THE KOLMOGOROV-SMIRNOV TEST FOR GOODNESS OF FIT

FRANK J. MASSEY, JR.
University of Oregon

The test is based on the maximum difference between an empirical and a hypothetical cumulative distribution. Percentage points are tabled, and a lower bound to the power function is charted. Confidence limits for a cumulative distribution are described. Examples are given. Indications that the test is superior to the chi-square test are cited.

1. INTRODUCTION

FREQUENTLY a statistician is called upon to test some hypothesis about the distribution of a population. If the test is concerned with the agreement between the distribution of a set of sample values and a theoretical distribution we call it a "test of goodness of fit."

Some tests have been developed in which the sampling distribution of the test statistic depends explicitly upon the form of, or the value of some parameter in, the distribution of the population. For example, in the test for normality which uses the g_1 and g_2 statistics (see Snedecor [11] page 176) the distributions of g_1 and g_2 are dependent on the form of the population. Similarly, the statistic $t=\sqrt{N}(\overline{X}-\mu)/s$ has Student's distribution only if the population is normal.

Attempts have been made to find test statistics whose sampling distribution does not depend upon either the explicit form of, or the value of certain parameters in, the distribution of the population. Such tests have been called non-parametric or distribution-free tests. Probably the most widely used of such tests is the χ^2 test.

In this paper an alternative distribution-free test of goodness of fit is discussed, and some evidence is presented indicating that when it is applicable it may be a better all-around test than the chi-square test. Also, a technique for estimating the cumulative distribution of a population is discussed, including a method of determining the necessary sample size for desired precision. Only the case where the cumulative distribution of the population is continuous is discussed. This, of course, excludes discrete populations.

The test for goodness of fit described here has been suggested by Kolmogorov [3], Smirnov [9], Scheffé [8], and Wolfowitz [14]. The limiting distribution of the test-statistic, d, was derived by Kolmogorov [3] and by Smirnov [9]. Feller [2] and Doob [1] have simplified

and unified the proofs. A table of the limiting distribution was given by Smirnov [10]. The method of evaluating the distribution of d for small samples was given by Massey [5], as was the construction of the lower bound to the power function [6].

2. THE TEST

Suppose that a population is thought to have some specified cumulative frequency distribution function, say $F_0(x)$. That is, for any specified value of x, the value of $F_0(x)$ is the proportion of individuals in the population having measurements less than or equal to x. The cumulative step-function of a random sample of N observations is expected to be fairly close to this specified distribution function. If it is not close enough, this is evidence that the hypothetical distribution is not the correct one.

If $F_0(x)$ is the population cumulative distribution, and $S_N(x)$ the observed cumulative step-function of a sample (i.e., $S_N(x) = k/N$, where k is the number of observations less than or equal to x), then the sampling distribution of $d = \text{maximum } |F_0(x) - S_N(x)|$ is known, and is independent of $F_0(x)$ if $F_0(x)$ is continuous.

Table 1 gives certain critical points of the distribution of d for various sample sizes. For example, at a 0.20 level of significance, the critical value of d for $N = 10$ is 0.322; this means that in 20 per cent of random samples of size 10, the maximum absolute deviation between the sample cumulative distribution and the population cumulative distribution will be at least 0.322. The values in Table 1 for $N \leqq 35$ were computed by the procedure described in [5]; those for $N > 35$ are from Smirnov's table [10]. The values in the table are believed not to be in error by more than 4 units in the last figure shown for $N \leqq 20$, and by not more than 0.005 for $N = 25, 30, 35$.

3. APPLICATIONS

Our procedure is to draw the hypothetical cumulative distribution function on a graph and to draw curves a distance $d_\alpha(N)$ above and below the hypothetical curve (see Figure 1). If $S_N(x)$ passes outside of this band at any point we will reject, at the α level of significance, the hypothesis that the true distribution is $F_0(x)$. Thus, in the example shown the hypothetical curve is rejected. Only part of the observed distribution has been plotted in Figure 1; if it were plotted completely, it would, of course, rise to 1.0 on the vertical scale. Once the observed curve passes out of the acceptance band, the theoretical curve is re-

TABLE 1. Critical values, $d_\alpha(N)$, of the Maximum Absolute Difference between Sample and Population Cumulative Distributions.

