J. Statist. Comput. Simul., 1978, Vol. 6, pp. 257–263

Printed in Great Britain

An Efficient Approximate Algorithm for the Kolmogorov–Smirnov and Lilliefors Tests†

TEOFILO GONZALES, SARTAJ SAHNI and W. R. FRANTA
Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, U.S.A.

(*Received October 28, 1976*)

In an earlier paper we presented a linear time algorithm for computing the Kolmogorov–Smirnov and Lilliefors test statistics. In this paper we present a linear time approximate algorithm which requires less memory than the previous algorithm.

KEYWORDS and PHRASES: Kolmogorov–Smirnov test, Lilliefors test, exact and approximate algorithms, time and space complexity.

CR Categories: 5.25, 5.5

1. INTRODUCTION

The Kolmogorov–Smirnov and Lilliefors tests allow us to evaluate the hypothesis that a collected data set, i.e., a random sample $X_1,\ldots,X_n$, was drawn from a specified continuous distribution function $F(x)$. For both tests, a determination is made of the numeric difference between the specified distribution function $F(X)$ and the sample distribution function (X) defined as:

$$S(X)=j/n, j=\{\text{number of points} \leqq X\}. \tag{1.1}$$

If the sample, $X_1,\ldots,X_n$, has been sorted into nondecreasing order so that $X_1 \leqq X_2 \leqq \ldots \leqq X_n$, then the Kolmogorov–Smirnov statistics $K^+_{\max}$ (maximum positive) $K^-_{\max}$ (maximum negative) and $K_{\max}$ (maximum absolute) deviations

†This research was supported in part by NSF grant DCR 74-10081.

are computed by the formulas:

$$K^{+}_{\max} = \sqrt{n} \max_{1 \leqq j \leqq n} \left\{ \frac{j}{n} - F(X_j) \right\}$$

$$K^{-}_{\max} = \sqrt{n} \max_{1 \leqq j \leqq n} \left\{ F(X_j) - \frac{j-1}{n} \right\} \tag{1.2}$$

$$K_{\max} = \max \{K^{+}_{\max}, K^{-}_{\max}\}$$

The distribution functions of $K^{+}_{\max}$, $K^{-}_{\max}$ and $K_{\max}$ are known and tabulated. For certain $F(X)$ (see Lilliefors, 1967, 1969, Stephens, 1974), tabulated values of the test statistic distributions are available for the case where the actual parameters of $F(X)$ have been replaced by estimates computed from the sample. The test also has application for certain spectral tests, see for example, [2, p. 197].

Previous algorithms (see Knuth, 1969, Lindgren, 1962, Miller and Freund, 1965) for computing these test statistics are essentially identical to algorithm K below:

Algorithm K $(K^{+}_{\max}, K^{-}_{\max}, K_{\max})$
//Knuth's algorithm for Kolomogorov–Smirnov test statistics [4] pp. 44//

Step 1 obtain the n observations $X_1, X_2, \ldots, X_n$

Step 2 sort them so that $X_1 \leqq X_2 \ldots \leqq X_n$

Step 3 compute $K^{+}_{\max}, K^{-}_{\max}$ and $K_{\max}$ using equation 1.2.

end K

Since step 2 sorts the observations, it requires $0(n \log n)$ time. The remainder of the algorithm takes $0(n)$ time (assuming $F(X)$ may be computed in a constant amount of time 0(1)). Hence, the total time required is $0(n \log n)$. The algorithm presented in Gonzalez, Sahni and Franta, (1977) computes the test statistics $K^{+}_{\max}$, $K^{-}_{\max}$ and $K_{\max}$ without explicitly sorting the X_i's and thus has a time complexity of $0(n)$. The tabulated acceptance/rejection values of these statistics are usually accurate only to three or four decimal places. Hence, there seems little point in computing these statistics to greater precision than the tabulated values. With this in mind, we present here an approximation algorithm which guarantees a certain closeness to the exact values of $K^{+}_{\max}$, $K^{-}_{\max}$ and $K_{\max}$. This approximate algorithm requires less storage space than the exact algorithm and so should be useful when n is large. The computing time is still $0(n)$. Empirical tests, Section 3, show that the approximation algorithm is actually slightly faster than the exact algorithm. The desired closeness of the approximate and exact solutions can be fixed through an algorithm parameter.

Both the exact and approximate algorithms apply equally well to the Lilliefors test (Conover, 1971) which is a modification of the Kolmogorov–Smirnov test.

2. APPROXIMATION ALGORITHM

In this section we present an algorithm to determine approximately the values of $K^+_{\max}$, $K^+_{\max}$ and $K_{\max}$. This algorithm is slightly faster than the exact algorithm of Gonzales, Sahni, and Franta, (1977) and requires at most 1/3 the space required by that algorithm and is very similar to the algorithm presented in Gonzales, Sahni and Franta (1977).

