Computing the Cumulative Distribution Function of the Kolmogorov–Smirnov Statistic

John H. Drew Andrew G. Glen Lawrence M. Leemis

Department of Mathematics

The College of William & Mary

Williamsburg, VA 23187–8795, USA

Email:{jhdrew, agglen, leemis}@math.wm.edu

16 June 1999

We present an algorithm for computing the cumulative distribution function of the Kolmogorov–Smirnov test statistic D_n in the all-parameters-known case. Birnbaum (1952) gives an n-fold integral for the CDF of the test statistic which yields a function defined in a piecewise fashion, where each piece is a polynomial of degree n. Unfortunately, it is difficult to determine the appropriate limits of integration for computing these polynomials. Our algorithm performs the required integrations in a manner that avoids calculating the same integrals repeatedly, resulting in shorter computation time. It can be used to compute the entire CDF or just a portion of the CDF, which is more efficient for finding a critical value or a p-value associated with a hypothesis test. If the entire CDF is computed, it can be stored in memory so that various characteristics of the distribution of the test statistic (e.g., moments) can be calculated. To date, critical tables have been approximated by various techniques including asymptotic approximations, recursive formulas, and Monte Carlo simulation. Our approach yields exact critical values and significance levels. The algorithm has been implemented in a computer algebra system.

Keywords: algorithm, computer algebra systems, continuous probability distributions, goodness-of-fit tests, probability.

Acknowledgment

The authors gratefully acknowledge the assistance of Professor Marina Kondratovitch for her early help with this work, and Professor Donald Barr for his assistance with the paper.

1 Introduction

The Kolmogorov–Smirnov (KS) test statistic D_n, is defined by

$$D_n = \sup_x |F(x) - F_n(x)|,$$

where n is the sample size, $F(x)$ is a hypothesized CDF with fixed parameters, and $F_n(x)$, also called the empirical distribution function, is a step-function that increases by $1/n$ at each data value. This statistic has been used for goodness-of-fit testing for continuous populations for decades. The KS test's appeal is its straightforward computation of the test statistic and the distribution-free characteristic of D_n. Its drawback is that its cumulative distribution function under the null hypothesis is difficult to determine, leaving one to calculate critical values with various approximation methods. We consider the distribution of the KS statistic in the case when all parameters are known.

Birnbaum (1952) gives the CDF of $D_n - \frac{1}{2n}$ as

$$P\left(D_n < \frac{1}{2n} + v\right) = n! \int_{\frac{1}{2n}-v}^{\frac{1}{2n}+v} \int_{\frac{3}{2n}-v}^{\frac{3}{2n}+v} \cdots \int_{\frac{2n-1}{2n}-v}^{\frac{2n-1}{2n}+v} g(u_1, u_2, \ldots, u_n)\, du_n \ \ldots \ du_2\, du_1$$

for $0 \le v \le \frac{2n-1}{2n}$, where

$$g(u_1, u_2, \ldots, u_n) = 1$$

for $0 \le u_1 \le u_2 \le \cdots \le u_n \le 1$, and is 0 otherwise. Birnbaum's rather harmless-looking integral is tedious to compute by hand, even for small values of n, due to the complexity of the region where g is nonzero. Evaluating his expression requires calculating many n-fold integrals whose limits are determined by considering a carefully chosen partition of the support of D_n. The difficult part of the process is to set the appropriate limits on these integrals.

An algorithm for computing the CDF of the KS test statistic in the all-parameters-known case for a positive integer parameter n is presented here. Computing the CDF provides a challenging calculus problem, even for small values of n. The CDF can be constructed from a set of piecewise degree-n polynomials. The algorithm can be used to plot the entire CDF or to just find particular fractiles or probabilities associated with the distribution.

2 Literature review

The literature available on the KS statistic is extensive. Stephens' article (Chapter 4 of D'Agostino and Stephens, 1986) contains comprehensive coverage on the use of the KS statistic, as well as other statistics based on the empirical distribution function. He calculates the power of these goodness-of-fit tests. Johnson, Kotz, and Balakrishnan (1995, p. 640) consider this source (D'Agostino, 1986) to be so complete that they have deleted KS discussions in their second edition and refer the reader to that compendium instead. For computing critical points of the KS distribution, we find five-digit accuracy as early as 1956 (Miller, 1956). Miller relies on asymptotic results that converge fairly quickly to produce these estimates. Birnbaum's article (1952) also explains how various sets of recursive formulas can be used to calculate certain critical points to reasonable levels of accuracy. Law and Kelton (1991, p. 387) indicate that critical points require computer techniques, and are only easily calculated for $n \leq 50$. There appears to be no source that produces exact distribution functions for any distribution where $n > 3$ in the literature. Birnbaum (1952, p. 441) gives the CDF of $D_n - \frac{1}{2n}$ for $n = 2$ and $n = 3$. Knuth (1981) provides a functional form for the CDFs for the two statistics D_n^+ and D_n^-, but does not provide the CDF for $D_n = \max\{D_n^+, D_n^-\}$, a harder problem given the dependence between the two random variables. Schröer and Trenkler (1995) give the distribution of the KS test statistic for unequal sample sizes in the two- and three-sample cases. As a consequence of the apparent complexity of the required integration and lack of literature on exact distributions, we believed early on that a computational algebra system, such as Maple, would be necessary to compute the polynomials needed for the distribution of the KS test statistic.