Values of $d_\alpha(N)$ such that $Pr[\max|S_N(x) - F_0(x)| > d_\alpha(N)] = \alpha$, where $F_0(x)$ is the theoretical cumulative distribution and $S_N(x)$ is an observed cumulative distribution for a sample of N.

Sample size (N)	Level of significance (α)				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
over 35	$\frac{1.07}{\sqrt{N}}$	$\frac{1.14}{\sqrt{N}}$	$\frac{1.22}{\sqrt{N}}$	$\frac{1.36}{\sqrt{N}}$	$\frac{1.63}{\sqrt{N}}$

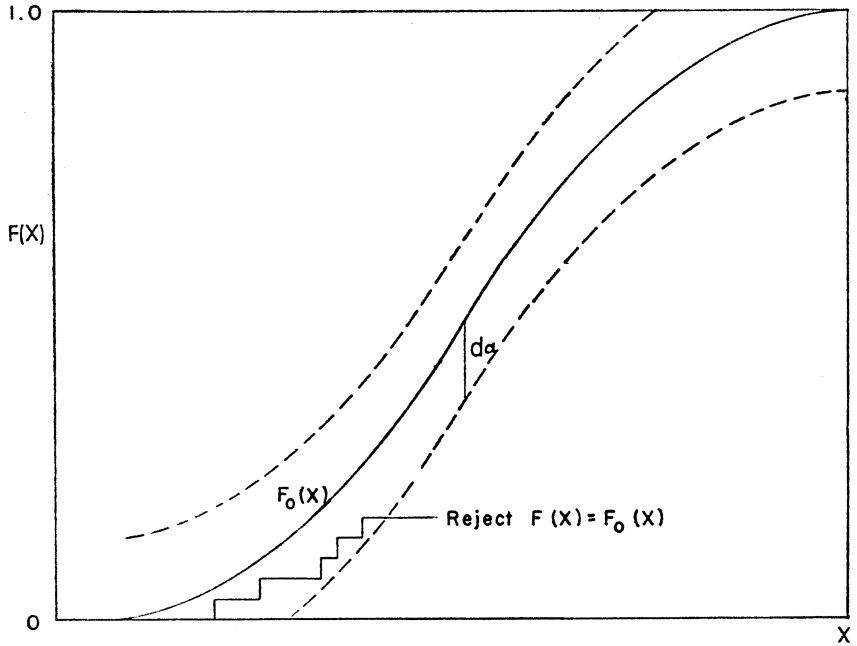

FIGURE 1. Graphical Method of Applying the d Test.

The continuous curve represents the theoretical distribution, and the broken curves are at distance $\pm d_\alpha(N)$ from it, $d_\alpha(N)$ being given by Table 1. The step-function represents part of the observed distribution. Reject unless the step-function lies entirely between the broken curves.

jected regardless of the later behavior of the observed curve. An alternative, and perhaps simpler, scheme is to record in a table the observed and hypothetical distributions and calculate the maximum deviation between them. If this exceeds $d_\alpha(N)$, we reject the hypothetical distribution.

As an example of the application of this test of goodness of fit we shall use data given by Snedecor ([11], p. 59). The results of a sampling experiment are compared with a theoretical normal distribution. His cumulative frequencies are recorded in Table 2.

