An Efficient Algorithm for the Kolmogorov-Smirnov and Lilliefors Tests

TEOFILO GONZALEZ, SARTAJ SAHNI, and W. R. FRANTA
University of Minnesota

A linear time algorithm for computing the Kolmogorov-Smirnov and Lilliefors test statistics is presented.

Key Words and Phrases: Kolmogorov-Smirnov test, Lilliefors test, 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 $S(X)$ defined as

$$S(X) = j/n, \quad j = \text{number of points less than or equal to } x. \qquad (1)$$

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

$$\begin{aligned} K^+_{\max} &= n^{1/2} \max_{1 \leq j \leq n} \{j/n - F(X_j)\} \\ K^-_{\max} &= n^{1/2} \max_{1 \leq j \leq n} \{F(X_j) - (j-1)/n\} \qquad (2) \\ K_{\max} &= \max \{K^+_{\max}, K^-_{\max}\}. \end{aligned}$$

The distribution functions of $K^+_{\max}$, $K^-_{\max}$, and $K_{\max}$ are known and tabulated. We cannot reject the null hypothesis that the sample was indeed drawn from the distribution $F(X)$ if the statistics computed do not exceed the critical values tabulated for the level of significance selected. For certain $F(X)$ (see [4, 5]), tabulated

This research was supported in part by the National Science Foundation under Grant DCR 74-10081.

Authors' address: Department of Computer, Information, and Control Sciences, 114 Main Engineering Building, Minneapolis, MN 55455.

values of the test statistic distributions are available for the case where the actual parameters of $F(X)$ have been replaced by estimates computed from the sample.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 ([3, 6, and 7]) for computing these test statistics are essentially identical to algorithm K below:

```
ALGORITHM K($K^+_{max}$, $K^-_{max}$, $K_{max}$)

Knuth's algorithm for Kolmogorov-Smirnov test statistics [3, pp. 44].

Step 1. Obtain the $n$ observations $X_1, X_2, \ldots, X_n$.
Step 2. Sort them so that $X_1 \le X_2 \le \cdots \le X_n$.
Step 3. Compute $K^+_{max}$, $K^-_{max}$, and $K_{max}$ by using eq. (2).
end K
```

Since step 2 sorts the observations, it requires $O(n \log n)$ time. The remainder of the algorithm takes $O(n)$ time (assuming $F(X)$ may be computed in a constant amount of time $O(1)$). Hence the total time required is $O(n \log n)$. The algorithm we present in Section 2 computes the test statistics K^+_{max}, K^-_{max}, and K_{max} without explicitly sorting the X_i. This algorithm has a time complexity of $O(n)$. It applies equally well to the Lilliefors test [1], which is very similar to the Kolmogorov-Smirnov test. In this test, the X_i in eq. (2) are replaced by $(X_i - \bar{X})/s$, where $\bar{X}$ and s^2 are the usual unbiased estimates of the mean and variance of X.

2. A LINEAR TIME ALGORITHM FOR THE KOLMOGOROV-SMIRNOV TEST

Our algorithm for computing the values of K^+_{max}, K^-_{max}, and K_{max} for the Kolmogorov-Smirnov test proceeds by dividing the range of the cumulative distribution function $F(X)$ into $n + 1$ intervals. The point y, $0 \le y \le 1$, lies in the interval $\lceil y*n \rceil$. For each of the n sample points $1 \le i \le n$, the value of $F(X_i)$ is computed. For each of the $n + 1$ intervals, the number of sample points for which $F(X)$ is in that interval is recorded, together with the minimum and maximum values of $F(X)$ achieved in that interval. Theorem 2.1 below shows that this information is sufficient to enable an accurate determination of the values of K^+_{max}, K^-_{max}, and K_{max}. We first formally present the algorithm. Lemma 2.2 below analyzes the time and space complexity of this algorithm.

```
ALGORITHM KS($n$, $K^+_{max}$, $K^-_{max}$, $K_{max}$)

This algorithm inputs $n$ sample points and performs the Kolmogorov-Smirnov test against the
cumulative distribution function $F(X)$. The outputs of the algorithm are:

                    $K^+_{max}$ :  the $K^+$ maximum deviate
                    $K^-_{max}$ :  the $K^-$ maximum deviate
                    $K_{max}$ :  the absolute maximum deviate.

Three vectors of size $n + 1$ each are made use of:
               $NUM_i$ :  number of samples in bin $i$   \
               $MAX_i$ :  maximum sample value in bin $i$  >  $0 \le i \le n$.
               $MIN_i$ :  minimum sample value in bin $i$   /
Step 1  [Initialize]
        for $i \leftarrow 0$ to $n$ do
            $XMIN_i \leftarrow 1$
            $XMAX_i \leftarrow 0$
            $NUM_i \leftarrow 0$
        end
Step 2  [Input observations and put into bins]
        for $i \leftarrow 1$ to $n$ do
            input $X$
            $f \leftarrow F(X)$
            $j \leftarrow \lceil f*n \rceil$  //compute bin for $X$//
            $NUM_j \leftarrow NUM_j + 1$
            if $MAX_j < f$ then $[MAX_j \leftarrow f]$
            if $MIN_j > f$ then $[MIN_j \leftarrow f]$
        end
Step 3  [Process each bin; find maximum positive and negative deviates]
        $j \leftarrow 0$;  $DP \leftarrow 0$;  $DN \leftarrow 0$;
        for $i \leftarrow 0$ to $n$ do
        if $NUM_i > 0$ then [$z \leftarrow MIN_i - j/n$
                             if $z > DN$ then $[DN \leftarrow z]$
                             $j \leftarrow j + NUM_i$
                             $z \leftarrow j/n - MAX_i$ ,
                             if $z > DP$ then $[DP \leftarrow z]$
                             ]
        end
Step 4  [Compute $K^+_{\max}$ , $K^-_{\max}$ , and $K_{\max}$]
        $K^+_{\max} \leftarrow n^{1/2}*DP$
        $K^-_{\max} \leftarrow n^{1/2}*DN$
        $K_{\max} \leftarrow \max\ \{K^+_{\max} , K^-_{\max}\}$
        return
        end of algorithm KS
```

