EDF Statistics for Goodness of Fit and Some Comparisons

M. A. Stephens

Journal of the American Statistical Association, Vol. 69, No. 347 (Sep., 1974), 730-737.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197409%2969%3A347%3C730%3AESFGOF%3E2.0.CO%3B2-L

Journal of the American Statistical Association is currently published by American Statistical Association.

EDF Statistics for Goodness of Fit and Some Comparisons

M. A. STEPHENS*

This article offers a practical guide to goodness-of-fit tests using statistics based on the empirical distribution function (EDF). Five of the leading statistics are examined—those often labelled D, W^2, V, U^2, A^2—and three important situations: where the hypothesized distribution $F(x)$ is completely specified and where $F(x)$ represents the *normal* or *exponential* distribution with one or more parameters to be estimated from the data. EDF statistics are easily calculated, and the tests require only one line of significance points for each situation. They are also shown to be competitive in terms of power.

1. INTRODUCTION

The goodness-of-fit problem is as follows: given a random sample $x_1, x_2, \cdots, x_n$, to test H_o; the sample comes from a population with distribution function $F(x)$. The classical test for this problem is the χ^2-test, which has certain advantages: (a) it is well adapted for the case when $F(x)$ is discontinuous, i.e., represents a discrete distribution, and (b) it is known (at least to a good approximation) how to adapt the statistic for the case when parameters of $F(x)$ must themselves be estimated from the sample.

This article deals with another class of goodness-of-fit statistics—EDF statistics, so-called because they are based on a comparison of $F(x)$ with the empirical distribution function $F_n(x)$. For the case when $F(x)$ is continuous and completely specified (Case 0 in Section 2), it has long been known that, in general, EDF statistics give more powerful tests of H_o than χ^2. The disadvantage is that they are neither well adapted for discrete distributions, nor for the case when parameters must be estimated from the sample. This last drawback, together with the fact that they are considered more difficult to compute than χ^2, has undoubtedly prevented their wider application in practice. This is doubtful, certainly if the techniques for estimating parameters and class intervals as discussed, say, in Cramér [1, Ch. 30] and Kendall and Stuart [6, Ch. 30] are followed. Recent work has made it possible to use EDF statistics very easily in Case 0 and also for two very important practical situations—when the distribution tested is *normal* or *exponential*, with parameters to be estimated.

Here we give a practical guide to the use of EDF statistics in these three situations. Power studies are also given for Case 0 and for testing for normality. For this last case, EDF statistics have suffered recently from comparison with the W-statistic of Shapiro and Wilk [17]; no doubt, this is because the power studies reported in that paper gave very low power to EDF statistics. We show that, when used as described later, the powers are much higher than previously reported, and those of W^2 and A^2 are comparable to that of W, with which they appear to be highly correlated. From a practical point of view, the user may still prefer EDF statistics since these do not require special coefficients for each n. On the theoretical side, slightly more is known about W^2, U^2 and A^2, and the close liaison with the W-statistic suggests theoretical questions to be investigated.

We discuss statistics usually called D (derived from D^+ and D^-), W^2, V, U^2 and A^2. A suffix is often added to represent sample size, but this will be omitted. Because of the practical emphasis, definitions of these statistics are omitted and only the computing formulas are given.

Once a test statistic has been calculated, a table is entered to make the test. The choice of table depends on what is known of $F(x)$, so this is classified first in Section 2. The formulas and illustrations are in Section 3. Comments on the tables and computational details are given in Section 4, and the power studies are presented in Sections 5 and 6.

2. KNOWLEDGE OF $F(x)$

The tables to be used with the statistics depend on knowledge of $F(x)$, classified as follows:

Case 0: $F(x)$ continuous, completely specified. This is the classical case, and tables of significance points for all the statistics exist in the literature. For references see Stephens [21]. The use of Table 1.0 as described in Section 3 permits us to dispense with these tables.

Case 1: $F(x)$ is the normal distribution, σ^2 known, μ estimated by $\bar{x}$.

Case 2: $F(x)$ is the normal distribution, μ known, σ^2 estimated by $\sum_i (x_i - \mu)^2/n$ ($= s_1^2$, say)

Case 3: $F(x)$ is the normal distribution, both μ and σ^2 unknown, estimated by $\bar{x}$ and $s^2 = \sum_i (x_i - \bar{x})^2/(n-1)$

Case 4: $F(x) = 1 - \exp(-\theta x)$, i.e., the test is for exponentiality, with θ estimated by $1/\bar{x}$.

These Case numbers are chosen to match those in [22], where for each Case the asymptotic percentage points

* M.A. Stephens is professor, Departments of Applied Mathematics and Mathematics, McMaster University, Hamilton, Ontario, Canada. This research was supported by the U.S. Office of Naval Research and the National Research Council of Canada. The author is grateful to E.S. Pearson for permission to reproduce Tables 1.0, 1.3 and 1.4 with slight changes. The author also wishes to thank H. Braun, I. Scarowsky, A. Mendrinos and Laurel Ward, who assisted with the computations.

© Journal of the American Statistical Association
September 1974, Volume 69, Number 347
Theory and Methods Section

for W^2, U^2 and A^2 are found theoretically. In a test for normality, Case 3 is the important practical situation, though Case 2 sometimes arises, e.g., if one wishes to test residuals in regression analysis, transformed to give linear combinations of residuals which are theoretically independently normal with mean zero, but with variance unknown.