The maximum deviation in the absolute frequencies, which occurs at the boundary score 30.5, is 12.41, which represents a difference in the proportions of $12.41/511 = 0.024$. The 5 per cent significance point is given in the last row of Table 1 as $1.36/\sqrt{511} = 0.060$. The observed value of d is less than the critical value, so we would accept, at the 5

TABLE 2. Comparison of Observed and Theoretical Frequencies in Sampling from a Normal Population
(Snedecor [11], p. 59)

Upper boundary of class	Cumulative frequency to upper boundary of class		
	Observed	Theoretical	Absolute difference
39.5	511	511.00	0
38.5	510	509.16	0.84
37.5	510	506.45	3.55
36.5	505	500.83	4.17
35.5	493	490.05	2.95
34.5	469	471.45	2.45
33.5	447	442.43	4.57
32.5	402	401.29	0.71
31.5	356	349.22	6.78
30.5	300	287.59	12.41*
29.5	228	223.36	4.64
28.5	162	161.73	0.27
27.5	114	109.66	4.34
26.5	73	68.52	4.48
25.5	43	39.50	3.50
24.5	24	20.90	3.10
23.5	14	10.12	3.88
22.5	9	4.50	4.50
21.5	2	1.79	0.21
20.5	2	0.61	1.39
19.5	1	0.20	0.80

* Maximum absolute difference = 12.41.
Hence $d = 12.41/511 = 0.024$.

per cent level of significance, the hypothesis that the population distribution is that recorded in Table 2.

Grouping observations into intervals tends to lower the value of d. For grouped data, therefore, the appropriate significance levels are smaller than those tabled. For large samples, grouping usually will cause little change in the appropriate significance levels. However, grouping into a very small number of categories can cause important changes for any sample size.

As another application, consider testing normality by observing whether or not the sample cumulative distribution drawn on arithmetic probability paper is approximately straight. There are no theoretical results, at present, which indicate how close to straight the

observed sample cumulative curve should be. The d test is correctly used only if the distribution is completely specified (i.e., not only as normal, but as normal with a specified mean and a specified standard deviation). The distribution of the maximum deviation is not known when certain parameters of the population have been estimated from the sample. It may be expected, however, that the effect of adjusting the population mean and standard deviation to those of the sample, either by calculation or by visually fitting a straight line on normal probability paper, will be to reduce the critical level of d. If the value of $d_\alpha(N)$ shown in Table 1 is exceeded in these circumstances, we may safely conclude that the discrepancy is significant, i.e., that the distribution is not normal.[1]

4. POWER OF THE TEST

Suppose we indicate by $F_1(x)$ an alternative form of the distribution function. Let Δ be the maximum absolute difference between $F_1(x)$ and $F_0(x)$. This measurement of distance between alternatives has been used by Mann and Wald [4] and by Williams [12].

For large samples it has been shown [6] that the power of the d test (i.e., the probability of rejecting the hypothetical distribution) is never less than

$$1 - (2\pi)^{-1/2} \int_{2[\Delta\sqrt{N}-d_\alpha(N)]}^{2[\Delta\sqrt{N}+d_\alpha(N)]} \exp(-t^2/2)dt. \tag{1}$$

Since this is a poor lower bound to the power, the actual power is likely to be much larger. As is shown in Section 5, however, it is of value in comparing the d test with the χ^2 test. Figure 2 shows this lower bound for the 5 and 1 per cent levels of significance.

Figure 2 can be used to indicate the sample size necessary so that, at the 5 per cent level of significance, the d test of $F_0(x)$ has power at least 0.50 against the alternative $F_1(x)$. Suppose that the maximum absolute difference between $F_0(x)$ and $F_1(x)$ is 0.2. Reading across from 0.50 on the vertical scale we see that the $\alpha = 0.05$ curve is intersected

[1] A sampling experiment was conducted by the writer in which 100 samples of size 10 were drawn from a known normal distribution. The cumulative distribution for each sample was plotted on arithmetic normal probability paper and a straight line was fitted by eye. The observed percentiles of the distribution of d were considerably lower than those given by Table 1. The 95th percentile was 0.29 as compared with 0.41 in Table 1, and the 90th percentile was 0.25 as compared with 0.37. This implies that deviations greater than those in Table 1 should, in these applications, be treated as very strong indication of departure from normality. In the sampling experiment only 1 observation of the 100 exceeded the 20 per cent critical value from Table 1 and no observation exceeded the 10 per cent critical value.

above $\Delta\sqrt{N}=1.36$. Solving this for N we find $N=(1.36/0.2)^2=46.24$. A sample of 47 would, therefore, be required.

