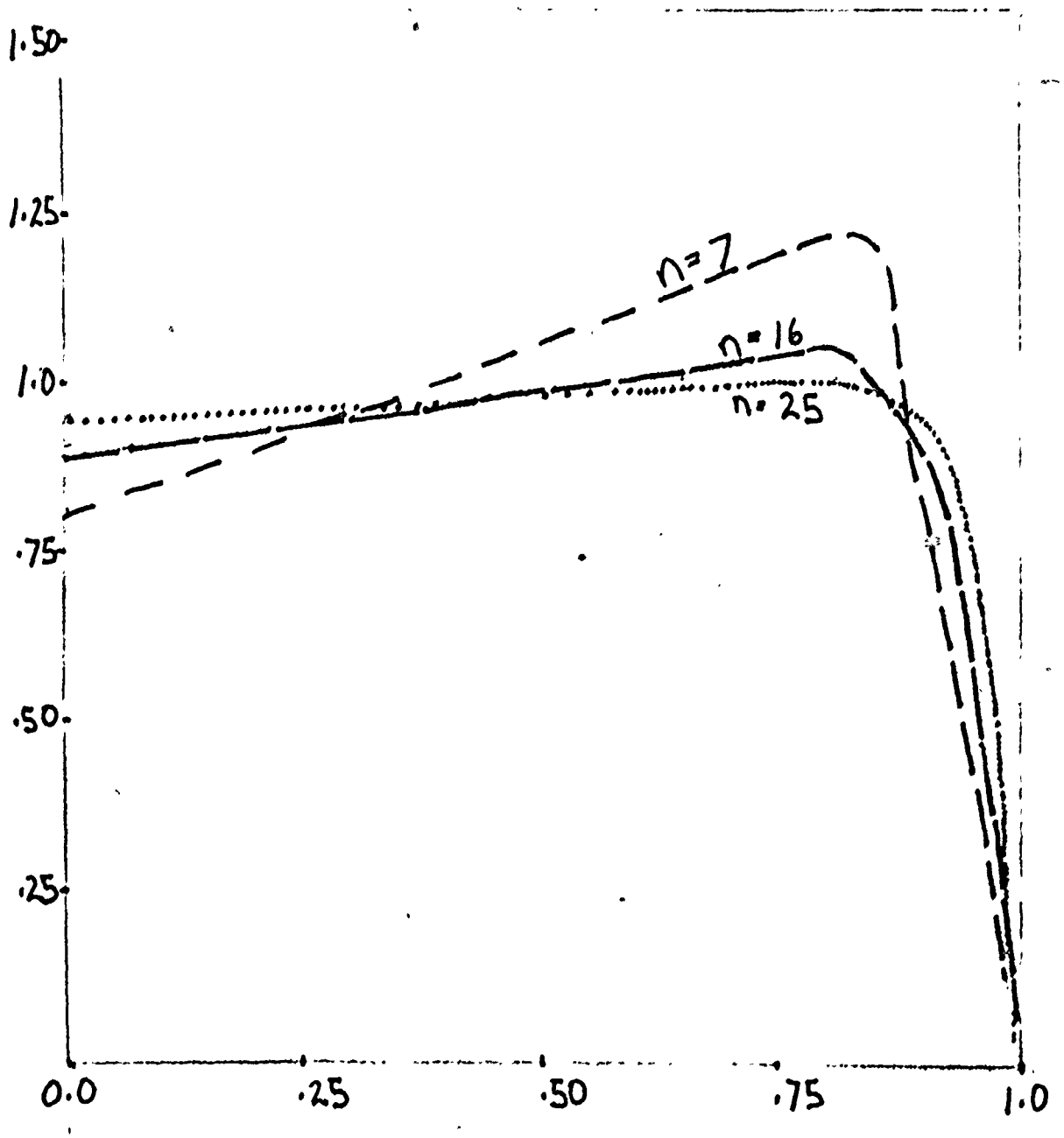


Figure 1b

The Density Function of a Single Transformed Observation from the Exponential Distribution, Population Mean Estimated by Sample Mean



lessening of the probability for the occurrence of a transformed observation well into the tails of the [0,1] interval. I have called the general behavior of samples for which the PIT is applied, with estimated parameters, "super-uniformity". This is meant to indicate the fact that in spite of not being themselves distributed uniformly, the observations tend toward a degree of regularity, of lack of clustering, and of lack of extremes which make samples appear uniformly distributed.

In the notation of Stephens (1974), three situations have been distinguished:

Case 0: The transformed observations are truly independently and identically distributed as Uniform (0,1).

Case 3: The transformed observations are obtained by applying the PIT to observations from the Normal Distribution, the parameters estimated by $\bar{x}$ and S^2, respectively.

Case 4: The transformed observations are obtained by applying the PIT to observations from the Exponential Distribution, the population mean estimated by the sample mean $\bar{x}$.

I will first examine some characteristics of super-uniform samples, and then methods of performing EDF tests in situations of super-uniformity.