3. TEST PROCEDURES

3.1 Steps in making a test

We suppose the given values are in ascending order

$$x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n.$$

(a) When necessary, estimate parameters as described previously.

(b) Calculate $z_i = F(x_i)$, $i = 1, 2, \cdots, n$, where $F(x)$ may contain estimated parameters. For Case 3, this is done, in practice, in two stages: first calculate w_i from

$$w_i = (x_i - \bar{x})/\sigma \quad \text{(Case 1)};$$
$$w_i = (x_i - \mu)/s_1 \quad \text{(Case 2)};$$
$$w_i = (x_i - \bar{x})/s \quad \text{(Case 3)};$$

then z_i is the cumulative probability of a standard normal distribution, to the value w_i, found from standard tables or computer routines. For Case 4, $z_i = 1 - \exp(x_i/\bar{x})$.

(c) For all cases, the required statistic is calculated from the z-values (which are in ascending order) as follows:

1. The Kolmogorov statistics D^+, D^-, D:
   $D^+ = \max_{1 \leqslant i \leqslant n} [(i/n) - z_i]$;
   $D^- = \max_{1 \leqslant i \leqslant n} [z_i - (i-1)/n]$.
   $D = \max(D^+, D^-)$.
2. The Cramér-von Mises statistic W^2:
   $W^2 = \sum_{i=1}^{n} [z_i - (2i-1)/2n]^2 + (1/12n)$
3. The Kuiper statistic V:
   $V = D^+ + D^-$.
4. The Watson statistic U^2:
   $U^2 = W^2 - n(\bar{z} - \frac{1}{2})^2$ where $\bar{z} = \sum_{i=1}^{n} z_i/n$.
5. The Anderson-Darling statistic A^2:
   $A^2 = -\{\sum_{i=1}^{n} (2i-1)[\ln z_i + \ln(1 - z_{n+1-i})]\}/n - n$.

(d) To make the test using one of the preceding statistics (call it T), enter Table 1, part K for Case K and calculate T^*, the *modified* statistic; reject H_0 at a chosen level of significance if T^* exceeds the significance point given on the same line.

The test described is the usual upper-tail test; on occasion, particularly when a transformation is used to produce uniform observations (Case 0), the lower tail is necessary (see, e.g., [15] and [13]).

3.2 Points on a circle

Only statistics U^2 and V should be used for such points, and any suitable origin may be chosen; the other statistics may take different values with different origins. U^2 and V may also, of course, be used for points on a line; they have different power properties from A^2, W^2 and D, particularly for Case 0 (cf. Section 5).

3.3 Illustrations

Illustration 1. Pearson [13] discusses the use of four of the preceding EDF statistics (A^2 is not included) on eight examples. For Example 1, Pearson has 20 values of warp breaks; these are transformed to a set of values to be tested to come from a uniform distribution between 0 and 1. Thus we have a Case 0 situation. For D, Pearson has the value 0.356. In Table 1.0, the modified D then becomes $D^* = 0.356(4.472 + 0.12 + 0.11/4.472) = 1.643$, and reference to Table 1.0 gives significance at a level just less than one percent (Pearson gives 0.9 percent).

For U^2, the value is 0.298 and modified value is $U^* = 0.304$, significant at the 0.005 level, in agreement with Pearson's results.

Illustration 2. The following values of men's weights in pounds, first given by Snedecor, were used by Shapiro and Wilk [17] as an illustration of a test for normality: 148, 154, 158, 160, 161, 162, 166, 170, 182, 195, 236. The mean is 172 and the standard deviation 24.95. For a test for normality (Case 3), the values of the modified statistics are: $D^* = 0.924$, $W^{2*} = 0.171$, $A^{2*} = 1.095$, $U^{2*} = 0.150$, $V^* = 1.544$. Table 1.3 gives α-values 0.035, 0.01, 0.01, 0.035. The one percent values given by the W^2 and A^2 statistics agree closely with the result of Shapiro and Wilk; this result, as well as the higher values of α given by D and V are in agreement with comments made on the powers of EDF statistics in Section 5.

Illustration 3. Proschan [14] gives 213 values of times t to failure of air-conditioning equipment in aircraft, believed to follow an exponential distribution with θ unknown. $\bar{t}$ is 93.14, so $\hat{\theta}$ in Case 4 is 0.0107. Proschan uses the well-known technique of adding D_α to $F_n(t)$ or subtracting it from $F_n(t)$ to obtain a confidence interval for $F(t)$ and uses this to construct a confidence interval for the fraction surviving after time t. D_α refers to the upper-tail critical value of the Kolmogorov statistic at level α, giving a $(1 - \alpha)$ percent confidence interval. In this case, Proschan uses $D_\alpha = 1.358/\sqrt{n} = 0.0931$; this corresponds to $\alpha = 0.05$, and 1.358 comes from the Case 0 Table 1.0. In fact, 1.358 should be replaced by 1.094 from Table 1.4; this gives a much narrower confidence interval in Proschan's Figure 1 and is a direct result of estimating the parameter. Proschan notes that although the test for exponentiality does not (apparently) lead to rejection, the variables nevertheless do not appear to be exponentially distributed. When the test is made, values of test statistics are $\sqrt{n}D = 1.067$, $W^2 = 0.324$, $U^2 = 0.190$, $V = 1.588$ and $A^2 = 1.691$; when Table 1.0 is used, none of these is significant at the 10 percent level—D not even at the 15 percent level. However, when the correct Table 1.4 is used, the significance levels α corresponding to these values are approximately 0.06, 0.012, 0.025, 0.07 and