As another example, suppose we have a sample of 400 and that $\Delta=0.10$. If we test the hypothetical distribution $F_0(x)$ at the 1 per cent level of significance we can determine a lower bound to the chance of rejecting it if the true distribution is $F_1(x)$. Here $\Delta\sqrt{N}=0.10\times 20=2$,

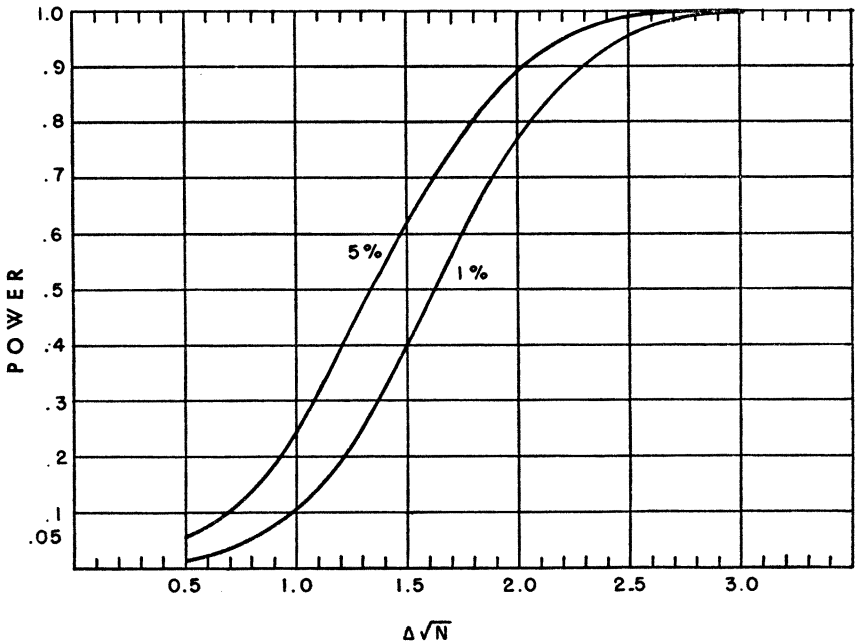

FIGURE 2. Lower bounds for the power of the d test for $\alpha=0.01$ and $\alpha=0.05$.

and above 2 on the horizontal scale we see that the 1 per cent curve is at a height of 0.77. The power of this test is, thus, at least 0.77 against the particular alternative $F_1(x)$; i.e., if $F_1(x)$ is correct, we have at least a 77 per cent chance of detecting that $F_0(x)$ is incorrect.

5. COMPARISON OF THE d TEST WITH THE χ^2 TEST

Mann and Wald [4] have given a technique for deciding on an optimum number of class intervals for the application of the χ^2 test for goodness of fit. The intervals and sample size are so chosen that the probability of rejecting $F_0(x)$ as the true distribution, if $F_1(x)$ is actually the true distribution, is never less than 0.5. More important for our purposes is the fact that there will be one alternative distribution,

$F_2(x)$, at a distance Δ from $F_0(x)$, such that for the χ^2 test the probability of rejecting $F_0(x)$, if $F_2(x)$ is the true distribution function, is as close as desired to 0.5.

Williams [12] has presented a table showing, for various sample sizes, minimum distances for which the power of the χ^2 test is not less than 0.5. Part of his table is reproduced in Table 3, together with minimum distances for which the d test has power not less than 0.5. The discrepancies detectable by the d test are all smaller than those for the χ^2 test. This implies that the d test, at least at the 50 per cent power level, will detect smaller deviations in cumulative distributions than will the χ^2 test.