2.1 Differences Between Samples of Case 0 and Case 3

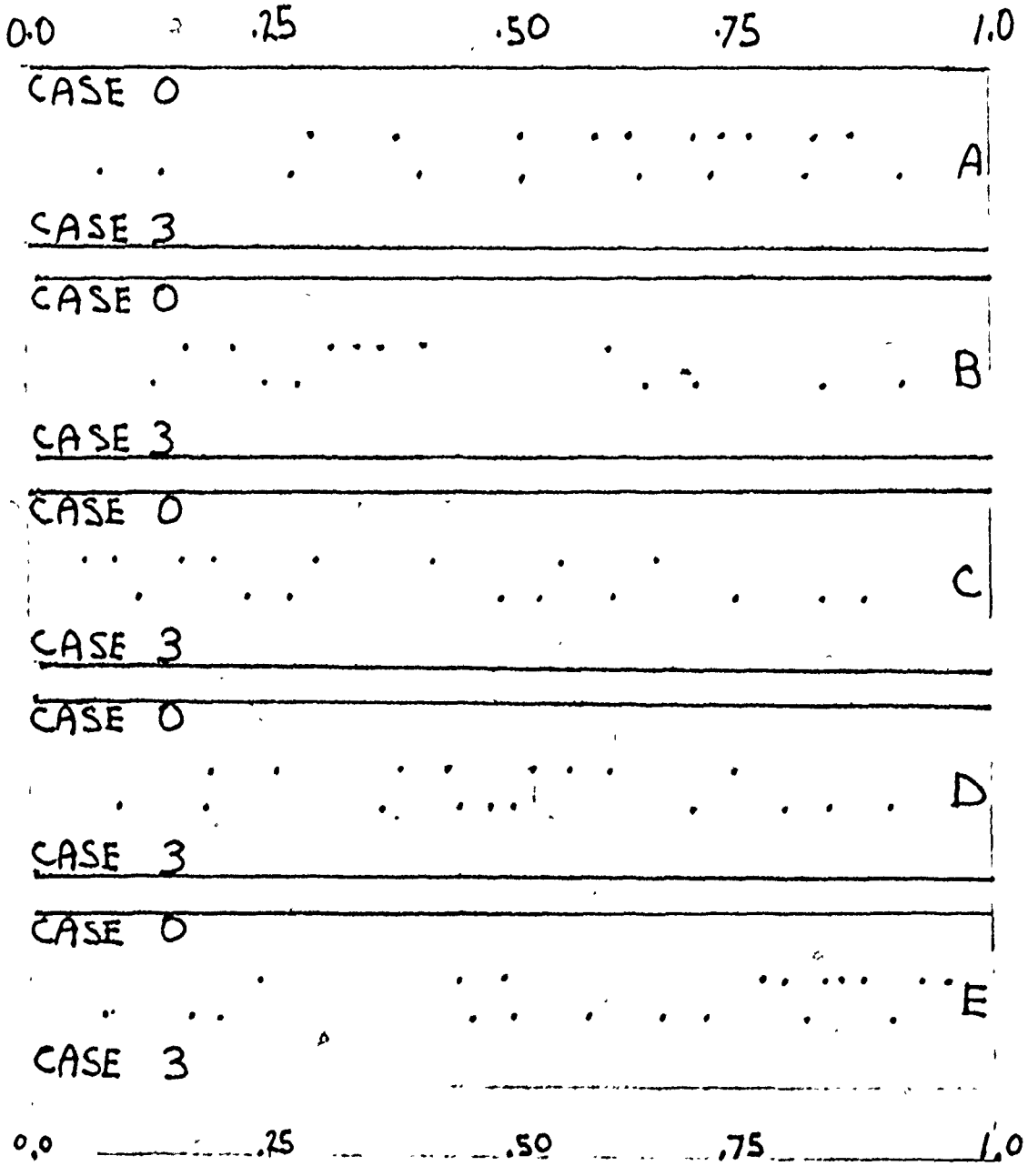
As a means of illustrating the differences in the transformed observations arising as a result of estimation of parameters when applying the PIT, several Case 0 and Case 3 samples have been presented in Figure 2. The samples shown were obtained, in Case 0, by applying the PIT using the known parameters, to observations from the Normal Distribution. In Case 3, these same observations were transformed using the PIT with parameters estimated. The samples depicted have all been chosen for the "non-uniform" appearance of the Case 0 samples.

From Figure 2, it appears that the Case 3 samples appear much more uniform than the truly Uniform samples. The observations, in Case 3, tend to disperse more fully throughout the [0,1] interval, with the observations not clustered together to the same extent as in Case 0. In Case 3, because of the estimation of parameters, the occurrence of observations distant into the tails of the [0,1] interval is lessened. Because in true situations of uniformity, especially for samples of small size, the variations of sampling may produce samples which do not appear as regular as expected, it is striking that super-uniform samples appear so "typically" uniform. They

are as samples from the Uniform Distribution, stripped of the seeming departures from uniformity which ordinarily might occur.

Figure 2

A Comparison of Case 0 Samples and Their Associated Case 3 Samples



2.2 The "Covariance" Between the Ordered Transformed Observations and Their Expected Values Cases 0, 3 and 4.

Because super-uniform observations appear to regularly space themselves throughout the [0,1] interval, without the large gaps, or clusters between observations, the super-uniform ordered observations occur much closer to their expectations than is usual for uniform ordered observations. To demonstrate this tendency, let us examine the "covariance" of the ordered transforms of Cases 0,3, and r and their expected values. We define this covariance by:

$$\text{Cov} = \sum_{i=1}^{n} (z_i - \bar{z}) \cdot (i/(n+1) - 1/2)$$

$$= \sum_{i=1}^{n} z_i \cdot (i/(n+1) - 1/2).$$

In Case 0 situations the first four moments were calculated by theoretical considerations. In order to obtain the significance points for various sample sizes we represent each z_i as the sum of spacings between uniform observations.