1A. Modifications to D, V, W^2, U^2, A^2, Cases 0, 3 and 4

Statistic T	Modified form T*	Percentage points for T*				
		15.0	10.0	5.0	2.5	1.0
	1.0 Modifications for the test when F(x) is completely known[a]					
$D^+(D^-)$	$D^+(\sqrt{n} + 0.12 + 0.11/\sqrt{n})$	0.973	1.073	1.224	1.358	1.518
D	$D(\sqrt{n} + 0.12 + 0.11/\sqrt{n})$	1.138	1.224	1.358	1.480	1.628
V	$V(\sqrt{n} + 0.155 + 0.24/\sqrt{n})$	1.537	1.620	1.747	1.862	2.001
W^2	$(W^2 - 0.4/n + 0.6/n^2)(1.0 + 1.0/n)$	0.284	0.347	0.461	0.581	0.743
U^2	$(U^2 - 0.1/n + 0.1/n^2)(1.0 + 0.8/n)$	0.131	0.152	0.187	0.221	0.267
A^2	For all $n \geq 5$:	1.610	1.933	2.492	3.070	3.857
	1.3 Modifications for a test for normality, μ and σ^2 unknown[a]					
D	$D(\sqrt{n} - 0.01 + 0.85/\sqrt{n})$	0.775	0.819	0.895	0.955	1.035
V	$V(\sqrt{n} + 0.05 + 0.82/\sqrt{n})$	1.320	1.386	1.489	1.585	1.693
W^2	$W^2(1 + 0.5/n)$	0.091	0.104	0.126	0.148	0.178
U^2	$U^2(1 + 0.5/n)$	0.085	0.096	0.116	0.136	0.163
A^2	$A^2(1 + 4/n - 25/n^2)$	0.576	0.656	0.787	0.918	1.092
	1.4 Modifications for a test for exponentiality, θ unknown[a]					
D	$(D - 0.2/n)(\sqrt{n} + 0.26 + 0.5/\sqrt{n})$	0.926	0.990	1.094	1.190	1.308
V	$(V - 0.2/n)(\sqrt{n} + 0.24 + 0.35/\sqrt{n})$	1.445	1.527	1.655	1.774	1.910
W^2	$W^2(1 + 0.16/n)$	0.149	0.177	0.224	0.273	0.337
U^2	$U^2(1 + 0.16/n)$	0.112	0.130	0.161	0.191	0.230
A^2	$A^2(1 + 0.6/n)$	0.922	1.078	1.341	1.606	1.957

[a] Tables 1.0, 1.3 and 1.4, with slight changes, have appeared in [13a]. Among the changes are revised versions of the modified forms for A^2 in Tables 1.3 and 1.4, those included in this article give more accurate results, though the differences are very small.

0.02, respectively. Thus, use of W^2, U^2 or A^2 would have led to rejection at the five percent level, and even D is nearly significant at this level. We show later that, for Case 3, W^2 and A^2 are preferred statistics in terms of better power and that D is poor; although power studies are not given here for Case 4, these studies indicate good performance for W^2 and A^2 for this Case also. Thus Proschan's conclusion would have been supported by the more powerful W^2, U^2 or A^2.

4. TABLES AND COMPUTATIONAL DETAILS

4.1 Tables

Table 1, in five parts, contains the formulas and percentage points with which tests will be made; Table 1.K is to be used with Case K. For Cases 0, 3 and 4, the three most practical situations, the Tables 1.0, 1.3 and 1.4 have been grouped together. Table 1.0 comes from Stephens [21] with the Anderson-Darling statistic replacing the A in that paper. The new A^2 converges so rapidly that no modification is required for any realistic situation ($n \geq 5$). This was suggested by Marshall [12] and has been confirmed by Monte-Carlo studies by Lewis [8]. Tables 1.1 to 1.4, the asymptotic points (those given for use with the modified forms), have been calculated theoretically for W^2, U^2 and A^2 [22].

For n finite, significance points for all statistics are difficult to find theoretically and have been found from Monte-Carlo studies by the author. These mostly involved 10,000 samples for each of many values of n. The points for a given significance level were plotted against $1/n$ or $1/\sqrt{n}$, and smoothed; for D and V they were also extrapolated to obtain the asymptotic points. The original five percent and one percent points, for all statistics except A^2, were given in [19, 20]. Since A^2 is found to be a powerful statistic, the points for A^2 are now added in Table 2. The modified forms were calculated from the smoothed Monte-Carlo points. For details on the general procedure, see [21]. Other Monte-Carlo studies, each one usually for only one or two statistics,

1B. Percentage Points for Cases 1 and 2

Statistic	n	Percentage level (%)				
		15	10	5	2.5	1
	1.1 Asymptotic points for W^2, U^2, A^2, Case 1 (exact)					
W^2			0.135	0.165	0.196	0.237
U^2			.128	.157	.187	.227
A^2			.908	1.105	1.304	1.573
	1.2 Percentage points for Case 2: (Monte Carlo results for D, V; exact results for W^2, U^2, A^2)					
$\sqrt{n}D$	10	1.050	1.138	1.270	1.380	1.530
	20	1.070	1.160	1.290	1.415	1.570
	50	1.080	1.170	1.310	1.432	1.595
	100	1.100	1.180	1.320	1.440	1.610
	∞	1.120	1.190	1.333	1.455	1.625
$\sqrt{n}V$	10	1.305	1.385	1.500	1.595	1.710
	20	1.345	1.410	1.535	1.642	1.770
	50	1.380	1.450	1.570	1.680	1.810
	100	1.390	1.470	1.590	1.697	1.825
	∞	1.410	1.490	1.612	1.720	1.845
W^2	all $n \geq 5$		.329	.443	.562	.723
U^2	all $n \geq 5$		.123	.153	.182	.221
A^2	all $n \geq 5$		1.760	2.323	2.904	3.690