3 Computing the distribution of D_n

When $0 < v < \frac{1}{2n}$, Birnbaum's integral is easy to compute since none of the intervals of integration overlap. Additionally, these intervals are all wholly contained within the interval from 0 to 1. Because the limits of Birnbaum's integral guarantee that $0 \le u_1 \le u_2 \le \cdots \le u_n \le 1$, we may replace the integrand $g(u_1, u_2, \ldots, u_n)$ with 1, and our computation of the KS CDF requires only a single n-fold integral:

$$P\left(D_n < \frac{1}{2n} + v\right) = n! \int_{\frac{1}{2n}-v}^{\frac{1}{2n}+v} \int_{\frac{3}{2n}-v}^{\frac{3}{2n}+v} \cdots \int_{\frac{2n-1}{2n}-v}^{\frac{2n-1}{2n}+v} 1 \, du_n \, \ldots \, du_2 \, du_1 = n!(2v)^n, \qquad 0 < v < \frac{1}{2n}.$$

When v has a fixed value greater than $\frac{1}{2n}$, it is also desirable to replace Birnbaum's integrand by 1. In order to justify this replacement, we must only allow integration over that region of n-dimensional space for which $0 \le u_1 \le u_2 \le \cdots \le u_n \le 1$. Since the intervals of integration specified in Birnbaum's integral for different u_i's can overlap, the smallest allowable value for any u_i is influenced by all u_j's with subscripts less than i that can take on values in u_i's interval of integration. This overlapping requires partitioning the interval from $\frac{1}{2n} - v$ to $\frac{2n-1}{2n} + v$ into subintervals (which we will henceforth refer to as u-subintervals), with the first u-subinterval starting at $\frac{1}{2n} - v$ and a new u-subinterval beginning whenever the left endpoint of one of Birnbaum's intervals of integration is encountered. When any u_i lies in a u-subinterval that consists entirely of negative values, Birnbaum's integrand is zero. For this reason, only u-subintervals that have a positive right endpoint contribute to the KS CDF.

Of course, the number of u-subintervals that have a positive right endpoint depends on the value of v. Because of this, the interval $0 < v < \frac{2n-1}{2n}$ must be subdivided at the following values of v: $\frac{1}{2n}$, $\frac{3}{2n}$, $\frac{5}{2n}$, ..., $\frac{2n-3}{2n}$. When the values of v remain within one of the resulting subintervals for v, the number of u-subintervals that have a positive right endpoint will remain fixed.

Another complication arises because it is necessary to know, for a fixed value of v and for the u-subintervals produced by this value, which variables of integration u_i can take on values in each of the u-subintervals of $[0, 1]$. The previous subdivision of the values of v is not fine enough to allow unambiguous specification of the range for each u_i. When $i < j$, u_i and

u_j have overlapping intervals of integration if the upper integration limit for u_i exceeds the lower integration limit for u_j, i.e., $\frac{2i-1}{2n} + v \geq \frac{2j-1}{2n} - v$. As a result, as v increases from 0 to $\frac{2n-1}{2n}$, new overlaps take place when v equals $\frac{1}{2n}$, $\frac{2}{2n}$, $\frac{3}{2n}$, ..., $\frac{n-1}{2n}$. The interval $0 < v < \frac{2n-1}{2n}$ must be divided into subintervals at these values of v as well as at the values of v listed previously. We will henceforth refer to the resulting subintervals as v-subintervals.

Indicator matrices will be used to summarize the interrelationships between the possible values for the variables of integration $u_1, u_2, \ldots, u_n$. For a fixed n and for values of v in the kth v-subinterval of $\left(0, \frac{2n-1}{2n}\right)$, the indicator matrix A_k will show, by the presence of a 1 or a 0 in row i and column j, whether or not u_i can take on values in the jth u-subinterval. Finally, by defining paths through these indicator matrices, we will produce a complete set of n-fold integrals which satisfy the requirement $0 \leq u_1 \leq u_2 \leq \cdots \leq u_n \leq 1$, and which produce the KS CDF when summed.

The algorithm for computing the cumulative distribution function of D_n is divided into four phases. In Phase 1, an appropriate partition of the support of $D_n - \frac{1}{2n}$ is determined. In Phase 2, we define matrices A_k that are instrumental in determining the limits of integration in the n-fold integrals. In Phase 3, these integrals and their associated limits are computed. To take advantage of the similarities in the limits of integration of these n-fold integrals, they are grouped for efficiency and all evaluated to the same level of integration before proceeding to the next level. Finally, Phase 4 consists of translating the support of $D_n - \frac{1}{2n}$ to the support of D_n. Thus the input to the algorithm is a positive integer n and the output is the piecewise CDF of D_n. The algorithm is available from the authors.

3.1 Phase 1: Partition the support of $D_n - \frac{1}{2n}$

The endpoints of the segments that define the support of the KS test statistic can be determined from the limits of integration given by Birnbaum. Using Birnbaum's formula, the *baseline* lower and upper limits of integration associated with $v = 0$ are

$$\left\{\frac{1}{2n}, \frac{3}{2n}, \frac{5}{2n}, \ldots, \frac{2n-1}{2n}\right\}.$$

As v increases, the support of $D_n - \frac{1}{2n}$ is broken into disjoint v-subintervals. The endpoints of these support v-subintervals consist of $v = 0$ and the values of v for which the endpoints of the n intervals of integration either:

- equal 0 or 1, which occurs at the v-values $\left\{\frac{1}{2n}, \frac{3}{2n}, \frac{5}{2n}, \ldots, \frac{2n-1}{2n}\right\}$, *or*
- coincide, which occurs at the v-values $\left\{\frac{1}{2n}, \frac{2}{2n}, \frac{3}{2n}, \ldots, \frac{n-1}{2n}\right\}$.

Thus the union of these two sets and 0 comprise the endpoints of the v-subintervals of the support of $D_n - \frac{1}{2n}$.

The first phase of the algorithm computes the above endpoints $v_0 = 0, v_1, v_2, \ldots, v_m$, where m is the number of v-subintervals on which the CDF of $D_n - \frac{1}{2n}$ is defined. For any n,

$$m = \left\lceil \frac{3n}{2} \right\rceil - 1,$$

where $\lceil\ \rceil$ denotes the ceiling function.