If

$$u_1 = z_1$$

$$u_j = z_j - z_{j-1} \quad (j = 2,3,\ldots,n)$$

then

$$z_i = \sum_{j=1}^{i} u_j, \text{ for all } i = 1,2,\ldots,n.$$

The covariance is then represented as a linear combination of these $u_j, j=1,2,\ldots,n$. By a result of Stephens (1972a) we have that the first four moments of the covariance are given by:

$$\mu = \sum_{i=1}^{n} a_i/n+1 \qquad \sigma^2 = \sum_{i=1}^{n} (a_i-\bar{a})^2/\{(n+1)\ (n+2)\}$$

$$\mu_3 = 2 \sum_{i=1}^{n} (a_i-\bar{a})^3/\{(n+1)\ (n+2)\ (n+3)\}$$

$$\mu_4 = [\sum_{i=1}^{n} (a_i-\bar{a})^4 + 3(\Sigma(a_i-\bar{a}_i)^2)^2]/\{(n+1)\ (n+2)\ (n+3)\ (n+4)\}$$

where $a_i = \sum_{j=i}^{n} \left(\frac{j}{n+1} - 1/2 \right)$.

These moments were used to fit Pearson curves to find significance points.

These Case 0 significance points have been verified by Monte Carlo study for several sample sizes, and found to be in agreement to three decimal places. To obtain the significance points for the covariance in Cases 3 and 4 it was necessary to rely totally on Monte Carlo simulations.

In Table 1, significance points of the covariance, for a few sample sizes, have been given. In examining these significance points, the most striking difference noted was the great lessening of spread between the significance points at various percentage levels for the super-uniform samples. In Case 3, for instance, the spread between upper and lower 1% points has dropped to less than one-half of the Case 0 values. With respect to the value of covariance, the super-uniform samples tend to act much more like each other than do samples from the uniform distribution. There is a much smaller incidence of "atypical" samples than with uniform observations. This fact is evidenced by the centralization of the Case 3 and Case 4 covariances.

TABLE 1

A comparison of significance points for covariance: Cases 0,3 and 4

Case 0

n	α .01	.05	.95	.99
10	.391	.481	.861	.918
15	.791	.931	1.317	1.392
20	1.094	1.223	1.788	1.859
30	1.835	1.991	2.661	2.777

Case 3

n	α .01	.05	.95	.99
10	.587	.643	.786	.793
15	.977	1.032	1.222	1.237
20	1.340	1.416	1.658	1.681
30	2.144	2.222	2.534	2.557

Case 4

n	α .01	.05	.95	.99
10	.429	.516	.865	.901
15	.763	.876	1.323	1.368
20	1.133	1.267	1.765	1.833
30	1.922	2.046	2.651	2.773

2.3 Behavior of Q Statistics in Cases 0,3 and 4

Further evidence of what may be described as the "pathological uniformity" displayed by super-uniform samples is demonstrated by the Q statistics, first introduced by Fisher (1932) as a means of both combining independent tests of significance and testing goodness-of-fit. The Q statistics bear some similarity to the EDF statistics in that all use the PIT to obtain the uniformly distributed transformed observations upon which tests are based. They are defined as follows:

Definition $Q = -2 \ln(\prod_{i=1}^{n} z_i)$,

$Q' = -2 \ln(\prod_{i=1}^{n} [1-z_i])$.

It may be readily shown that under a true Null Hypothesis both Q and Q' follow a chi-squared distribution with degrees of freedom equal to 2n. A fuller description of these statistics may be found in Pearson (1938).

The Q statistics have received very little use as goodness-of-fit tests when parameters are estimated from the sample. As mentioned earlier, when parameters must be estimated from the sample the transformed observations are neither uniformly nor independently distributed. Let us examine how this will affect the Q statistics.

Table 2 provides a comparison of the means and variances of the Q statistics for Cases 0, 3 and 4, and several sample sizes, based on a Mohte Carlo study of 500 samples of each size. Notice that for Case 3, all sample sizes, there is a tremendous drop in the variance of the statistics. The mean values are very close to their theoretical (Case 0) values, but the variability drops to about 1% of that for a true chi-square variate. In Case 4 there is still a sizeable lowering of the variance, but this is of nowhere near the same magnitude. Some insight into reasons for this may be gained by examining the Case 3 and Case 4 density functions (figures 1(a) and 1(b)).

It has already been noted (section 2.0) that for both these density functions there is a lessening of probability of the occurrence of an observation in the extremes of the [0,1] interval. It is the nature of the Q statistics that it is a preponderance of either "large" or "small" transforms which accounts for extreme values of the statistic. With samples containing only observations more closely distributed about the midrange, the ln product of these transforms will also attain less extreme values. While the Case 3 density function is symmetric, with a low probability of occurence to transforms at both ends of the [0,1] interval, in Case 4 there

is still sizeable probability of a transform's occurence in the lower tail. This opens up the possibility of a Case 4 sample having transforms with small values, and thus more room for variability of the statistic Q. Notice that the form of the Q' statistic forces, in Case 4, the attainment of the exact theoretical mean for all samples*.

$$*Q' = -2 \ln \prod_{i=1}^{n} (1-z_i) = -2 \sum_{i=1}^{n} \ln (1-z_i)$$

$$= -2 \sum_{i=1}^{n} \ln \left[1-\left(1-\exp\left(\frac{x_i}{\bar{x}}\right)\right)\right]$$

$$= -2 \sum_{i=1}^{n} \frac{-nx_i}{\bar{x}} = 2n$$

TABLE 2

Means and Variances of the Q Statistics:

Cases 0, 3 and 4

Q

n	Case 0		Case 3		Case 4	
	m	v	m	v	m	v
10	20.2402	42.1088	19.4265	.1089	19.1228	21.7195
15	29.8813	63.0159	29.4286	.2217	291320	32.6346
20	39.9190	77.2628	39.4411	.3540	39.2020	48.7158

Q'

n	Case 0		Case 3		Case 4	
	m	v	m	v	m	v
10	20.0417	42.5339	19.4321	.1086	20.0	———
15	29.8033	56.7333	29.4391	.2190	30.0	———
20	39.6855	78.2921	39.4311	.3545	40.0	———

2.4 An Illustrative Example of Super-Uniformity

Super-uniform observations often arise as a direct result of parameter estimation, but may arise in other ways. Let us consider an example by Pearson (1963), based on the years of ascension to the throne by English monarchs between the years 1050 and 1950 A.D. The dates are standardized, and testing is for uniformity on the [0,1] interval. It was not expected that this "time series" would behave as independently distributed variates from the Uniform Distribution. As Pearson says: "If a King reigned for a long time, his son would be old when he in turn succeeded." Because proximate reigns are highly interdependent, the observations violate the assumption of independence. If we apply the goodness-of-fit tests using EDF statistics, the following levels of significance for D, V, W^2 and U^2 are recorded: .03, .025, .07, .034. For each statistic the data yield a value well into the lower tail. For each statistic the nonrandomness of the model is picked out by subnormal variation. When the EDF has drifted in value from the CDF, the correlations between the observations cause a quick return to where it should be. Super-uniform samples may be described as too "good", and these dates have this appearance.

CHAPTER III

METHODS OF TESTING GOODNESS OF FIT IN THE PRESENCE OF UNKNOWN PARAMETERS.

3.0 INTRODUCTION

From the preceding discussions it is hoped clear that the samples that have been labelled super-uniform differ greatly from samples that truly come from the Uniform Distribution. Often we are confronted with the problem of testing goodness-of-fit when the Null Hypothesis specifies only the form of the Null CDF. Because parameters do not have known values we are faced with the situation in which we may actually be testing for super-uniformity. Several methods of circumventing this difficulty in goodness-of-fit testing are presented below.

3.1 Application of Case 0 Statistics in Situations of Super-Uniformity

While the EDF statistics were designed for use in tests of simple hypotheses, they may be modified for use in situations of estimated parameters. To do this, the PIT is applied to the sample observations using sample estimates for the parameters. The question naturally arises as to how this estimation of parameters is to be performed. Consider, for instance, the Kolmogorov-Smirnov statistic D. It has been modified by Lilliefors (1967) to obtain a statistic $\hat{D}$, for use in Case 3 situations. The parameters μ and σ are estimated using maximum likelihood estimates $\bar{x}$ and S_1, where $\bar{x}$ is defined as usual, and $S_1^2 = \Sigma(x_i-\bar{x})^2/n$. The statistic is calculated, as in Case 0, and is compared to significance points obtained by Monte Carlo methods. An alternative modification to the statistic D has been proposed by Srinivasan (1970), namely $\tilde{D}$ which uses Minimum Variance Unbiased Estimation of the parameters in applying the PIT. Results originally published seemed to indicated that the statistic $\tilde{D}$ has favorable power properties over $\hat{D}$. It has more recently been shown by Schafer, Finkelstein and Collins (1972) that this was due to errors in the significance points used. Their results seemed to indicate that both

adaptations to D have essentially the same power properties, with little reason for preferring one method to the other. Similar modifications may be made for each of the EDF statistics.

Table 3 provides a comparison of significance points, found by Monte Carlo simulation, for several sample sizes in Cases 0, 3 and 4. The estimates used in applying the PIT were the unbiased estimate of population variance, ^{2}, and the usual estimator of the mean. For each of the statistics examined, (ie. D, V, W^2, U^2, AD) a substantial drop in the critical points needed for rejection of a Null Hypothesis was noted. The most drastic drop occurred for the Anderson-Darling statistic in Case 3. To explain this substantial decline in critical values, we remember that the A-D statistic gives heavy importance to differences between the EDF and CDF occuring in the tails of the sample space. When parameters are estimated from the sample, the low probability to the occurrence of a transformed observation far into the tails of the [0,1] interval accounts for extremal order statistics being much closer to their expected values. Large differences between the EDF and CDF do not occur here, and hence the statistic seldom achieves values as large as it does in Case 0. For all the statistics a much less pronounced difference between the EDF and CDF is enough to warrant rejection of the Null

Hypothesis. Goodness-of-fit testing is essentially finding a distribution function which fits the sample. When we are able to estimate parameters, we are able to much more closely fit the data to the distributional form we are testing. The closeness of fit shows up in the smaller values of the EDF statistics.

TABLE 3

A Comparison of the Significance Points for the EDF Statistics: Cases 0,3 and 4

$\alpha = .05$

	n	Case 0	Case 3	Case 4
	10	.4094	.2616	.3265
D	20	.2941	.1924	.2358
	30	.2418	.1592	.1944
	10	.5149	.4289	.4911
V	20	.3732	.3164	.3555
	30	.2078	.2623	.2929
	10	.4531	.1200	.2205
W^2	20	.4575	.1229	.2222
	30	.4588	.1259	.2228
	10	.1821	.1105	.1585
U^2	20	.1846	.132	.1597
	30	.1854	.1141	.1601
	10	2.492	.7843	1.2651
A-D	20	2.492	.6919	1.3019
	30	2.492	.7119	1.3147

TABLE 3 (cont.)