TABLE 3. Minimum Deviation of Actual from Assumed Population that is Detectable with Probability 0.50 by the χ^2 and d Tests at the 5 per cent and 1 per cent Levels of Significance*

	$\alpha = .05$		$\alpha = .01$	
N	χ^2	d test	χ^2 test	d test
200	0.1605	0.096	0.1847	0.115
250	0.1469	0.086	0.1657	0.103
300	0.1343	0.079	0.1577	0.094
350	0.1284	0.073	0.1479	0.087
400	0.1213	0.068	0.1369	0.082
450	0.1157	0.064	0.1315	0.077
500	0.1112	0.061	0.1273	0.073
550	0.1052	0.058	0.1209	0.070
600	0.1024	0.055	0.1184	0.067
650	0.1000	0.053	0.1137	0.064
700	0.0961	0.051	0.1120	0.062
750	0.0945	0.050	0.1083	0.060
800	0.0914	0.048	0.1051	0.058
850	0.0887	0.047	0.1022	0.056
900	0.0877	0.045	0.0997	0.054
950	0.0855	0.044	0.0974	0.053
1000	0.0834	0.043	0.0953	0.052
1100	0.0812	0.041	0.0918	0.049
1200	0.0782	0.039	0.0888	0.047
1300	0.0757	0.038	0.0862	0.045
1400	0.0734	0.036	0.0841	0.044
1500	0.0715	0.035	0.0823	0.042
2000	0.0629	0.030	0.0728	0.036

* The deviation between two populations is measured by the maximum absolute difference between their cumulative distributions. The values for the χ^2 test are taken from [12]; those for the d test are computed from formula (1).

Other points of comparisons between the χ^2 and d tests may be noted:

(*i*) In general, the power of the χ^2 test is not known (Mann and Wald [4] considered only the case where it is 0.5), whereas a lower bound to the power of the d test for any alternative can be read from Figure 2.

(*ii*) The d test treats individual observations separately and thus does not lose information by grouping, as the χ^2 test necessarily does. In small samples this loss of information in χ^2 procedures is large, since wide class intervals must be used; and for very small samples χ^2 is not applicable at all. This, together with the information in Table 3, suggests that the d test may be always more powerful than χ^2 tests.

(*iii*) d will usually require less computation than χ^2. This is especially true when a graphical test is used, as illustrated in Figure 1, for if the hypothesis is rejected the computation stops at the point of rejection. Graphing might be convenient if a standard hypothesis is tested repeatedly, since a master test chart could be prepared. There are also instances where individuals can be ranked easily according to size, and then the individuals can be measured one at a time starting with the smallest. After each individual is measured, the cumulative distribution can be checked to see if d exceeds $d_\alpha(N)$. Using this sequential procedure it might be possible to avoid the actual measurement of many of the individuals. This might be especially useful if the ranking technique were fast and inexpensive while actually measuring was slow and expensive.

(*iv*) In cases where parameters must be estimated from the sample the χ^2 test is easily modified by reducing the number of degrees of freedom. The d test has no such known modifications so, except for the remarks in Section 3, is not applicable in such cases.

(*v*) As yet the d test cannot be applied to discrete populations, whereas the χ^2 can be.

6. CONFIDENCE LIMITS FOR THE TRUE CUMULATIVE DISTRIBUTION FUNCTION[2]

Table 1 can be used to find confidence limits for the true cumulative distribution function, say $F(x)$. Thus 100 $(1-\alpha)$ per cent confidence limits for $F(x)$ are

$$S_N(x) - d_\alpha(N) < F(x) < S_N(x) + d_\alpha(N).$$

[2] See reference [13].

For example, for a sample of size 100, one can be 95 per cent sure that $S_N(x)$ will stay within $1.36/\sqrt{100}=0.136$ of the true distribution.