3.2 Phase 2: Define the A matrices

Two book-keeping steps are needed in this phase of the algorithm. They are:

1. Define $c_1, c_2, \ldots, c_m$ as the midpoints of the v-subintervals of support for $D_n - \frac{1}{2n}$.
2. Define $x_1, x_2, \ldots, x_n$ as the midpoints of the intervals of integration for $u_1, u_2, \ldots, u_n$ in Birnbaum's n-fold integral. Thus

$$x_i = \frac{2i-1}{2n} \qquad\qquad i = 1, 2, \ldots, n.$$

Part of the algorithm involves defining $n \times n$ indicator matrices $A_1, A_2, \ldots, A_m$ corresponding to the v-subintervals which form the support of $D_n - \frac{1}{2n}$. The rows of an A matrix refer to the variables of integration $u_1, u_2, \ldots, u_n$. The columns of an A matrix refer to the u-subintervals, with the jth column corresponding to the interval from $x_j - v$ to $x_{j+1} - v$, except for the nth column, which corresponds to the interval from $x_n - v$ to $x_n + v$. If the (i, j) element of an A matrix equals 1, then the range of u_i includes at least part of the jth u-subinterval, as will be explained in Phase 3 below.

The integrals that need to be computed for each segment of support of $D_n - \frac{1}{2n}$ can be visualized as a path in A, consisting of a sequence of moves from the (n, n) position of A to a nonzero element in the first row of A. All moves in the A matrix from row i to row $i-1$ (for $i = n, n-1, \ldots, 2$) require that the following conditions are met:

1. The move must be from one of the 1's in row i to one of the 1's in row $i-1$.

2. The move is either directly vertical or vertical and one or more spaces to the left.

That is, if the move begins at a 1 element in column j of row i, it must end at a 1 element in column 1 through j of row $i-1$.

Example 3.1 For $n = 3$ and $k = 3$, the support of the CDF includes the v-subinterval $\frac{2}{6} < v < \frac{3}{6}$ and has the following A_3 matrix:

$$A_3 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

There are five different paths from the (3, 3) element to the top row of the matrix. In the five matrices below, these five paths to the top row of the A matrix are displayed by using boldface type on the path.

$$\begin{bmatrix} \mathbf{1} & 1 & 1 \\ 0 & \mathbf{1} & 1 \\ 0 & 0 & \mathbf{1} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{1} & 1 \\ 0 & \mathbf{1} & 1 \\ 0 & 0 & \mathbf{1} \end{bmatrix}, \begin{bmatrix} \mathbf{1} & 1 & 1 \\ 0 & 1 & \mathbf{1} \\ 0 & 0 & \mathbf{1} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{1} & 1 \\ 0 & 1 & \mathbf{1} \\ 0 & 0 & \mathbf{1} \end{bmatrix}, \begin{bmatrix} 1 & 1 & \mathbf{1} \\ 0 & 1 & \mathbf{1} \\ 0 & 0 & \mathbf{1} \end{bmatrix}.$$

Each path corresponds to a triple integral, whose limits will be specified in Phase 3. The sum of these integrals produces the CDF on $\frac{2}{6} < v < \frac{3}{6}$, as will be shown in Example 3.3 below.

Table 1 shows how rapidly the number of n-fold integrations required to compute the CDF of $D_n - \frac{1}{2n}$ grows with n. Note that the five triple integrals in the previous example corresponds to $n = 3$ and the third v-subinterval ($k = 3$), shown in boldface in the table. Note that the fifth column in Table 1 is the product of the first and fourth columns.

Table 1. Computational requirements for computing the D_n CDF for small n.

n	m	Number of n-fold integrals required for each v-subinterval	Total number of n-fold integrals required	Total number of single integrals required
1	1	1	1	1
2	2	1, 2	3	6
3	4	1, 4, **5**, 3	13	39
4	5	1, 8, 13, 9, 4	35	140
5	7	1, 16, 34, 27, 28, 14, 5	125	625
6	8	1, 32, 89, 81, 89, 48, 20, 6	366	2196

Example 3.2 The method for determining the integers in the third column of the previous table is somewhat analogous to determining the entries in a row of Pascal's triangle by using the entries of the previous row. Beginning at each nonzero A matrix element in rows 2 through n, there are a certain number of paths whose moves to the top row of the matrix satisfy the two conditions stated previously. We will assign such elements a number p (for paths) that denotes the number of paths from that element to the top of the matrix. Thus if $p_{2,2} = 2$ (as is the case for the (2, 2) element of the A_3 matrix for $n = 3$ and $k = 3$ in Example 3.1) then no matter how we reached this element of the A matrix, there are only two possible paths remaining to the top. The $p_{i,j}$ values are assigned as follows. Only those elements of matrix A with value 1 are assigned a $p_{i,j}$ value. If the jth entry in the first row of A is nonzero, assign $p_{1,j} = 1$. For the second and subsequent rows ($i = 2, 3, \ldots, n$), $p_{i,j} = \sum_{q=1}^{j} p_{i-1,q}$. Upon completion, $p_{n,n}$ represents the total number of paths from the lower right corner of the matrix to the top row of the matrix. Consider, for example, the case of $n = 5$ and $k = 4$.

The A_4 matrix with superscripts denoting the $p_{i,j}$ values is

$$A_4 = \begin{bmatrix} 0 & 1^1 & 1^1 & 1^1 & 0 \\ 0 & 1^1 & 1^2 & 1^3 & 1^3 \\ 0 & 0 & 1^3 & 1^6 & 1^9 \\ 0 & 0 & 0 & 1^9 & 1^{18} \\ 0 & 0 & 0 & 0 & 1^{27} \end{bmatrix}.$$

Here we see that $p_{5,5} = 27$, meaning there are 27 possible paths to the top row of A_4, consistent with the $n = 5$ and $k = 4$ element in Table 1. The $p_{i,j}$ values parallel the work on the two-sample case in Schröer and Trenkler (1995, page 187).