$\alpha = .10$

	n	Case 0	Case 3	Case 4
	10	.380	.2394	.2965
D	20	.2651	.1760	.2144
	30	.2179	.1457	.1765
	10	.4794	.3992	.4547
V	20	.3461	.2945	.3288
	30	.2854	.2441	.2708
	10	.3495	.0990	.1742
W^2	20	.3490	.1015	.1756
	30	.3485	.1023	.1761
	10	.1497	.0914	.1280
U^2	20	.1509	.0937	.1290
	30	.1513	.0944	.1293
	10	1.933	.5704	1.0170
A-D.	20	1.933	.5767	1.0466
	30	1.933	.5934	1.0569

3.2 Tests Based on Correlated Uniform Observations

It has already been stated that when the PIT is applied to an observation, using the sample estimates for parameters, the transformed observation follows a known, non-uniform density function (section 2.0). Consider a Case 3 test of normality. In this situation, the density of a single w_i is given by:

$$p(w_i) = \frac{\sqrt{2\pi n}}{n-1} \quad \frac{1}{B(1/2,1/\{2(n-2)\})} \cdot$$

$$\left[1 - \frac{nw_i^2}{(n-1)^2}\right]^{(n-4)/2} e^{-w_i^2/2}$$

where $w_i = \frac{x_i - \bar{x}}{S}$. The proof of this lies in first finding the joint density of $x_i, \bar{x}, S$, ie. $f(x_i, \bar{x}, S)$ and then the conditional density $f(x_i | \bar{x}, S)$. Because the density of a transformed observation may be readily determined, an approach to testing goodness-of-fit in the presence of unknown parameters may be developed based on a further transformation to achieve uniformity. It is not expected that such a transformation would be capable of eliminating the dependencies between transformed observations: thus the testing procedures would still be based on variates not satisfying the conditions required for EDF testing.

Let us examine how such transformations could be performed.

Let $\underset{\sim}{Y} = \underset{\sim}{R}\underset{\sim}{X}$ define a new set of variables

$(y_1, y_2, \ldots, y_n)$ where $\underset{\sim}{R}$ is an orthogonal matrix

$$\underset{\sim}{R} = \begin{vmatrix} c(1-1/n), -c/n, \ldots, -c/n \\ \\ 1/\sqrt{n},\ 1/\sqrt{n}, \ldots, 1/\sqrt{n} \end{vmatrix} \quad \text{with } c = \{n/(n-1)\}^{\frac{1}{2}}.$$

$$\underset{\sim}{Y}'\underset{\sim}{Y} = (\underset{\sim}{R}\underset{\sim}{X})'\,(\underset{\sim}{R}\underset{\sim}{X}) = \underset{\sim}{X}'\underset{\sim}{R}'\underset{\sim}{X} = \underset{\sim}{X}'\underset{\sim}{X}$$

Representing both sides of the above expression as a sum of algebraic terms, we have:

$$(1)\quad y_1^2 + y_2^2 + \cdots + y_n^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

From the definition of the y variates, it follows that y_1^2 and y_n^2 can be represented as:

$$y_1^2 = (x_1 - \bar{x})^2 \frac{n}{n-1}\ ; \quad y_n^2 = \frac{(x_i + x_2 + \cdots + x_n)^2}{(\sqrt{n})^2} = n\bar{x}^2$$

substracting y_n^2 or its equivalent from both sides of (1) we have:

$$(2)\quad y_1^2 + y_2^2 + \cdots + y_{n-1}^2 = x_1^2 + x_2^2 + \cdots + x_n^2 - n\bar{x}^2$$

$$= (n-1)S^2$$

if we subtract y_1^2 from both sides of (2), and then divide by y_1^2 we obtain:

$$\frac{y_2^2+y_3^2+\ldots+y_{n-1}^2}{y_1^2} = \frac{(n-1)^2S^2}{n(x_1-\bar{x})^2} - 1$$

If we divide both sides of the above expression by n-2 we obtain:

$$q_1^2 = \frac{(y_2^2+\ldots+y_{n-1}^2)/n-2}{y_1^2} = \frac{(n-1)^2S^2}{n(n-2)\ (x_1-\bar{x})^2} - \frac{1}{n-2}$$

and note that q_1^2 is the ratio of χ^2_{n-2} variable divided by its degrees of freedom and a χ^2_1 variable; q_1^2 has a distribution which is $F_{n-2,1}$. A similar procedure may be performed for each observation of our sample, to obtain transformed observations which are each distributed as $F_{n-2,1}$. From each of these transformed observations, we may obtain variates which are uniformly distributed by applying the PIT. These uniform variates will not be distributed independently, but it is felt that as sample size increases the correlations among them will decrease, and that for large size samples, tests may be made by Case 0 EDF statistics. For small sample sizes (up to 30), the effect of correlations among transformed observations is quite important. In able 4 a chart of percentages of samples from the Normal Distribution found significant by the EDF statistics is presented. The tests were

performed using the critical points of Case 0, with level of significance $\alpha = .05$ and $.10$. There are wide disparities between the proportions of samples found significant and the proportions which would be found significant in true Case 0 situations. Without a recomputing of critical points for EDF statistics under these transformations, it is impossible to gauge the effectiveness of these transformations. To perform tests using the given percentage points would be an extremely conservative testing procedure.