We now prove that the above algorithm does in fact give the correct results. First we state a lemma that will be useful in proving the correctness of our algorithm KS.

LEMMA 2.1. *Let $F(x)$ be an increasing function over the interval* $[a, b]$, *and let $x_1 \leq \cdots \leq x_k$ be k points in the interval. Suppose $F(b) - F(a) \leq 1/n$ for arbitrary n. Then, for any c and c',*

$$\max_{1 \leq i \leq k} ((i + c)/n - F(x_i)) = (k + c)/n - F(x_k)$$

$$\max_{1 \leq i \leq k} (F(x_i) - (i + c')/n) = (1 + c')/n - F(x_1).$$

PROOF. Obvious.

THEOREM 2.1. *Algorithm KS gives the correct values for $K^+_{\max}$, $K^-_{\max}$, and $K_{\max}$.*

PROOF. The proof is in two parts. First, we show that it is sufficient to consider only the smallest and largest sample points in each bin and then that algorithm KS determines accurately the index of these sample points if all sample points were sorted into nondecreasing order.

(i) Since $F(\ )$ is a cumulative distribution function, it must increase monotonically, i.e. $x > y$ iff $F(x) > F(y)$. Also, by construction, the maximum deviation of $F(\ )$ in any bin is $1/n$. Hence Lemma 2.1 applies and it is sufficient to consider only the maximum and minimum $F(\ )$ in each bin.

(ii) Again, since $F(\ )$ increases monotonically, all sample points in bin l are less than all sample points in bin $l + 1$, $1 \leq l < n$. Therefore if X is the small-

Table I. Mean Execution Times for the Kolmogorov-Smirnov Test

			Computing time (msec)				
			K		KS		
Distribution	Sample size, n	Number of experiments	mean	standard deviation	mean	standard deviation	K/KS
Exponential	100	50	17.8	1.0	9.0	.9	1.98
	500	30	107	2.7	44.3	1.8	2.42
	1000	20	231	2	88	2	2.63
	5000	5	1342	8.9	439	5.9	3.06
Uniform	100	50	15.9	.8	7.2	.7	2.21
	500	30	97.9	2.5	35.2	1.6	2.78
	1000	20	213.5	4.6	71.8	2.9	2.97
	5000	5	1260	5	352	8	3.58
Normal	100	50	25.9	1.3	17	1	1.52
	500	30	147	3.7	84	2.9	1.75
	1000	20	308	10.3	168	5.7	1.83
	5000	5	1760	31	838	20	2.1

est and Z the largest sample point in bin l, then the number of sample points less then X is $\sum_{0\le i<l} NUM_i$ and the number greater than or equal to Z is $\sum_{0\le i\le l} NUM_i$.

(i) and (ii) show that the correct values for $K^{+}_{\max}$ and $K^{-}_{\max}$ are obtained. By definition of $K_{\max}$, the correct $K_{\max}$ is also obtained.

LEMMA 2.1. *The time complexity of algorithm KS is $O(n)$. The space required is $3n + c$ for some constant c.*

PROOF. Obvious.

3. EMPIRICAL TESTS

In order to determine the relative performance of our algorithm on practical sample sizes, we programmed algorithms KS and K in Fortran and ran several tests on the 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 two to three times the time required by our algorithm KS. This difference will, of course, become larger for larger sample sizes.

4. CONCLUSIONS

We have presented a linear time algorithm for the Kolmogorov-Smirnov and Lilliefors tests. While this algorithm is faster than those of [1, 3, 6, and 7], one should note that this speedup is obtained by avoiding a sort of the sample. If the sample is already known to be sorted or has to be sorted for some other reason, then the values of $K^{+}_{\max}$, $K^{-}_{\max}$, and $K_{\max}$ can be computed more efficiently by a direct application of eq. (2). Thus we recommend the use of algorithm KS when the sample is neither sorted to begin with nor has to be sorted for some other purpose.

REFERENCES

1. CONOVER, W.J. *Practical Nonparametric Statistics*. Wiley, New York, 1971.
2. FISHMAN, G.S. *Concepts and Methods in Discrete Event Digital Simulation*. Wıley, New York, 1973.
3. KNUTH, D.E. *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*. Addison-Wesley, Reading, Mass., 1969.
4. LILLIEFORS, H.W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. ASA 62* (1967), 339–402.
5. LILLIEFORS, H.W. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *J. ASA 64* (1969), 387–389.
6. LINDGREN, B.W. *Statistical Theory*. Macmillan, New York, 1962.
7. MILLER, I. AND FREUND, J.E. *Probability and Statistics for Engineers*. Prentice-Hall, Englewood Cliffs, N.J., 1965.

Received October 1975