3.3 Phase 3: Set limits on the appropriate integrals

The indicator matrix A shows, by the presence of a 1 in row i and column j, for $j < n$, that u_i can assume values from the maximum of $x_j - v$ and 0 to the minimum of $x_{j+1} - v$ and $x_i + v$. The presence of a 1 in row i and column n, means that u_i assumes values from $x_n - v$ to the minimum of $x_i + v$ and 1. Each path in A, as described in Section 3.2, represents a set of allowable intervals of integration for the variables u_1 to u_n in a particular n-fold integral. For a particular path, the limits of integration for each u_i are those given above that correspond to entry $a_{i,j}$ of A in row i that lies on the path, with one exception: if both $a_{i,j}$ and $a_{i-1,j}$ are on the same path, the lower limit of integration for u_i must be u_{i-1}.

For each path through the matrix A, and for each nonzero entry $a_{i,j}$ on that path, a single integration with respect to u_i must be performed. If $a_{i-1,j}$ is on the same path as $a_{i,j}$, then the lower limit of integration will be the variable u_{i-1}. If $a_{i-1,j}$ is not on the same path as $a_{i,j}$, then the lower limit of integration will be a fixed number: the maximum of $x_j - v$ and 0. Thus for each path which passes through $a_{i,j}$, either a variable or fixed lower limit of integration might be appropriate.

Example 3.3 Using the rules above, the five paths displayed in Example 3.1

correspond, respectively, to the following integrals:

$$\int_0^{\frac{3}{6}-v} \int_{\frac{3}{6}-v}^{\frac{5}{6}-v} \int_{\frac{5}{6}-v}^{1} 1\, du_3\, du_2\, du_1,$$

$$\int_{\frac{3}{6}-v}^{\frac{5}{6}-v} \int_{u_1}^{\frac{5}{6}-v} \int_{\frac{5}{6}-v}^{1} 1\, du_3\, du_2\, du_1,$$

$$\int_0^{\frac{3}{6}-v} \int_{\frac{5}{6}-v}^{\frac{3}{6}+v} \int_{u_2}^{1} 1\, du_3\, du_2\, du_1,$$

$$\int_{\frac{3}{6}-v}^{\frac{5}{6}-v} \int_{\frac{5}{6}-v}^{\frac{3}{6}+v} \int_{u_2}^{1} 1\, du_3\, du_2\, du_1,$$

$$\int_{\frac{5}{6}-v}^{\frac{1}{6}+v} \int_{u_1}^{\frac{3}{6}+v} \int_{u_2}^{1} 1\, du_3\, du_2\, du_1.$$

When these five integrals are summed, Birnbaum's expression reduces to

$$P\left(D_3 < \frac{1}{6} + v\right) = -4v^3 + \frac{11}{3}v - \frac{11}{27}, \qquad \frac{2}{6} < v < \frac{3}{6}.$$

Note that the inside integral

$$\int_{u_2}^{1} 1\ du_3$$

on the third, fourth, and fifth triple integral is identical. The number of identical integrals of this type grows rapidly as n increases. The algorithm we have developed avoids recalculation of duplicate integrals. Continuing in this fashion for the other A matrices (i.e., A_1, A_2, and A_4), yields the CDF

$$P\left(D_3 < \frac{1}{6} + v\right) = \begin{cases} 48v^3 & 0 < v < \frac{1}{6} \\ -12v^3 + 8v^2 + v - \frac{1}{9} & \frac{1}{6} < v < \frac{2}{6} \\ -4v^3 + \frac{11}{3}v - \frac{11}{27} & \frac{2}{6} < v < \frac{3}{6} \\ 2v^3 - 5v^2 + \frac{25}{6}v - \frac{17}{108} & \frac{3}{6} < v < \frac{5}{6}. \end{cases}$$

Note that Birnbaum's (1952) CDF contains a sign error in the fourth v-subinterval of support.

Example 3.4 To illustrate the role of fixed and variable limits, consider again

the case of $n = 5$ and $k = 4$ as in Example 3.2. The A_4 matrix shown below has its 1's replaced with $\mathcal{F}$, $\mathcal{V}$, or $\mathcal{B}$, indicating whether a fixed-limit integral, a variable-limit integral, or both need to be computed for each entry.

$$A_4 = \begin{bmatrix} 0 & \mathcal{F} & \mathcal{F} & \mathcal{F} & 0 \\ 0 & \mathcal{V} & \mathcal{B} & \mathcal{B} & \mathcal{F} \\ 0 & 0 & \mathcal{B} & \mathcal{B} & \mathcal{B} \\ 0 & 0 & 0 & \mathcal{B} & \mathcal{B} \\ 0 & 0 & 0 & 0 & \mathcal{B} \end{bmatrix}.$$

In general, when the A matrix contains a zero, neither the fixed nor variable lower limits need to be computed. Now consider the elements of the A matrix that contain a 1. The positions associated with the first 1 in each column of A require only a fixed lower limit to be calculated. The positions below the first 1 in the first nonzero column require only a variable lower limit to be calculated. All other positions in the A matrix associated with 1 elements require both a fixed a variable lower limit to be calculated. Table 2 shows the computational efficiency of performing the integrations for various values of n by using the algorithm.

Table 2. Computational efficiency associated with using the F and V arrays.

n	m	Total number of single integrals required (Table 1)	Total number of single integrals required (algorithm)
1	1	1	0
2	2	6	4
3	4	39	23
4	5	140	48
5	7	625	108
6	8	2196	170
10	14	442540	800
15	22	318612735	2793

The number of single integrals (algorithm) for any n is bounded above by

$$\frac{n(n+1)}{2} \cdot 2 \cdot \left(\left\lceil \frac{3n}{2} \right\rceil - 2 \right) \approx \frac{3}{2} n^3,$$

where the first factor corresponds to the maximum number of 1's in any A matrix, the second factor accounts for computing both the fixed and variable matrices (described subsequently), and the last factor is one less than m since no integration is required for the first v-subinterval.

