

An Efficient Approximate Algorithm for the Kolmogorov–Smirnov and Lilliefors Testst

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, \dots, 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:

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

If the sample, X_1, \dots, X_n , has been sorted into nondecreasing order so that $X_1 \leq X_2 \leq \dots \leq X_n$, 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:

$$\begin{aligned} K_{\max}^+ &= \sqrt{n} \max_{1 \leq j \leq n} \left\{ \frac{j}{n} - F(X_j) \right\} \\ K_{\max}^- &= \sqrt{n} \max_{1 \leq j \leq n} \left\{ F(X_j) - \frac{j-1}{n} \right\} \\ K_{\max} &= \max \{ K_{\max}^+, K_{\max}^- \} \end{aligned} \quad (1.2)$$

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values of the test statistic distributions are available for the case where the actual parameters of $F(X)$ have been replaced by estimates computed from the sample.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 Kolmogorov-Smirnov test statistics [4] pp. 44//

Step 1 obtain the n observations X_1, X_2, \dots, X_n

Step 2 sort them so that $X_1 \leq X_2 \leq \dots \leq X_n$

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

end K

Since step 2 sorts the observations, it requires $O(n \log n)$ 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 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 $O(n)$. 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 $O(n)$. 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.

