A PROCEDURE TO FIND EXACT CRITICAL VALUES OF KOLMOGOROV-SMIRNOV TEST

Silvia Facchinetti

Dipartimento di Scienze statistiche, Università Cattolica del Sacro Cuore, Milano, Italia

Abstract The compatibility of a random sample of data with a given distribution can be checked with a goodness of fit test. Kolmogorov (1933) and Smirnov (1939) proposed the $D_n$ statistic based on the comparison between the hypothesized distribution function $F_0(x)$ and the empirical distribution function of the sample $S_n(x)$: $D_n = \sup_{-\infty < x < \infty} |S_n(x) - F_0(x)|$. If $F_0(x)$ is continuous and under the null hypothesis, the distribution of $D_n$ is independent of $F_0(x)$, i.e. the test is distribution-free. In this paper we introduced a procedure providing the exact critical values of the Kolmogorov-Smirnov test for fixed significance levels. These values are obtained by a modification of the procedure proposed by Feller (1948). In particular, the distribution function of the test statistic is obtained by the solution of a linear system of equations whose coefficients are proper marginal and conditional probabilities. Moreover, a Matlab program provides the computation of the cumulative distribution function’s value of $D_n$ statistic $P(D_n < D)$ for given values of $n$ and $D$.

Keywords: Goodness of fit tests, Percentiles of Kolmogorov-Smirnov’s statistic, Empirical distribution function.

1. INTRODUCTION

The Kolmogorov-Smirnov goodness-of-fit test involves the examination of a random sample from an one-dimensional and continuous random variable, in order to test if the data were really extracted from a hypothesized distribution $F_0(x)$. The test is about the null hypothesis against a generic alternative:

$$\left\{ \begin{array}{ll} H_0 : F(x) = F_0(x) & \text{for every } x \\ H_1 : F(x) \neq F_0(x) & \text{for some } x \end{array} \right. \quad (1)$$

where $F(x)$ is the true cumulative distribution function.

1 Silvia Facchinetti, email: silvia.facchinetti@unicatt.it
Let $X$ be the random variable with the continuous cumulative distribution function

$$F(x) = Pr(X \leq x)$$

and let $(x(1), x(2), \ldots, x(n))$ be the order statistic of the random sample \(\{x_i \sim IID(F), i = 1, 2, \ldots, n\}\), so that \(x(1) \leq x(2) \leq \ldots \leq x(n)\).

The empirical distribution function is defined as follows:

$$S_n(x) = \begin{cases} 
0 & \text{for } x < x(1) \\
k/n & \text{for } x(k) \leq x < x(k+1) \text{ with } k = 1, 2, \ldots, n-1. \\
1 & \text{for } x \geq x(n) 
\end{cases}$$

This is a step function with jumps occurring at the sample values.

Glivenko (1933) and Cantelli (1933), applying the strong law of large numbers, proved that \(S_n(x)\) converges to \(F_0(x)\) under \(H_0\) with probability one as \(n \to \infty\).

In the same year Kolmogorov (1933) introduced the statistic:

$$D_n = \sup_{-\infty < x < \infty} |S_n(x) - F_0(x)|$$

for which the critical region of size \(\alpha\) to reject the null hypothesis in (1) is:

$$R = \left\{ D_n : D_n > D_{\alpha,n} = \frac{d_\alpha}{\sqrt{n}} \right\}$$

where \(d_\alpha\) depends only on \(\alpha\).

Since \(X\) is a continuous random variable, \(D_n\) depends on the null probability integral transformation of the sample values, i.e. \(F_0(x_i)\), and the probability distribution of \(D_n\) is independent of \(F_0(x)\), thus the test is distribution-free.

For large samples the Author found that \(D_n\) has the following limiting distribution:

$$\lim_{n \to \infty} Pr\left(D_n < \frac{d_\alpha}{\sqrt{n}}\right) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-k^2 d_\alpha^2} = L(d_\alpha).$$

Moreover, for \(n \geq 35\), the approximation

$$Pr\left(D_n < \frac{d_\alpha}{\sqrt{n}}\right) \simeq 1 - 2e^{-d_\alpha^2}$$

has been found to be close enough to its limit for practical purposes.

Smirnov (1939 A; 1948) proposed an alternative proof for the limiting distribution, and tabulated the values of the function \(L(d_\alpha)\) in (4). Moreover, the Au-
A Procedure to Find Exact Critical Values of Kolmogorov-Smirnov Test

339

As the original proofs of Kolmogorov and Smirnov are very intricated and are based on different approaches, Feller (1948) presented simplified and unified proofs based on methods of great generality. See also Kendall & Stuart (1967) for a description of the procedure.