Table 3 lists the coefficients of the polynomials that define the CDF of $D_n - \frac{1}{2n}$ as computed by the algorithm which has been implemented in Maple for $n = 1, 2, \ldots, 6$. The ability to integrate polynomials and to store fractions exactly is required for these calculations. Note that D_1 is uniformly distributed as expected.

Table 3. CDFs of $D_n - \frac{1}{2n}$ for $n = 1, 2, \ldots, 6$.

n	m	k	Coefficients of CDF polynomials	Subinterval
1	1	1	$2, 0$	$0 < v < \frac{1}{2}$
2	2	1	$8, 0, 0$	$0 < v < \frac{1}{4}$
		2	$-2, 3, -\frac{1}{8}$	$\frac{1}{4} < v < \frac{3}{4}$
3	4	1	$48, 0, 0, 0$	$0 < v < \frac{1}{6}$
		2	$-12, 8, 1, -\frac{1}{9}$	$\frac{1}{6} < v < \frac{2}{6}$
		3	$-4, 0, \frac{11}{3}, -\frac{11}{27}$	$\frac{2}{6} < v < \frac{3}{6}$
		4	$2, -5, \frac{25}{6}, -\frac{17}{108}$	$\frac{3}{6} < v < \frac{5}{6}$
4	5	1	$384, 0, 0, 0, 0$	$0 < v < \frac{1}{8}$
		2	$-48, 0, 15, -\frac{9}{8}, \frac{3}{256}$	$\frac{1}{8} < v < \frac{2}{8}$
		3	$16, -40, 21, -\frac{5}{8}, -\frac{29}{256}$	$\frac{2}{8} < v < \frac{3}{8}$
		4	$6, -7, -\frac{27}{16}, \frac{293}{64}, -\frac{853}{2048}$	$\frac{3}{8} < v < \frac{5}{8}$
		5	$-2, 7, -\frac{147}{16}, \frac{343}{64}, -\frac{353}{2048}$	$\frac{5}{8} < v < \frac{7}{8}$
5	7	1	$3840, 0, 0, 0, 0, 0$	$0 < v < \frac{1}{10}$
		2	$0, -288, \frac{624}{5}, -\frac{96}{25}, -\frac{36}{125}, \frac{6}{625}$	$\frac{1}{10} < v < \frac{2}{10}$
		3	$160, -160, \frac{24}{5}, \frac{616}{25}, -\frac{332}{125}, \frac{6}{125}$	$\frac{2}{10} < v < \frac{3}{10}$
		4	$-20, 64, -\frac{318}{5}, \frac{542}{25}, \frac{343}{500}, -\frac{273}{1250}$	$\frac{3}{10} < v < \frac{4}{10}$
		5	$12, 0, -\frac{62}{5}, \frac{6}{5}, \frac{2391}{500}, -\frac{3413}{6250}$	$\frac{4}{10} < v < \frac{5}{10}$
		6	$-8, 18, -\frac{52}{5}, -\frac{19}{5}, \frac{1838}{250}, -\frac{10527}{25000}$	$\frac{5}{10} < v < \frac{7}{10}$
		7	$2, -9, \frac{81}{5}, -\frac{729}{50}, \frac{6561}{1000}, -\frac{9049}{50000}$	$\frac{7}{10} < v < \frac{9}{10}$
6	8	1	$46080, 0, 0, 0, 0, 0, 0$	$0 < v < \frac{1}{12}$
		2	$2880, -3360, 660, 60, -\frac{95}{12}, \frac{5}{24}, -\frac{5}{5184}$	$\frac{1}{12} < v < \frac{2}{12}$
		3	$320, 480, -700, \frac{5620}{27}, -\frac{125}{36}, -\frac{275}{216}, \frac{2195}{46656}$	$\frac{2}{12} < v < \frac{3}{12}$
		4	$-280, 420, -\frac{335}{2}, -\frac{1675}{54}, \frac{26005}{864}, -\frac{3125}{1728}, -\frac{20645}{373248}$	$\frac{3}{12} < v < \frac{4}{12}$
		5	$104, -188, \frac{1235}{6}, -\frac{7435}{54}, \frac{36245}{864}, -\frac{7327}{5184}, -\frac{69797}{373248}$	$\frac{4}{12} < v < \frac{5}{12}$
		6	$-20, 22, \frac{45}{4}, -\frac{2005}{108}, -\frac{185}{1728}, \frac{57971}{10368}, -\frac{406469}{746496}$	$\frac{5}{12} < v < \frac{7}{12}$
		7	$10, -33, \frac{925}{24}, -\frac{3065}{216}, -\frac{22175}{3456}, \frac{134807}{20736}, -\frac{632863}{1492992}$	$\frac{7}{12} < v < \frac{9}{12}$
		8	$-2, 11, -\frac{605}{24}, \frac{6655}{216}, -\frac{73205}{3456}, \frac{161051}{20736}, -\frac{278569}{1492992}$	$\frac{9}{12} < v < \frac{11}{12}$

Rather than detail every aspect of the logic of Phase 3 of the algorithm, we illustrate the evolution of the $n \times n$ matrices F (for fixed limits) and V (for variable limits) in a row-by-row fashion for a particular combination of n and k.