TABLE 4

Percentages of N(0,1) Samples Found Significant by the Methods of Correlated Uniformly Distributed Observations

$\alpha = .05$

n	D	V	W^2	U^2	A-D
10	.006	.025	.003	.022	.004
15	.009	.028	.003	.026	.004
20	.010	.027	.003	.027	.005
30	.011	.030	.008	.031	.008

$\alpha = .10$

n	D	V	W^2	U^2	A-D
10	.022	.052	.012	.053	.016
15	.025	.053	.015	.052	.019
20	.023	.056	.014	.057	.016
30	.034	.058	.022	.059	.026

3.3 Exact Tests in the Presence of Unknown Parameters

Another method of testing goodness-of-fit for composite hypotheses has been proposed by Csorgo, Seshadri, and Yalovsky (1973). The method, based on characterizations of distributions, involves the use of transformations to the data so as to produce transforms which follow a completely specified distribution function if and only if the observations themselves come from the Null CDF. This method avoids the problem of unknown parameters by replacing the test of a composite hypothesis with that of a simple one. This method does, however, have several undesirable properties which should become apparent in discussion of the method with respect to a test of normality. References to proofs of theorems underlying the method may be found in the paper by Csorgo et al.

Suppose a sample of size n has been obtained on a random variable X with mean μ and variance σ^2. We define:

$$v_i = (x_1+x_2+\cdots+x_i-i\cdot x_{i+1})/\sqrt{i(i+1)},\ i=1,2,\ldots,n-1$$

$$y_k = k\, v_{k+1}\left(v_1^2+v_2^2+\cdots+v_k^2\right)^{\frac{1}{2}},\ k = 1,2,\ldots,n-2$$

Then the sample observations $x_1,x_2,\ldots,x_n$ are distributed $N(\mu,\sigma)$ if and only if $y_1,y_2,\ldots,y_{n-2}$ are distributed according to student's t distribution with degrees of

freedom equal to 1, 2,...,n-2, respectively. Since we have, under the Null Hypothesis, variates from a completely specified distribution, our test of fit may be based on these transformed variates. Any of the methods for testing simple hypotheses may be applied, such as the EDF test statistics.

While from a theoretical stand point there is little that is objectionable to basing goodness-of-fit tests on the above transformations (the transformed observations are not identically distributed) from the practical point of view, it is an extremely poor method. It is found to be extremely insensitive at detecting departures from normality, and is found to be extremely sensitive to the order in which the sample observations have occurred, different orderings of the same data yeilding markedly different results. An example may serve to demonstrate this.

Consider the following 15 observations to be tested for normality, mean and variance unspecified: 1.3, 2.4, 3.4, 4, 4.5, 5.5, 6.1, 7.2, 8.3, 9.4, 10.7, 11.8, 98.2, 98.9, 99.9. This sample has been constructed to bear no resemblance to a sample truly obtained from the Normal Distribution. When this sample is tested for normality by the case 3 EDF statistics at an α level of 5% the value of each test statistic fell in the extreme upper tail. Testing the

sample, as it stands, for normality by the Csorgo transformations, the Null Hypothesis was once again rejected. Yet when these transformations were applied to the data in different, random orderings of the sample, the following percentages of acceptance for the Null Hypothesis were recorded: D = 34%, V = 14%, W^2 = 55%, U^2 = 19%, A-D = 44%. The advantage gained by the use of tests of simple hypotheses seems small in contrast to the loss of reliability. Instead of sharpening the information contained by the data, the use of these transformations seems to dull it and "wash it out".

3.4 Goodness-of-fit testing without estimation of parameters

While there are many goodness-of-fit methods, none is totally effective at determining some large departures from the Null Hypothesis. Perhaps this stems from the fact that samples of small size from many different distributions may all appear alike. Goodness-of-fit tests are only able to tell us that it would have been improbable to obtain the sample actually obtained, under our Null CDF. The tests can not tell us that the sample actually was drawn from the Null CDF. In real situations, samples never are from a particular distribution, and so we are more concerned with being able to act as if they were. In light of this, it may be wise to adopt an extremely pragmatic approach to goodness-of-fit testing. We shall seek a Null Hypothesis which is in agreement with the sample data. Instead of accepting or rejecting the Null Hypothesis, when parameters are estimated, based on a test criterion evaluated at these parameter values, let us seek whether, under the Null Distribution Function, there are parameter values for which this agreement does exist. For these values, there is no basis for disbelieving the Null Hypothesis, and we may act as if the Null Hypothesis is

true.

Since this method entails the determination of a region of acceptance for the Null Hypothesis, goodness-of-fit testing must be carried out throughout the sample space. A single composite hypothesis is replaced by myriad tests of simple hypotheses, and the question arises as to which test criterion to use. Should the acceptance region be based on a single criterion, or should a joint acceptance region, based on several test statistics be computed? Stephens (1974) has shown that among the EDF statistics there are fairly high correlations; that is, samples rejected by one EDF statistic are often rejected by the others. This would seem to indicate that, overall, little difference would be made by requiring a joint acceptance region.

Because our acceptance region will point out the parameter values at which our Null Hypothesis is tenable (or perhaps not untenable), the question of size of an acceptance region should not arise. Should our acceptance region consist of but a single point, there is no reason to find our Null Hypothesis false at this particular point. It is here that the problem of super-uniformity once again comes into play. Since the smallest values of EDF test statistics are associated with parameter values near the sample estimates, the acceptance region might consist only

of points associated with the situation of super-uniformity. Although in cases of super-uniformity the acceptance of a Null Hypothesis requires much smaller values of the EDF test statistics than in Case 0, it is felt that the test should be conducted as usual. We merely seek whether there is contradiction, for a set of postulated values of parameters, between the Null Hypothesis and the sample data.