Doob (1949) proposed a heuristic approach of a proof based on results concerning the Brownian process and its relation with the Gaussian process.

Besides, a method of evaluating the distribution of \( D_n \) for small samples \( (n \leq 35) \) was proposed by Massey (1950) who obtained a system of recursive formulas for computing \( P(D_n < c/n) \), equivalent with the formulas (14)-(17) proposed by Kolmogorov (1933), as well as a procedure for replacing them with a system of difference equations. A table of percentage points was also given by the same Author (Massey, 1951) for different values of \( \alpha \) and \( n = 1, 2, \ldots, 35 \).

Also Birnbaum (1952) has tabulated \( P(D_n < c/n) \) for \( n = 1, 2, \ldots, 100 \) and \( c = 1, 2, \ldots, 15 \) by a method of computation that involves a truncation of Kolmogorov’s recursive formulas.

Some years later, Miller (1956) introduced some more extensive tables of the percentage points of \( D_n \) distribution by empirical modification of function (4).

Moreover, for \( D_n \) and \( D_n^+ \) Stephens’ modifications (Stephens, 1970) are available for every \( n \) as simple function for the asymptotic percentage points.

For a complete coverage of the history, development, and outstanding problems related to the Kolmogorov-Smirnov statistic, as well as other statistics based on the empirical distribution function, other contributions are worth mentioning.

In particular, Darling (1957) made a review of the goodness of fit tests introduced by Kolmogorov-Smirnov and Cramér-von Mises, and Durbin (1973) summarized and extended the results of numerous authors who had made progress on the problem from 1933 to 1973.
Facchinetti S. (1986), in chapter 4 (due to Stephens), presented a comprehensive coverage on the use of some statistics based on the empirical distribution function.

D’Agostino & Stephens (1986) presented an algorithm for computing the cumulative distribution function of the Kolmogorov-Smirnov test statistic with all parameters known, extending the Birnbaum’s procedure (Birnbaum, 1952) to calculate $P(D_n < D)$ as a spline function. Moreover, Marsaglia, Tsang & Wang (2003) implemented a C procedure that provided the probability $P(D_k < D)$ with great precision and assessed an approximation to limiting form.

Finally, if $X$ is a discontinuous random variable, $D_n$ does not depends on the probability integral transformation of the sample values, and the probability distribution of $D_n$ depends on $F_0(x)$, thus the test is not distribution-free.


3. A PROCEDURE TO CALCULATE THE EXACT CRITICAL VALUES OF KOLMOGOROV-SMIRNOV TEST

Let $X$ be a Uniform random variable on $(0,1)$. The empirical cumulative distribution function $S_n(x)$ may be displayed on the same graph along with the hypothesized cumulative distribution function of $X$, $F_0(x)$, as shown in Figure 1.

In the figure the differences

$$d(x) = S_n(x) - F_0(x) = \frac{k}{n} - x$$

correspond to the vertical deviations between the two functions. Consequently, $D_n$ is the value of the largest absolute vertical difference between them.
A Procedure to Find Exact Critical Values of Kolmogorov-Smirnov Test

Figure 1: Hypothesized cumulative distribution function $F_0(x)$ and empirical cumulative distribution function $S_n(x)$, for a sample size $n = 4$

For a fixed value $0 \leq D_{\alpha,n} = D \leq 1$, the probability

$$F_{D_n}(D) = Pr(D_n \leq D)$$

refers to all samples $(x_1, x_2, \ldots, x_n)$ whose empirical law, for $0 \leq x \leq 1$, is included between the two lines:

$$\begin{align*}
    & y = x + D, 	ext{ upper line } r_1 \\
    & y = x - D, 	ext{ lower line } r_2
\end{align*}$$

which are parallel to $F_0(x) = x$.

If the statistic $D_n$ assumes a value outside the region included between these two lines, the null hypothesis that the true distribution is $F_0(x)$ can be rejected at the $\alpha$ level of significance.