Example 3.5 Consider again the case of $n = 3$ and $k = 3$, which corresponds to the v-subinterval $\frac{2}{6} < v < \frac{3}{6}$. The x_i values associated with $n = 3$ are $x_1 = \frac{1}{6}$, $x_2 = \frac{3}{6}$, and $x_3 = \frac{5}{6}$. The v_i values associated with $n = 3$ are $v_0 = 0$, $v_1 = \frac{1}{6}$, $v_2 = \frac{2}{6}$, $v_3 = \frac{3}{6}$, and $v_4 = \frac{5}{6}$. The centers of the v-subintervals are $c_1 = \frac{1}{12}$, $c_2 = \frac{3}{12}$, $c_3 = \frac{5}{12}$, and $c_4 = \frac{8}{12}$. The A_3 matrix is

$$A_3 = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

The F and V matrices are computed simultaneously, row by row, starting with the last row and ending with the first. Thus all of the elements of the third row of F and of V are computed before computing the elements of the second row is begun. In general, integrals which are calculated in a particular row become integrand candidates for the integration to be carried out in the row immediately above. The F and V matrices are designed to store the intermediate results of integration of the n-fold integrals so that the necessary inner integration operations are performed only once. Entries in the ith rows of F and V each result from $n + 1 - i$ single integrations, one integration corresponding to the ith row and one integration for every lower row. The F matrix stores the results of all integration in which the last integral had a fixed lower limit and the V matrix stores the results of all integration in which the last integral had a variable lower limit. In this example, for the inner-most integration variable u_3, there is only one possible fixed integration to compute and one possible variable integration to compute on the third u-subinterval. These computations are seen in the (3,

3) element of the respective matrices:

$$F = \begin{bmatrix} F_{1,1} & F_{1,2} & F_{1,3} \\ 0 & F_{2,2} & F_{2,3} \\ 0 & 0 & \int_{\frac{5}{6}-v}^{1} 1\,du_3 \end{bmatrix}, \qquad V = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} \\ 0 & V_{2,2} & V_{2,3} \\ 0 & 0 & \int_{u_2}^{1} 1\,du_3 \end{bmatrix}.$$

The second variable, u_2, has four possible combinations of limits and integrands. Elements (2, 2) and (2, 3) of the F matrix store integration results that have all fixed lower-limits of integration, but the integrands are fixed and variable, respectively. Elements (2, 2) and (2, 3) of the V matrix have a variable lower limit of integration u_1. $F_{2,2}$ covers the case in which u_2 varies over the second u-subinterval and u_3 varies over the third u-subinterval. For $F_{2,3}$, both u_2 and u_3 vary over the third u-subinterval, so u_3 must have u_2 as its variable lower limit; hence the integrand is $V_{3,3}$, not $F_{3,3}$. Similar cases are covered by $V_{2,2}$ and $V_{2,3}$, except that variable lower limits are used in anticipation of the fact that u_1 and u_2 can both vary over the same u-subinterval, making it necessary to have u_1 as the (variable) lower limit for u_2. The integrations in the second row of each matrix are shown below (note, the integration for the (3, 3) element, previously discussed, has been carried out).

$$F = \begin{bmatrix} F_{1,1} & F_{1,2} & F_{1,3} \\ 0 & \int_{\frac{1}{2}-v}^{\frac{5}{6}-v} F_{3,3}\,du_2 & \int_{\frac{5}{6}-v}^{\frac{1}{2}+v} V_{3,3}\,du_2 \\ 0 & 0 & \frac{1}{6}+v \end{bmatrix}, \qquad V = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} \\ 0 & \int_{u_1}^{\frac{5}{6}-v} F_{3,3}\,du_2 & \int_{u_1}^{\frac{1}{2}+v} V_{3,3}\,du_2 \\ 0 & 0 & 1-u_2 \end{bmatrix}.$$

For the third and last variable, u_1, only fixed-limit integration takes place so only the F matrix is updated. The integrals in this first row are shown below (note again, all previously discussed integration have been carried out in the second and third rows of F and V):

$$F = \begin{bmatrix} \int_{0}^{\frac{1}{2}-v} (F_{2,2}+F_{2,3})\,du_1 & \int_{\frac{1}{2}-v}^{\frac{5}{6}-v} (V_{2,2}+F_{2,3})\,du_1 & \int_{\frac{5}{6}-v}^{\frac{1}{6}+v} V_{2,3}\,du_1 \\ 0 & \frac{1}{3}v+\frac{1}{18} & -\frac{1}{9}+\frac{2}{3}v \\ 0 & 0 & \frac{1}{6}+v \end{bmatrix},$$

$$V = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} \\ 0 & -v^2 + \frac{2}{3}v + \frac{5}{36} - vu_1 - \frac{1}{6}u_1 & \frac{1}{2}v^2 + \frac{1}{2}v + \frac{3}{8} + \frac{1}{2}u_1^2 - u_1 \\ 0 & 0 & 1 - u_2 \end{bmatrix}.$$

The completely evaluated F matrix is given by

$$F = \begin{bmatrix} -v^2 + \frac{5}{9}v - \frac{1}{36} & -\frac{1}{36} + \frac{5}{18}v & -\frac{2}{3}v^3 + v^2 - \frac{2}{9}v - \frac{1}{81} \\ 0 & -\frac{1}{3}v + \frac{1}{18} & -\frac{1}{9} + \frac{2}{3}v \\ 0 & 0 & \frac{1}{6} + v \end{bmatrix}.$$

Finally, the CDF on the 3rd v-subinterval is the sum of the elements in the first row of the F matrix, all multiplied by $3! = 6$.

3.4 Phase 4: Shift the distribution

At this point, the CDF is computed in the form $P(D_n < v + \frac{1}{2n})$. Now we convert the distribution into the more usable form $P(D_n < y) = F_{D_n}(y)$ by making the substitution $y = v + \frac{1}{2n}$ in both the polynomials and the v-subintervals of the CDF. Specifically, we add $\frac{1}{2n}$ to each endpoint of the v-subintervals and we substitute $(y - \frac{1}{2n})$ for v in the CDF polynomials. We then take these two lists of numbers and polynomials respectively, simplify them [using Maple's `simplify()` command], and create the CDF representation in the form of the "list-of-lists" representation for distributions outlined in Glen (1998). This enables us to use the distribution in the ways that all other distributions are used in the software proposed by Glen (1998). Specifically, we can now verify critical values for the distribution, but more importantly, we can calculate exact significance levels of any given statistic. An