2. Percentage Points for A^2: Cases 3 and 4 (Monte Carlo Points for Finite n; Exact Asymptotic Points)

n	Percentage level (%)				
	15	10	5	2.5	1
			Case 3		
10	.514	.578	.683	.779	.926
20	.528	.591	.704	.815	.969
50	.546	.616	.735	.861	1.021
100	.559	.631	.754	.884	1.047
∞	.576	.656	.787	.918	1.092
			Case 4		
10	.887	1.022	1.265	1.515	1.888
20	.898	1.045	1.300	1.556	1.927
50	.911	1.062	1.323	1.582	1.945
100	.916	1.070	1.330	1.595	1.951
∞	.922	1.078	1.341	1.606	1.957

have been reported for Cases 3 and 4 as follows: Lilliefors [9, 10], statistic D; Van Soest [24], statistics D and W^2; Koerts and Abrahamse [7] and Louter and Koerts [11], statistic V. The points given by these authors agree well with the values given by using Tables 1.3 or 1.4, except for some differences in estimates of asymptotic points for D and V. Those given here are based on sample sizes up to 100; other authors have $n \leqslant 40$. In any event the practical difference is very small. The various Monte-Carlo points in the literature may be used to obtain an estimate of α', the true significance level of a point calculated from the modified forms at level α. For $n = 10$ and 25, the estimated difference $|\alpha' - \alpha|$ never exceeds 0.002 for $\alpha = 0.05$ and 0.01 and only once exceeds 0.005 for $\alpha = 0.10$.

For Case 1, the most unlikely to arise in practice, only asymptotic points, obtained from [22], are known. These are given in Table 1.1. For Case 2, the asymptotic points are supported by some Monte-Carlo points given in Table 1.2; no modifications have been calculated.

5. POWER COMPARISONS: TEST FOR UNIFORMITY

5.1 Results of Power Studies for Case 0

Table 3 gives results based on at least 1,000 Monte-Carlo samples, drawn from given distributions and tested for uniformity. This table also gives the percentage of samples declared significant by various test statistics.

If $F(x)$ is completely specified, the z_i should be uniformly distributed between 0 and 1, written $U(0, 1)$. Power studies have therefore been confined to a test of this hypothesis concerning z, when the z_i are in fact drawn from alternative distributions. If the variance of the hypothesized $F(x)$ is correct but the mean is wrong, the points z_i will tend to move toward 0 or 1; if the mean is correct but the variance wrong, the points will move to each end, or will move towards 0.5.

The following alternatives A, B, C were chosen to give patterns of z-values corresponding to these situations:

$$A: F(z) = 1 - (1 - z)^k, \qquad 0 \leqslant z \leqslant 1;$$

$$B: F(z) = 2^{k-1}z^k, \qquad 0 \leqslant z \leqslant 0.5;$$
$$F(z) = 1 - 2^{k-1}(1 - z)^k, \qquad 0.5 \leqslant z \leqslant 1,$$

$$C: F(z) = 0.5 - 2^{k-1}(0.5 - z)^k, \qquad 0 \leqslant z \leqslant 0.5;$$
$$F(z) = 0.5 + 2^{k-1}(z - 0.5)^k, \qquad 0.5 \leqslant z \leqslant 1.$$

A gives points closer to zero than expected on the hypothesis of uniformity; B gives points near 0.5; C gives two clusters close to 0 and 1.

Table 3 shows that statistics D, W^2 and A^2 will detect a change in mean better than the others, and V and U^2 will detect a change in variance. This is to be expected from the geometry associated with their null distributions. W^2 and A^2 tend to be better than D, and U^2 slightly better than V. In practice, it would always seem worth while to look at W^2, U^2 and A^2. Historically, D has been the most used EDF statistic, but of the four, it tends to be the least powerful, overall. For references to earlier work on Case 0, see [6].

Included in Table 3 are some results for X^2 (with expected number 5 per cell, i.e., degrees of freedom 3 for $n = 20$, 7 for $n = 40$). X^2 is not at all as powerful as EDF statistics. Results are given also for statistic $Q = \sum_i \ln z_i$ which, on H_o, has the χ^2_{2n} distribution. Q is included for comparison, since it is most powerful against alternative distribution A, and it is rare in goodness-of-fit work to have such a standard available. Some interesting comparisons, but only for D and V, are given by Koerts and Abrahamse [7] for tests for a normal distribution $N(0, 1)$, against $N(\mu, \sigma^2)$, with 16 pairs of μ, σ. Durbin and Knott [5] and Stephens [23] give theoretical asymptotic powers for normal and exponential tests against the same family of alternatives.