The algorithm divides the range of the cumulative distribution function, $F(X)$, into $m+1$ intervals. An arbitrary point y, $0 \leqq y \leqq 1$ lies in the interval $[y^*m]$.† For each sample point x_i, $i=1,\ldots,n$ we compute $F(x_i)$ and determine the interval into which $F(x_i)$ falls. A counter variable, NUM_i, for that interval is then incremented. On the basis of NUM_i, together with knowledge of the maximum, MAX_i, and minimum, MIN_i, values to fall in each interval we can determine $K^+_{\max}$, $K^-_{\max}$, and $K_{\max}$. For the approximation algorithm we approximate the maximum and minimum values by setting $\mathrm{MAX}_i \approx \mathrm{MIN}_i \approx (i-0.5/m)$, where $m+1$ is the number of intervals (bins) used as specified in the formal statement of the algorithm given next.

Algorithm APPROX_$KS(n, \hat{K}^+_{\max}, \hat{K}^-_{\max}, \hat{K}_{\max}, m)$

//Find approximations to $K^+_{\max}, K^-_{\max}, K_{\max}$ using only $m+1$ *bins*.//

Step 1 //initialize bins//
for $i \leftarrow 0$ *to* m *do*
$\mathrm{NUM}_i \leftarrow 0$
end

Step 2 //input and count number of sample points in each bin//
for $i \leftarrow 1$ *to* n *do*
input X
$f \leftarrow F(X); j \leftarrow \lceil f^*m \rceil$; //compute bin for X//
$\mathrm{NUM}_j \leftarrow \mathrm{NUM}_j + 1$
end

---

†The notation $\lceil x \rceil$ denotes the ceiling or least integer function of x, that is, the minimum k such that $k \geqq x$.

Step 3 //process each bin finding approximate values for maximum positive and negative deviates from $F(X)$//

$DN \leftarrow 0$;

$DP \leftarrow \text{NUM}_0/n$

$j \leftarrow \text{NUM}_0$

for $i \leftarrow 1$ *to* m *do*

if $\text{NUM}_i > 0$ then $[z \leftarrow (i-.5)/m - j/n$

if $z > DN$ *then* $[DN \leftarrow z]$

$j \leftarrow j + \text{NUM}_i$

$z \leftarrow j/n - (i-.5)/m$

if $z > DP$ *then* $[DP \leftarrow z]$

]

end

Step 4 //Compute $\hat{K}^+_{\max}, \hat{K}^-_{\max}, \hat{K}_{\max}$//

$K^+_{\max} \leftarrow \sqrt{n} * DP$

$K^-_{\max} \leftarrow \sqrt{n} * DN$

$K_{\max} \leftarrow \max\{\hat{K}^+_{\max}, \hat{K}^-_{\max}\}$

return

end of algorithm APPROX_KS

THEOREM 2.1 *The following relations hold between the approximate values* $\hat{K}^+_{\max}, \hat{K}^-_{\max}$ *and* $\hat{K}_{\max}$ *as given by algorithm APPROX_KS and the exact values* $K^+_{\max}$, $K^-_{\max}$ *and* $K_{\max}$ *given by algorithm of Gonzales, Sahni and Franta* (1977):

$$(i) \quad |\hat{k}^+_{\max} - K^+_{\max}| \leqq \sqrt{n}/(2m)$$

$$(ii) \quad |\hat{K}^-_{\max} - K^-_{\max}| \leqq \sqrt{n}/(2m)$$

and

$$(iii) \quad |\hat{K}_{\max} - K_{\max}| \leqq \sqrt{n}/(2m)$$

Proof (i) follows from the observation that for any bin, i, if $j = \Sigma_{0 \leqq l \leqq i} \text{NUM}_l$; if X is a sample point such that $\lceil F(X) * m \rceil = i$ and if there are

k sample points $\leqq X$, then

$$\begin{aligned}&\sqrt{n}\,(k/n - F(X)) - -\sqrt{n}\,(j/n) - (i-0.5)/m)\\ &= \sqrt{n}\,((k-j)/n + (i-.5)/m - F(X))\\ &\leqq \sqrt{n}/(2m).\end{aligned}$$

This, together with (1.2), proves (i). The proofs for (ii) and (iii) are similar. Theorem 2.1 dictates that k digits of accuracy requires $n > m \geqq (10^k/2)\sqrt{n}$, which requires that $\sqrt{n} \geqq 10^k/2$ or $n \geqq 10^{2k}/4$.