As another example, suppose it is desired to perform a large scale ("Monte Carlo" method [7]) sampling experiment to study some distribution. To be 99 per cent sure of estimating the cumulative sampling distribution within, say, 2 percentage points for the entire curve,

TABLE 4. Confidence Limits for the Cumulative Distribution Function Shown in Table 2

Upper Boundary of Class	Lower Confidence	Observed Cumulative Proportion	Upper Confidence Limit
39.5	0.940	1.000	1.000
38.5	0.938	0.998	1.000
37.5	0.938	0.998	1.000
36.5	0.928	0.988	1.000
35.5	0.905	0.965	1.000
34.5	0.858	0.918	0.978
33.5	0.815	0.875	0.935
32.5	0.727	0.787	0.847
31.5	0.637	0.697	0.757
30.5	0.527	0.587	0.647
29.5	0.386	0.446	0.506
28.5	0.257	0.317	0.377
27.5	0.163	0.223	0.283
26.5	0.083	0.143	0.203
25.5	0.024	0.084	0.144
24.5	0	0.047	0.107
23.5	0	0.027	0.087
22.5	0	0.018	0.078
21.5	0	0.004	0.064
20.5	0	0.004	0.064
19.5	0	0.002	0.062
18.5	0	0	0.060

the necessary sample size is found as follows: From Table 1, we find $d_{0.01}=1.63/\sqrt{N}=0.02$. Hence $\sqrt{N}=81.5$ and $N=6643$.

As a final example, consider the data in Table 2. Ninety-five per cent confidence limits for the true distribution curve are obtained by adding and subtracting $1.36/\sqrt{511}=0.060$ from the observed distribution as shown in Table 4. Note that the theoretical cumulative distribution, recorded in Table 2, is completely inside the limits and thus would be accepted at the 5 per cent level of significance.

7. REFERENCES

[1] Doob, J. L., "Heuristic Approach to the Kolmogorov-Smirnov Theorems." *Annals of Mathematical Statistics*, 20 (1949), 393–403.

[2] Feller, W., "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions," *Annals of Mathematical Statistics*, 19 (1948), 177–189.

[3] Kolmogorov, A., "Sulla Determinazione Empirica di una Legge di Distributione," *Giornale dell'Istituto Italiano degli Attuari*, 4 (1933), 1–11.

[4] Mann, H. B. and Wald, A., "On the Choice of the Number of Intervals in the Application of the Chi-square Test," *Annals of Mathematical Statistics*, 13 (1942), 306–317.

[5] Massey, Frank J., Jr., "A Note on the Estimation of a Cumulative Distribution Function by Confidence Intervals," *Annals of Mathematical Statistics*, 21 (1950), 116–119.

[6] Massey, F. J., Jr., "A Note on the Power of a Non-Parametric Test," *Annals of Mathematical Statistics*, 21 (1950), 440–443.

[7] Metropolis, N. and Ulam, S., "The Monte Carlo Method," *Journal of American Statistical Association*, 44 (1949), 335–342.

[8] Scheffé, H., "Statistical Inference in the Non-Parametric Case," *Annals of Mathematical Statistics*, 14 (1943), 305–332.

[9] Smirnov, H., "Sur les Écarts de la Courbe de Distribution Empirique," *Recueil Mathématique* (*Matematiceskii Sbornik*), N.S. 6 (1939), 3–26.

[10] Smirnov, N., "Table for Estimating the Goodness of Fit of Empirical Distributions," *Annals of Mathematical Statistics*, 19 (1948), 279–281.

[11] Snedecor, George W., *Statistical Methods*, Fourth Edition, Ames, Iowa (Iowa State College Press) 1946.

[12] Williams, C. A., Jr., "On the Choice of the Number and Width of Classes for the Chi-square Test for Goodness of Fit," *Journal of the American Statistical Association*, 45 (1950), 77–86.

[13] Wald, A. and Wolfowitz, J., "Confidence Limits for Continuous Distribution Functions," *Annals of Mathematical Statistics*, 10 (1939), 105–118.

[14] Wolfowitz, J., "Non-parametric Statistical Inference," *Proceedings of the Berkeley Symposium of Mathematical Statistics and Probability*. Berkeley (University of California Press) 1949, 93–113.