3. Power Comparisons, Test for Uniformity (case 0)[a]

N	n	D	W^2	V	U^2	A^2	Q	X^2
A, k = 1.5	10	23	27	18	19	24	43	—
	20	38	46	25	28	46	68	—
	40	60	70	43	43	—	89	40
A, k = 2.0	10	54	60	35	35	58	—	—
	20	78	87	61	60	87	97	59
	40	98	99	91	89	—	100	89
B, k = 1.5	10	9	7	22	23	6	—	—
	20	13	11	32	34	10	11	—
	40	19	22	57	61	—	—	39
B, k = 2.0	10	9	7	40	44	6	—	—
	20	25	25	71	77	28	25	—
	40	56	72	96	98	—	—	85
B, k = 3.0	10	21	21	81	86	18		
	20	63	79	99	99	84		
C, k = 1.5	20	25	20	36	37	28		
	40	36	32	58	63	—		
C, k = 2.0	20	47	44	71	77	54		
	40	71	80	96	98	—		

[a] This table gives the percentage of samples significant, when the population is as shown and each sample has size n. The test is at the 10% level.

5.2 Correlation between test statistics

Clearly, there are fairly strong correlations between the various test statistics. To gain some information on these, a matrix R was produced for each power study, giving, in cell (r_{ij}), the number of samples significant by two of the test statistics, say T_i and T_j; cell (r_{ii}) contains the number significant by T_i alone.

A typical R matrix is in Table 6 of [19]. We include another in Table 4.

4. Typical Output Matrix of Power Studies[a]

Statistic	D	W^2	V	U^2	A^2
D	94				
W^2	71	90			
V	48	40	95		
U^2	48	43	69	87	
A^2	70	83	43	47	98

[a] The test is for uniformity, n = 20, α = .10. The samples are actually from a uniform distribution. The matrix gives number of 1000 samples significant by both statistics (row and column).

6. POWER COMPARISONS: TEST FOR NORMALITY

6.1 Other statistics for testing for normality (Case 3)

In Section 6.2 we discuss power results for the important problem of testing for normality when the parameters μ and σ are unknown. Tables 5 and 6 contain results for EDF statistics and for statistics W, D_A and W'. These statistics will be briefly described. W is a statistic introduced by Shapiro and Wilk [17], and D_A, W' are subsequent extensions introduced by d'Agostino [2, 3] (there called D), and Shapiro and Francia [16].

The W statistic is based on a comparison of two estimates of σ^2: the usual s^2, and the estimate $\hat{\sigma}$ obtained by least squares estimation of the slope, when the ordered observations x_i are plotted against expected values of order statistics from a standard normal distribution. From a practical viewpoint, this procedure has some disadvantages. For each n, a different set of coefficients is required for the estimation of $\hat{\sigma}$; these are not available for $n > 50$. Exact coefficients are given by Shapiro and Wilk, for $n \leqslant 20$, and approximate values for $20 < n \leqslant 50$. Further, a set of significance points is needed for each n.

The statistics D_A and W' are esssentially introduced to extend the W statistic for use beyond $n = 50$. Both use estimators $\hat{\sigma}$ which are asymptotically less efficient than that used in W; d'Agostino needs no special coefficients, and Shapiro and Francia need the expected values of standard normal order statistics as coefficients. These are, of course, available. The null distribution theory of these three statistics is difficult; even asymptotic theory is lacking. Thus, for the W and W' statistics, Monte-Carlo methods were used to provide the significance points in the papers cited; for D_A, approximate points are given, using moments in connection with Cornish-Fisher expansions or the fitting of Pearson curves.

Significant values of these statistics are in the lower tail for W, and in both tails for D_A. Low values have been used for W' as for W. Significant values are not always readily interpretable in terms of properties of the parent population; the authors, in introducing W and D_A, have decided on the critical regions from the results of Monte-Carlo studies.

5. Power Comparisons, Test for Normality (Case 3)[a]

Population (β_1,β_2)	n	D	W^2	V	U^2	A	χ^2	W
χ_1^2	10	51	64	65	63	67	—	69
(8,15)	20	86	94	94	93	—	44	97
	30	98	100	100	100	—	75	100
Exponential	10	30	38	36	37	41	—	43
(4,9)	20	59	74	71	70	82	27	85
	30	76	90	88	86	95	52	97
χ_3^2	10	23	—	—	—	—	—	—
(2.67,7)	20	40	55	50	—	—	—	—
	30	57	—	68	—	—	—	—
χ_4^2	20	33	45	—	—	—	—	50
(2,6)								
χ_{10}^2	20	18	23	—	—	—	—	29
(.8,4.2)								
Lognormal	10	45	56	53	53	59	—	60
(38,114)	20	78	88	84	85	91	40	93
	30	94	99	97	98	99	70	99
Uniform	20	12	16	17	18	21	8	21
(0,1.8)	30	17	26	25	29	—	12	42
	50	28	47	44	52	—	21	88
Cauchy	10	58	62	60	61	62	—	59
—	20	86	88	87	88	98	—	87
Laplace	10	13	16	14	15	16	—	14
(0,6)	20	22	26	22	25	26	12	25
	30	29	35	31	34	—	26	30
Student-t_1	20	95	88	—	—	—	—	88
Student-t_3	10	17	18	18	18	20	—	—
	20	23	28	25	29	32	—	—
Student-t_4	20	17	21	18	20	23	—	24
Student-t_6	20	10	12	11	11	14	—	15

[a] The table gives the percentage of samples significant, when the population is as shown, and the samples have size n. The test is at the 5% level.