CHAPTER IV

AN ILLUSTRATIVE EXAMPLE OF THE GOODNESS-OF-FIT METHODS

To indicate how the various goodness-of-fit methods perform in an actual testing situation, I will use an example common to the literature, from Snedecor (1946). A test of normality is performed on the following sample of size 11 of the weight in pounds of men: 148, 154, 169, 161, 162, 166, 170, 182, 195, 236. The test is of the composite hypothesis of normality. As an indication of the truth of the Null Hypothesis, the Shapiro-Wilk statistic was computed, and found to have a value of 0.79, which is at just below the 1% level of significance. On the basis of this test, it is felt that the Null Hypothesis is false, the sample was not obtained from the Normal Distribution.

If we calculate the sample estimates of the mean and variance, we may perform the EDF tests, Case 3. When this is done, the following values for the statistics are obtained: $D = .259$, $V = .427$, $W^2 = .164$, $U^2 = .143$, $A\text{-}D = .974$. These values correspond to roughly the following α-values, respectively: .035, .035, .01, .035, .01. On the basis of the EDF tests, Case 3, once again the Null Hypothesis is found to be false. Notice too, the close

agreement between the level of significance of W^2 and A-D to that of the Shapiro-Wilk statistic. This may be taken as partial indication of their similar power properties.

The procedure of transforming observations to obtain correlated uniform transforms was also conducted. The following values of the EDF statistics were recorded: $D = .248$, $V = .369$, $W^2 = .159$, $U^2 = .112$, $A\text{-}D = 1.024$. These correspond to α levels of approximately, .07, .12, .03, .05, .01 using Case 3 critical points. Because in a sample of so small a size, the effects of correlations among the transformed observations are felt to be important, and because there was no means of adequately evaluating the significance points of the test statistics when this procedure has been used, it is difficult to analyze what these values mean. It was noticed that the significance points for each of the EDF statistics, under the method of transformations to correlated uniform observations, fell between those of Cases 0 and 3, but closer to those of Case 3. With this as a guide, it is felt that the Null Hypothesis would be rejected at the 10% level of significance, though possibly not at the 5% level. It is not felt that this method will be particularly good at detecting departures from the Null Hypothesis.

When the data were analyzed using the transformations suggested by Csorgo, et al, results were obtained which once again point out the tremendous effect that ordering of the data plays. For a test based on the data in the order given, all EDF statistics occurred in the extreme upper tail. While this might seem to be in accordance with the results of other methods, this is misleading. What the method so readily picked out is the lack of independence of the observations. When other, random orderings of the observations were tested by this method The Null Hypothesis was most often accepted (the A-D statistic accepts the Null Hypothesis of normality 17 out of 20 times).

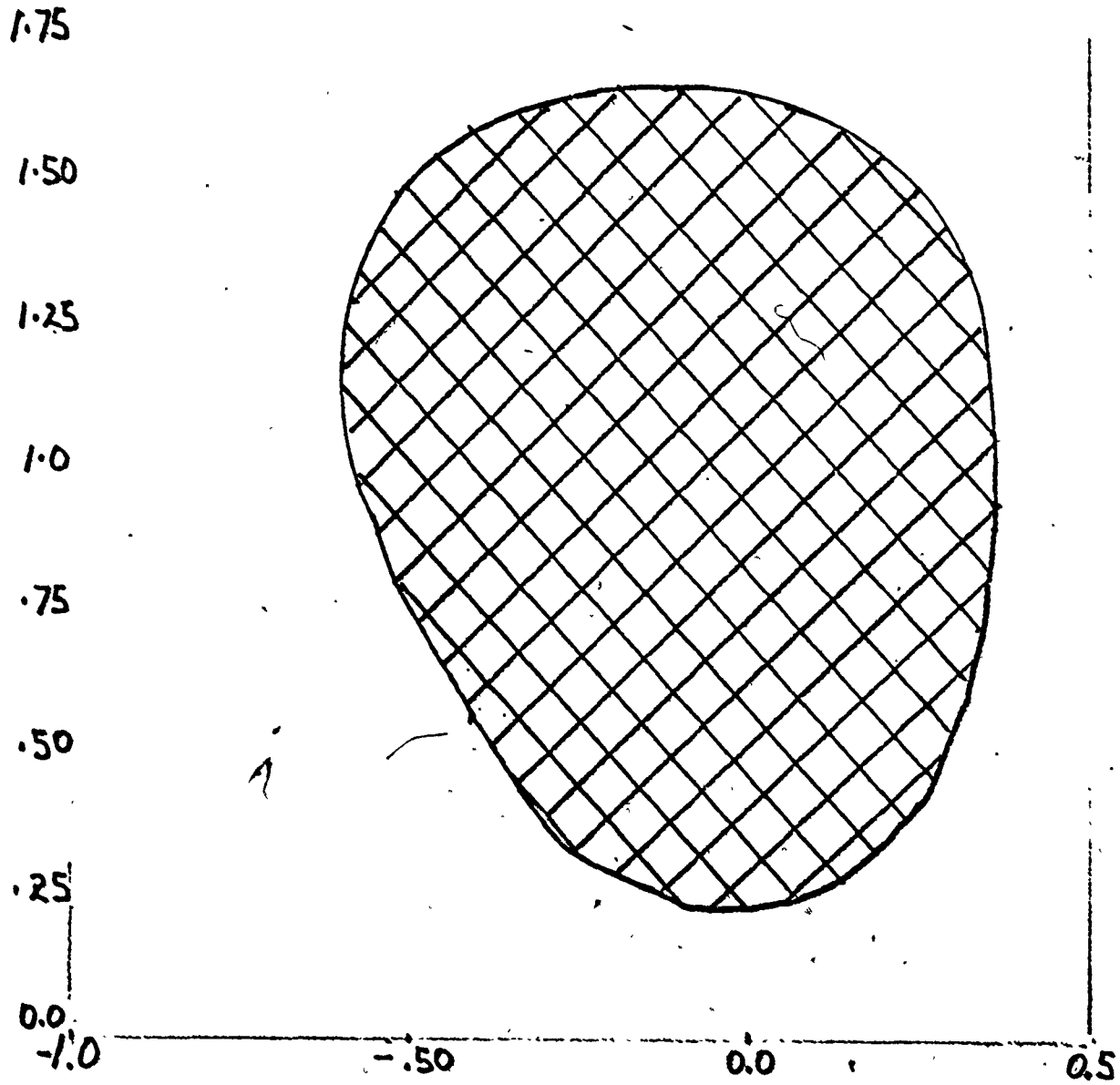
A calculation of an acceptance region for the hypothesis of normality was conducted as follows:

The data were standardized by the usual procedure. The standardized observations were tested for normality with values of the parameters ranging from (-1.5,1.5) for μ and (.20,2.20) for σ^2, the acceptance region based on all 5 EDF statistics at the 5% level of significance (figure 3) A surprisingly large region of acceptance was noted. This points out one of two things.

(1) For samples of small size, the traditional methods do not perform well. A small sample could have conceivably been drawn from an extremely wide range of

Figure 3

A Joint Acceptance Region of The Null Hypothesis of Normality



distributional models.

(2) For samples of small size, there is little justification for any goodness-of-fit testing. Because of wide variability based on test criteria, there is little risk in assuming any distributional assumptions which are convenient.

For completeness and a further emphasis of differences between uniformity and super-uniformity, I have calculated the "covariance" of the Case 3 transforms, and the Q and Q' statistics.

The covariance of the sample had the value cov. = .664. When compared to the significance points of Case 0 covariance, this is found to be near the modal value of the statistic. Even in Case 3, the covariance is at level of significance $\alpha = .07$. The covariance could conceivably by used as a test of fit criterion (in preliminary investigations, the covariance was found extremely ineffective as a test criterion), but the main value lies in pointing out the tremendous change which occurs when the PIT is applied using sample estimates.

Similarly, the Q and Q' statistics obtained values of 20.189 and 22.400 respectively. When compared to the tables of chi-square, degrees of freedom equal to 22, these correspond to levels of significance between the 50th and 25th percentages. According to this, by the misuse

of traditional methods, the Null Hypothesis is readily accepted. In preliminary studies, it was found that the Q and Q' statistics, using the proper Case 3 significance points which were obtained by Monte Carlo methods were adequate test criteria. Because they are easily computed, they may deserve further consideration as a possible test of fit statistic.

BIBLIOGRAPHY

Anderson, T. W. and Darling, D. A (1952). "Asymptotic Theory of Certain Goodness-of-Fit Criteria Based on Stochastic Processes". Ann. Math. Statis. 23, 193-212.

Birnbaum, Z. W. (1952). "Numerical Tabulation of Kolmogorov's Statistic For Finite Sample Size". Journ. Amer. Statist. Assoc. 47, 425-440.

Csorgo, M., Seshadri, V., and Yalovsky, M. (1973). "Some Exact Tests For Normality in the Presence of Unknown Parameters." Biometrika. 60, 507-532.

David, F. N., and Johnson, N. L. (1947). "The Probability Integral Transformation When Parameters are Estimated From the Sample". Biometrika, 35, 182-193.

Fisher, R. A. (1932) Statistical Methods For Research Workers. London and Edinburgh: Oliver and Boyd.

Kolmogorov, A. (1933). "Sulla Determinazion Empirica Di Una Legge Di Distribuzione". G. Inst. Ital. Attauri 4.

Kuiper, N. H. (1960). "Tests Concerning Random Points on a Circle". Proc. Koninkl. Nederl. Akad. Van Wetenschappen, series A 4 383.

Lillefors, H. W. (1967). "On the Kolmogorov-Simirnov Test For Normality with Mean and Variance Unknown". Journ. Amer. Statist. Assoc., 62, 399-402.

Pearson, E. S. (1938). "The Probability Integral Transformation For Testing Goodness-of-Fit and Combining Independent Tests of Significance". Biometrika, 30, 134-140.

Pearson, E. S. (1963). "Comparison of Tests For Randomness of Points on a Line". Biometrika, 50, 315-332.

Pearson E. S. and Stephens, M. A. (1962). "The Goodness-of-Fit Tests Based on W_n^2 and U_n^2". Biometrika, 49, 397-402.

Shafer, R. W., Finkelstein, J. M. and Collins, J. (1972). "On a Goodness-of-fit test for the Exponential Distribution with mean unknown" Biometrika, 59, 222-224.

Snedecor, G. U. (1946). Statistical Methods, Iowa State College Press, Ames, Iowa.

Srinivasan, R. (1970). "An approach to testing the Goodness-of-fit of incompletely specified distributions". Biometrika, 57, 604-611.

Stephens, M. A. (1965). :The goodness-of-fit statistic V_n: distribution and significance points." Biometrika, 52.

Stephens, M. A. (1972a). "Linear Functions of uniform order statistics". Technical report, Stanford University.

Stephens, M. A. (1974). "EDF statistics for goodness-of-fit and some comparisons. Journ. Amer. Statist. Assoc,. 69, 730-737.

Watson, G. S. (1961). "Goodness-of-fit tests on a circle." Biometrika, 48, 109.