example of the shifted distribution for $n = 6$ is:

$$F_{D_6}(y) = \begin{cases} 0 & y < \frac{1}{12} \\ 46080\,y^6 - 23040\,y^5 + 4800\,y^4 - \frac{1600}{3}\,y^3 + \frac{100}{3}\,y^2 - \frac{10}{9}\,y + \frac{5}{324} & \frac{1}{12} \le y < \frac{1}{6} \\ 2880\,y^6 - 4800\,y^5 + 2360\,y^4 - \frac{1280}{3}\,y^3 + \frac{235}{9}\,y^2 + \frac{10}{27}\,y - \frac{5}{81} & \frac{1}{6} \le y < \frac{1}{4} \\ 320\,y^6 + 320\,y^5 - \frac{2600}{3}\,y^4 + \frac{4240}{9}\,y^3 - \frac{785}{9}\,y^2 + \frac{145}{27}\,y - \frac{35}{1296} & \frac{1}{4} \le y < \frac{1}{3} \\ -280\,y^6 + 560\,y^5 - \frac{1115}{3}\,y^4 + \frac{515}{9}\,y^3 + \frac{1525}{54}\,y^2 - \frac{565}{81}\,y + \frac{5}{16} & \frac{1}{3} \le y < \frac{5}{12} \\ 104\,y^6 - 240\,y^5 + 295\,y^4 - \frac{1985}{9}\,y^3 + \frac{775}{9}\,y^2 - \frac{7645}{648}\,y + \frac{5}{16} & \frac{5}{12} \le y < \frac{1}{2} \\ -20\,y^6 + 32\,y^5 - \frac{185}{9}\,y^3 + \frac{175}{36}\,y^2 + \frac{3371}{648}\,y - 1 & \frac{1}{2} \le y < \frac{2}{3} \\ 10\,y^6 - 38\,y^5 + \frac{160}{3}\,y^4 - \frac{265}{9}\,y^3 - \frac{115}{108}\,y^2 + \frac{4651}{648}\,y - 1 & \frac{2}{3} \le y < \frac{5}{6} \\ -2\,y^6 + 12\,y^5 - 30\,y^4 + 40\,y^3 - 30\,y^2 + 12\,y - 1 & \frac{5}{6} \le y < 1 \\ 1 & y \ge 1. \end{cases}$$

The CDF and PDF of D_6 are shown in Figure 1 and Figure 2. We have noted that there is a discontinuity in the PDF at $y = \frac{1}{6}$, and more generally, at $y = \frac{1}{n}$. This corresponds to the value $v = \frac{1}{2n}$ in Birnbaum's original integral. This is the smallest value of v for which the n-dimensional hypercube centered at $\left(\frac{1}{2n}, \frac{3}{2n}, \ldots, \frac{2n-1}{2n}\right)$ makes contact with the boundary of the region in n-space that satisfies $0 \le u_1 \le u_2 \le \cdots \le u_n \le 1$. The CDFs for D_1 through D_{30} are available in ASCII form (readable by Maple) at `www.math.wm.edu/~eemis.html`. The CDF for $n = 30$ requires about 20 printed pages of text.

4 Conclusions and further research

The algorithm described here computes the exact CDF of the Kolmogorov–Smirnov test statistic in the all-parameters-known case, effectively eliminating the need for tables when the sample size is small. Computing the CDF provides a challenging calculus problem, even for small values of n. The resultant CDF is set of piecewise degree-n polynomials. While the algorithm is slow for large n, once the CDF is generated and stored, it may be used with virtually no computation time. The CDFs for $n = 1$ to $n = 30$ have been stored. Once the CDF has been generated and stored, subsequent computations of characteristics (e.g.,

p-values, critical points, moments) are fast. The mean of D_{25}, for example, is exactly:

$$\frac{19789174192091655069533351340633597236049}{11823431123048067092895507812500000000000000}$$

We have compared the accuracy of Birnbaum's approximate methods in his Table 2 and found errors in the third digit of the CDF values Furthermore, asymptotic methods typically only find the CDF associated with a single point, whereas our algorithm exactly determines the entire distribution of the test statistic.

The algorithm has been implemented in Maple and incorporated into a larger package of procedures known as APPL (A Probability Programming Language). This package can be used to solve problems such as the one that follows. Let X be a KS random variable (all parameters known) with $n = 6$. Let Y be a KS random variable (all parameters known) with $n = 4$. Assuming that X and Y are independent, find $Var[\max\{X, Y\}]$. One would typically be limited to Monte Carlo simulation in order to solve this question without the algorithm given in the paper. The APPL code to solve this problem is

```
X := KSRV(6);
Y := KSRV(4);
Z := Maximum(X, Y);
Variance(Z);
```

which yields the variance as exactly

$$\frac{10251047454659775800001920152 79}{83793210145582989309719976345600}$$

or approximately 0.0122337.

References

[1] D'Agostino, R.B., and M.A. Stephens, *Goodness-of-Fit Techniques* (Marcel Dekker, New York, 1986).

[2] Birnbaum, Z.W., Numerical Tabulation of the Distribution of Kolomogorov's Statistic for Finite Sample Size, *Journal of the American Statistical Association*, 47 (1952) 425–441.

[3] Glen, A.G., A Probability Programming Language: Development and Application (Ph.D. Dissertation, Department of Mathematics, The College of William & Mary, 1998).

[4] Johnson, N.L., S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Volume 2,* Second Edition (John Wiley & Sons, New York, 1995).

[5] Knuth, D.E., *The Art of Computer Programming,* Second Edition (Addison–Wesley, Reading, Mass, 1981).

[6] Law, A.M. and W.D. Kelton, *Simulation Modeling and Analysis,* Second Edition (McGraw–Hill, New York, 1991).

[7] Miller, L.H., Table of Percentage Points of Kolmogorov Statistics, *Journal of the American Statistical Association,* 51 (1956) 111–121.