6.2 Power comparisons

Tables 5 and 6 give the percentage of M Monte-Carlo samples, each of size n and drawn from the population given, which were declared significant by the statistics quoted when the test for normality, of size α, was applied. In Table 5, M was at least 1,000, and in Table 6, M was at least 2,000 for $n = 50$ and at least 500 for $n = 90$. With these values one can make a good comparison of relative power. In Table 5, $\alpha = 0.05$; then comparisons can be made, for $n = 20$, with results reported by Shapiro and Wilk [17]. More extensive results were later reported by Shapiro, Wilk and Chen [18]. These results included power studies for older statistics used in testing for normality, such as χ^2, b_1 and b_2, or u = (range/standard deviation). These are on the whole much inferior to W. Shapiro and Wilk [17] also included power studies for D, W^2 and A^2 (there called D, CVM, $WCVM$) and found

6. Power Comparisons, Test for Normality (Case 3)[a]

Population (β_1,β_2)	n	D_A^U	D_A^L	D_A	W	W'	W^2	A^2
Uniform	50	69	0	69	95	—	61	75
(0,1.8)	90	96	0	96	—	98	88	95
Cauchy	50	30	0	30	100	—	98	100
	90	55	0	55	—	100	100	100
Exponential	50	0	92	92	100	—	100	100
(4,9)	90	0	99	99	—	100	100	100
χ_4^2	50	0	66	66	99	—	89	94
(2,6)	90	0	82	82	—	100	100	100
Laplace	50	0	69	69	50	—	63	64
(0,6)	90	0	91	91	—	90	86	86
Lognormal (38,114)	50	0	99	99	100	—	100	100
Weibull, K = 2	50	5	15	20	59	—	32	45
(.63,3.25)	90	4	18	22	—	83	64	76
Tukey, $\lambda = 5$	50	0	14	14	24	—	38	37
(0,2.9)	90	1	18	19	—	41	64	62
Student-t_4	50	0	58	58	43	—	48	52
	90	0	78	78	—	78	66	69

[a] The table gives the percentage of samples significant, when the population is as shown, and each sample is of size n. The test is at the 10% level. The columns headed D_A^U and D_A^L show those samples significant at the upper and lower 5% levels for D_A

these statistics also greatly inferior to W. However, their results are misleading. This is because, in using EDF statistics to test for normality, when in fact the sample came from another distribution, it was supposed that the true mean and variance were known to the tester. Then Case 0 was assumed and, effectively, Table 1.0 was used. But for a true comparison with W (and later, with D_A and W') we should allow the tester to estimate his own mean and variance, and follow the procedure of Case 3. This has been done to produce the results in Tables 5 and 6.

Other comparisons have been given by Koerts and Abrahamse [7] for statistics D, V, by van Soest [24] for statistics D, W^2 and by Lilliefors [9] for statistic D only. The results for χ_3^2, χ_4^2, χ_{10}^2 and Student-t_1 are taken from these authors: their other values agree with Table 5, except for Lilliefors' Student-t_3 results for D, which appear to be incorrect.

The tables show that EDF statistics, when used as described for Case 3, have powers roughly comparable with those of W. The most widely known statistic, the Kolmogorov D, gives the poorest performance, as for Case 0. The difference between V and U^2 on the one hand, and D and W^2 on the other, noted for Case 0, largely disappears when one is allowed to estimate the mean and variance of the normal distribution (as might be expected). Overall, A^2 and W^2 appear to be the best pair of EDF statistics.

Table 6 includes the statistics D_A and W'. The value $\alpha = 0.10$ has been used to enable comparisons to be made with those of d'Agostino [2] which concern only D_A and W for $n = 50$. They match closely the results in Table 6 for these two statistics. Table 6 includes, in columns D_A^U and D_A^L, results using D_A at level $\alpha = 0.05$ in the upper or lower tail only; they bear out d'Agostino's comments that, overall, a two-tail test is needed when D_A is used. However, D_A gives relatively poor results; a possible explanation is as follows. In the calculation of W, generalized least-squares, known to have good properties, is used for one estimate of σ (and the usual s for the other). This, however, necessitates using a set of numerical coefficients different for each n. In W' a very good approximation is used, but again numerical values are needed; the advantage is that the coefficients are available for $n > 50$. In D_A, these coefficients are replaced by values which can easily be calculated, but a price is paid, and power drops considerably.

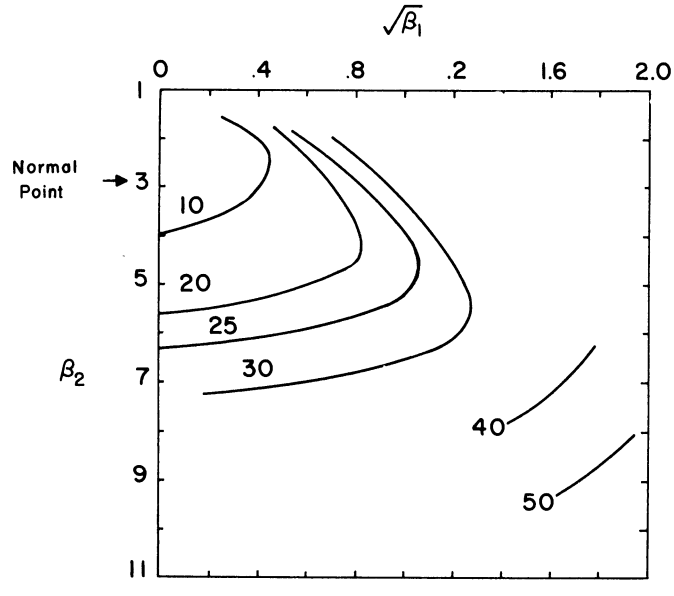
Statistics W, for n up to 50, and W', for n beyond 50, appear slightly better than the best EDF competitors W^2 and A^2; the user must decide whether the extra power in using W or W' is worth the extra effort in calculation. The closeness of the results for W^2 and, particularly, A^2 to those for W suggests strong correlation between these three statistics; the **R** matrices strongly bear this out. On the whole, W^2 and A^2 give very good performance considering that for all n, only one formula is needed for each statistic. Further, if the modified forms of Table 1 are used, only one line of significance points is required for each statistic. Furthermore significant values have a straightforward interpretation.