(It has been suggested, but not verified, that using the formulas

$$\mathrm{MAX}_i = (1/m) * (i - 1 + (i * \mathrm{NUM}_i)/(\mathrm{NUM}_i + 1)$$

$$\mathrm{MIN}_i = (1/m) * ((i-1) * \mathrm{NUM}_i + i)/(\mathrm{NUM}_i + 1)$$

to estimate MAX_i and MIN_i from NUM_i and then using Step 3 of the algorithm of Gonzales, Sahni and Franta (1977) would, in practice, yield a better approximation.)

LEMMA 2.1 *The computing time for algorithm APPROX_KS is* $0(n)$, *and the space required is* $m + c$ *for* $m \leqq n$ *and* c *a constant.*

Proof Obvious.

For comparison, note that the space requirement for the exact algorithm given in Gonzales, Sahni and Franta (1977) is $3n + c$.

3. EXPERIMENTAL RESULTS

In order to determine the relative performance of the algorithm on practical sample sizes, we programmed the algorithms of Gonzales, Sahni and Franta (1977) and APPROX_KS and algorithm K in FORTRAN and ran several experiments on a Cyber 74. The sorting method used for algorithm K was heapsort. Three distribution functions: normal, exponential and uniform were tried so as to reflect the differences in the computing times for $F(\ )$. Table I presents the results obtained for various sample sizes. The times are the mean computing times over several experiments. As can be seen from this Table, algorithm K required from about 2 to 3 times the time required by our algorithms. This difference will, of course, become larger for larger sample sizes. Algorithm APPROX_KS took roughly the same time as the algorithm of Gonzales, Sahni and Franta (1977) but used considerably less storage. The

TABLE I

Mean execution times for the Kolmogorov–Smirnov test

Distribution	Sample Size, n	Number of Experiments	Computing Time									
			Algorithm of Gonzales, Sahni, Franta (1977)				APPROX KS					
			K				$m=n$		$m=n/4$		$m=n/16$	
			Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Exponential	100	50	17.8	1.0	9.0	0.9	7.9	0.8	6.8	0.8	6.4	0.5
	500	30	107	2.7	44.3	1.8	35.7	2.0	33	1.5	31.2	1.2
	1000	20	231	2	88	2	78	2	66	2.2	63	2
	5000	5	1342	8.9	439	5.9	387	4.9	329	6.2	308	2.4
Uniform	100	50	15.9	0.8	7.2	0.7	5.8	0.7	4.6	0.7	4.4	0.8
	500	30	97.9	2.5	35.2	1.6	29.1	1.7	23.4	1.1	21.5	1.1
	1000	20	213.5	4.6	71.8	2.9	59.9	2.6	46.9	1.8	43.9	2.2
	5000	5	1260	5	352	8	294	4.4	234	4	213	4.6
Normal	100	50	25.9	1.3	17	1	16	0.9	15.2	0.8	14.8	0.9
	500	30	147	3.7	84	2.9	80	2.9	75	2.3	72	2.5
	1000	20	308	10.3	168	5.7	158	4.9	149	5.5	145	4.8
	5000	5	1760	31	838	20	801	4.5	738	4.6	731	6.9

Times in milliseconds

observed difference between the exact and approximate values of the test statistics was about half the theoretical bound of Theorem 2.1.

4. CONCLUSIONS

We have presented an approximate linear time algorithm for the Kolmogorov–Smirnov and Lilliefors tests. For algorithm APPROX_KS as well as the algorithm of Gonzales, Sahni and Franta (1977), we note that the speed up is obtained by avoiding a sort of the sample. Algorithm APPROX_KS is thus recommended in cases where n is large, available storage small, the sample is unsorted, and the acceptance/rejection values of K^+_{max}, K^-_{max} and K_{max} are themselves known only approximately (i.e., only a few digits of significance is desired). The value of m to use can be determined using Theorem 2.1.

References

Conover, W. J. (1971), *Practical Nonparametric Statistics*, John Wiley and Sons, Inc.

Fishman, G. S. (1973), *Concepts and Methods in Discrete Event Digital Simulation*, John Wiley and Sons.

Gonzales, T., Sahni S. and Franta, W. R. (1977), "An efficient algorithm for the Kolmogorov–Smirnov and Lilliefors tests", *ACM Transactions on Mathematical Software*, Vol. 3, No. 1, pp. 60–64.

Knuth, D. E. (1969), *Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, Addison-Wesley.

Lilliefors, H. W. (1967), "On the Kolmogorov–Smirnov test for normality with mean and variance unknown", *JASA*, **62,** 339–402.

Lilliefors, H. W. (1969), "On the Kolmogorov–Smirnov test for the exponential distribution with mean unknown," *JASA*, **64,** 387–389.

Lindgren, B. W. (1962), *Statistical Theory*, The MacMillan Company, New York.

Miller, I. and Freund, J. E. (1965), *Probability and Statistics for Engineers*," Prentice-Hall.

Stephens, M. A. (1974), "EDF statistics for goodness of fit and some comparisons," *JASA*, **69,** 347, 730–737.