[8] Owen, D.B., *Handbook of Statistical Tables*, (Addison–Wesley, Reading, Mass, 1962).

[9] Schröer, G. and Trenkler, D., Exact and Randomized distributions of Kolmogorov–Smirnov Tests for Two or Three Samples, *Computational Statistics and Data Analysis,* 20 (1995), 185–202.

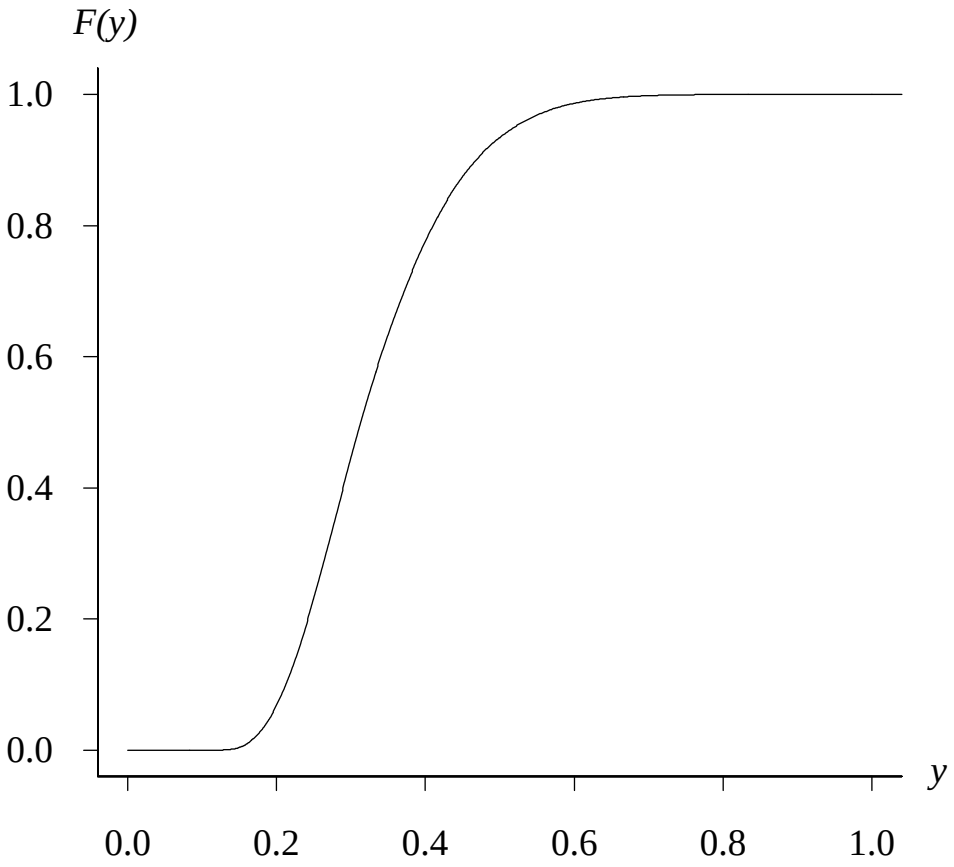

Figure 1: The CDF of the D_6 random variable.

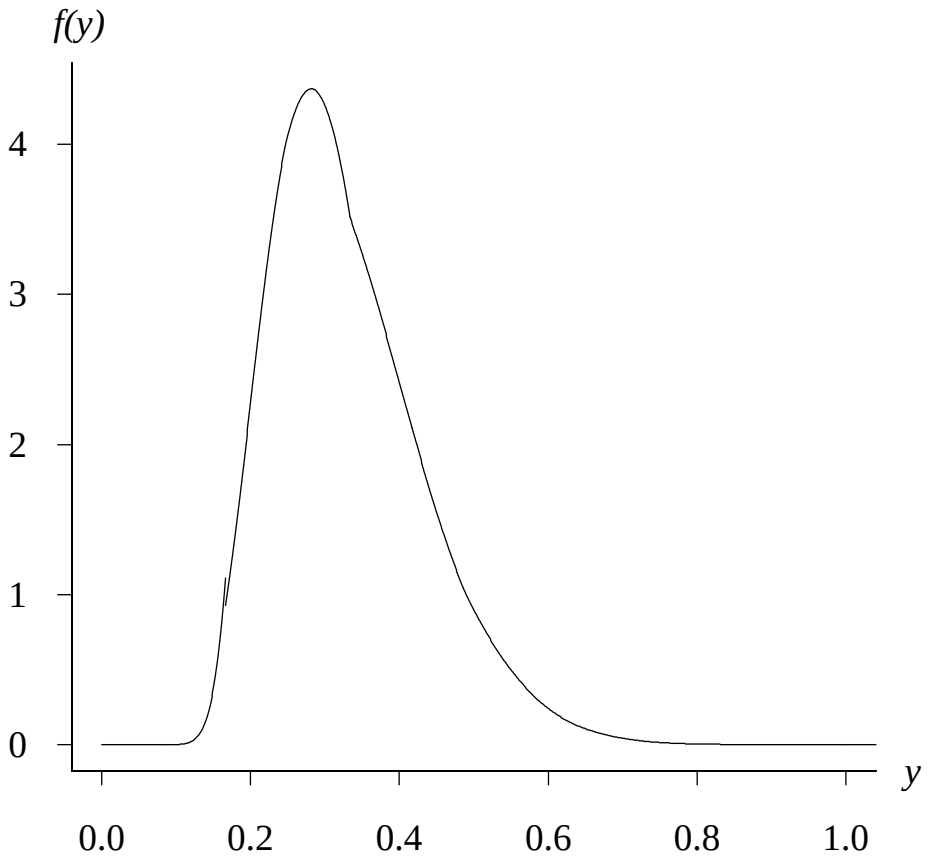

Figure 2: The PDF of the D_6 random variable. Note the discontinuity at $y = 1/6$.