It is an interesting result of these studies that, in a test of this type, it is better *not* to have the true mean and variance available but to estimate it from the data! It appears that since one is trying, in effect, to fit a density of a certain shape to the data, the precise location and scale is relatively unimportant, and being tied down to fixed values, even correct ones, is more a hindrance than a help. The paper by Durbin and Knott [5] deals with Case 0, but brings out that W^2 (and similarly A^2 and U^2) can be split into components which can be used separately to test for location, scale and other (e.g. shape) effects; see also [23]. Estimation of μ and σ presumably reduces the influence of the early components; it will be interesting to see the extension of their work to the Case 3 situation.

6.3 Contour Maps of Power

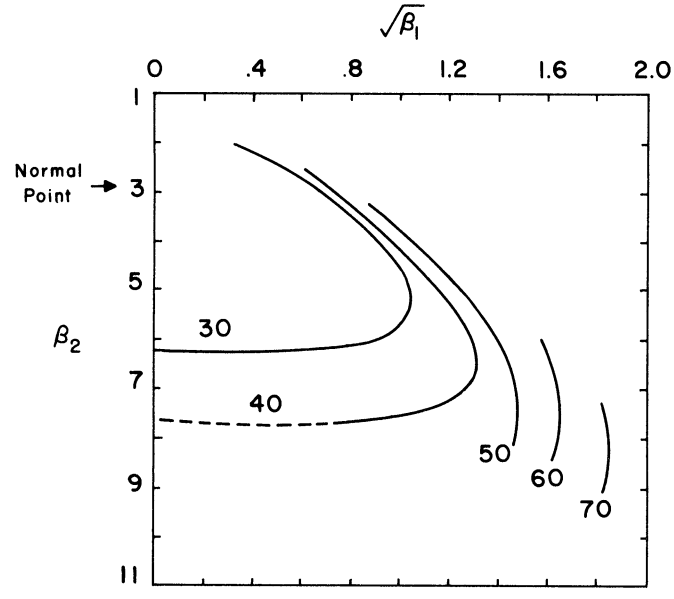
An effective way to demonstrate the influence of the higher moments of the true distribution on power, and also to compare the different statistics, is to plot contour maps of power on a β_1, β_2 diagram; this has been done for D, W and A^2 in Figures A–C. The maps shown are for $n = 20$, $\alpha = 0.05$; the diagram uses the conventional $\sqrt{\beta_1}$, β_2 grid as given in the discussion of Pearson curves in [13a]. The lines give *isodynes*, i.e., lines of constant power which are drawn from the results of the Monte-Carlo studies in Tables 5 and 6, supplemented by others. The power (%) is marked on the lines. The alternatives were produced from combinations of easily generated

A. Isodynes[a]: Kolmogorov-Smirnov Statistic D



[a] The test is a test for normality (Case 3), with $N = 20, \alpha = 0.05$.

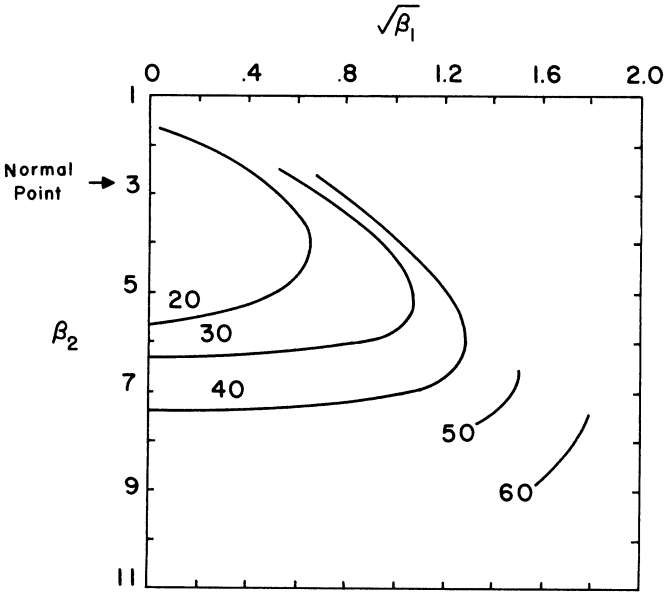
C. Isodynes[a]: Shapiro-Wilk Statistic W



[a] See note to Figure A.

distributions, and their (β_1, β_2) values form a grid of approximately 40 points on the diagram where the lines are drawn. These maps, although rough, give an easier guide for power comparisons than extensive tables for different statistics. Different true distributions with approximately the same (β_1, β_2) point gave nearly the same powers for any one statistic, as expected. One interesting feature is that, for fairly large β_2, a change of β_1 from 0 to 1 hardly affects power at all for W and A^2; this probably demonstrates the fact that both these statistics give considerable weight to observations in both tails.

B. Isodynes[a]: Anderson-Darling Statistic A^2



[a] See note to Figure A.

6.4 Final remarks

W (and W') uses a linear combination of order statistics to estimate σ; one could guard against various alternatives to normality by varying the weights in these combinations, and so produce different powers. Similarly, slight adaptations of existing EDF statistics will vary the powers against different alternatives. Although there has been a tendency recently to propose such statistics, there often seems to be very little justification. The mathematics is often intractable, and Monte-Carlo methods must be used for both null percentage points and power studies. This in itself is not an objection if the statistic proposed can be shown to have strong computational or other advantages, or systematically better power against a wide class of alternatives when compared with its best opposition. Power studies in support of new statistics have often been skimpy; since, after all, we don't usually know the exact alternative to the null distribution, it is urged that a range of alternatives comparable to those used by Shapiro and Wilk [17] or in Table 6 of this article should always be investigated. Even if a new statistic is proposed and claimed to have advantages only for a certain type of alternative (say very skew, or long-tailed), for a real comparison with statistics of the W or A^2 type, we need to see how the new statistic fares when used on other alternatives also.

[*Received February 1973. Revised January 1974.*]

REFERENCES

[1] Cramér, H., *Mathematical Methods of Statistics*, Princeton, N.J.: Princeton University Press, 1946.

[2] d'Agostino, Ralph B., "An Omnibus Test of Normality of Moderate and Large Size Samples," *Biometrika*, 58 (August 1971), 341–48.

[3] ———, "Small Sample Probability Points for the D Test of Normality," *Biometrika*, 59 (April 1972), 219–21.
[4] Durbin, J., "Some Methods of Constructing Exact Tests," *Biometrika*, 48 (June 1961), 41–55.
[5] ——— and Knott, M., "Components of Cramer-Von Mises Statistics. 1," *Journal of the Royal Statistical Society*, Ser. B, 34, No. 2 (1972), 290–307.
[6] Kendall, M.G. and Stuart, A., *The Advanced Theory of Statistics*, Vol. 2, London: Charles W. Griffin and Co., Ltd., 1961.
[7] Koerts, J. and Abrahamse, A.P.J., *On the Theory and Application of the General Linear Model*, Rotterdam University Press, 1969.
[8] Lewis, Peter A.W., "Distribution of the Anderson-Darling Statistic," *Annals of Mathematical Statistics*, 32 (December 1961), 1118–24.
[9] Lilliefors, H.W., "On the Kolmogorov-Smirnov Tests for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, 62 (June 1967), 399–402.
[10] ———, "On the Kolmogorov-Smirnov Tests for the Exponential Distribution with Mean Unknown," *Journal of the American Statistical Association*, 64 (March 1969), 387–89.
[11] Louter, A.S. and Koerts, J., "On the Kuiper Test for Normality with Mean and Variance Unknown," *Statistica Neerlandica*, 24 (1970), 83–7.
[12] Marshall, A.W., "The Small-Sample Distribution of nw_n^2," *Annals of Mathematical Statistics*, 29 (March 1958), 307–9.
[13] Pearson, E.S., "Comparison of Tests for Randomness of Points on a Line," *Biometrika*, 50 (December 1963), 315–25.
[13a] ——— and Hartley, H.O., eds., *Biometrika Tables for Statisticians, Vol. 2*, 2nd ed., New York: Cambridge University Press, 1972.
[14] Proschan, F., "Theoretical Explanation of Observed Decreasing Failure Rate," *Technometrics*, 5 (August 1963), 375–83.
[15] Seshadri, V., Csorgo, M. and Stephens, M.A., "Tests for the Exponential Distribution using Kolmogorov-Type Statistics," *Journal of the Royal Statistical Society*, Ser. B, 31, No. 3 (1969), 499–509.
[16] Shapiro, S.S. and Francia, R.S., "An Approximate Analysis of Variance Test for Normality," *Journal of the American Statistical Association*, 67 (March 1972), 215–16.
[17] ——— and Wilk, M.B., "An Analysis-of-Variance Test for Normality (Complete Samples)," *Biometrika*, 52 (December 1965), 591–611.
[18] ———, Wilk, M.B. and Chen, H.J., "A Comparative Study of Various Tests for Normality, "*Journal of the American Statistical Association*, 63 (December 1968), 1343–72.
[19] Stephens, M.A., "Test for Normality," Technical Report No. 152, Department of Statistics, Stanford University, 1969.
[20] ———, "Kolmogorov-Type Tests for Exponentiality when the Scale Parameter Is Unknown," Technical Report No. 154, Department of Statistics, Stanford University, 1970.
[21] ———, "Use of Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics Without Extensive Tables," *Journal of the Royal Statistical Society*, Ser. B, 32, No. 1 (1970) 115–22.
[22] ———, "Asymptotic Results for Goodness-of-Fit Statistics With Unknown Parameters," Technical Reports 159, 180, Department of Statistics, Stanford University, 1971.
[23] ———, "Components of Goodness-of-Fit Statistics," *Annales de l'Institut Henri Poincaré*, Ser. B, 10 (April 1974), 37–54.
[24] Van Soest, J., "Some Experimental Results Concerning Tests of Normality," *Statistica Neerlandica*, 21 (1967) 91–7.