In order to obtain the probability:

$$1 - F_{D_n}(D) = Pr \{ D_n > D \}.$$

we can observe that $D_n$ may be greater than $D$ with respect to the upper or the lower line. In particular, if for a value $x$

$$S_n(x) - F_0(x) > D$$  \hspace{1cm} (7)

this inequality holds for all values of $x$ in the interval $I_k = [x_{(k)}, x_{1k})$ (where $x_{1k}$ is the point of intersection of $S_n(x)$ with $r_1$), at whose upper endpoint $x_{1k}$ we have:

$$S_n(x_{1k}) - F_0(x_{1k}) = D.$$  \hspace{1cm} (8)
Since $F_0(x) = x$, also $F_0(x_{1k}) = x_{1k}$, and the equation (8) becomes:

$$\frac{k}{n} - x_{1k} = D.$$ 

Consequently the inequality (7) holds if and only if for some $k$

$$x_{(k)} < x_{1k} = \frac{k}{n} - D$$

for $k = 0, 1, \ldots, n$ and with $x_{(0)} = 0$.

Similarly, if for a value $x$:

$$S_n(x) - F_0(x) < -D \quad (9)$$

this inequality holds for all values of $x$ in the interval $2I_k = (x_{2k}, x_{(k+1)})$ (where $x_{2k}$ is the point of intersection of $S_n(x)$ with $r_2$), at whose lower endpoint $x_{2k}$ we have:

$$S_n(x_{2k}) - F_0(x_{2k}) = -D. \quad (10)$$

As in this case $F_0(x_{2k}) = x_{2k}$, the equation (10) becomes:

$$\frac{k}{n} - x_{2k} = -D.$$ 

Thus the inequality (9) holds if and only if for some $k$

$$x_{(k+1)} > x_{2k} = \frac{k}{n} + D$$

for $k = 0, 1, \ldots, n$ and with $x_{(n+1)} = 1$.

By denoting the events:

$$\begin{array}{ll}
A_{1k} & \text{if } D_n > D \\
A_{2k} & \text{if } D_n < -D
\end{array}$$

for $k = 0, 1, \ldots, n$, we observe that the statistic $D_n$ will exceed $D$ if and only if at least one of the $2n + 2$ events:

$$A_{10}, A_{20}, A_{11}, A_{21}, A_{12}, A_{22}, \ldots, A_{1n}, A_{2n},$$

(11)

occurs.

Actually, the events $A_{10}$ and $A_{2n}$ are impossible because the overrun of the two lines cannot occur.
Thus we have the formal equivalence of events

\[
\{ D_n > D \} \iff \left\{ \bigcup_{k=0}^{n} A_{1k} \cup \bigcup_{k=0}^{n} A_{2k} \right\}.
\]  

(12)

We must be aware that the possible events are only those that occur inside the unit square, i.e. \(0 < x_{ik} < 1\), for \(i = 1, 2\) and \(k = 0, 1, \ldots, n\). As a consequence, the following conditions must be satisfied:

- for the upper line: \(x_{1k} > 0 \iff (k - nD)/n > 0 \iff k > nD\), thus the minimum value of \(k\) is:
  \[m_1 = \lfloor nD \rfloor + 1\]
  where \(\lfloor nD \rfloor = \text{int}(nD)\), hence \(k = m_1, m_1 + 1, \ldots, n\);
- for the lower line: \(x_{2k} < 1 \iff (k + nD)/n < 1 \iff k < n - nD\), thus the maximum value of \(k\) is:
  \[m_2 = n - (\lfloor nD \rfloor + 1)\]
  where \(\lfloor nD \rfloor = \text{int}(nD)\), hence \(k = 0, 1, \ldots, m_2\).

Summarizing:

\[
\begin{align*}
0 < x_{1k} < 1 & \iff k = m_1, m_1 + 1, \ldots, n \\
0 < x_{2k} < 1 & \iff k = 0, 1, \ldots, m_2
\end{align*}
\]

with \(m_1 + m_2 = n\).

The events \(A_{1k}\) and \(A_{2k}\) are defined on the two distinct sets:

\[
\begin{align*}
A_{1k} & \text{ for } k = m_1, m_1 + 1, \ldots, n \\
A_{2k} & \text{ for } k = 0, 1, \ldots, m_2.
\end{align*}
\]

(13)

Since the union extended to impossible events does not alter the final results, we have the equivalence of the events:

\[
\{ D_n > D \} \iff \left\{ \bigcup_{k=0}^{n} A_{1k} \cup \bigcup_{k=0}^{m_2} A_{2k} \right\} \iff \left\{ \bigcup_{k=m_1}^{n} A_{1k} \cup \bigcup_{k=0}^{m_2} A_{2k} \right\}
\]  

(14)

Then it is possible to define the \(2n + 2\) mutually exclusive events \(U_r \subset A_{1r}\) and \(V_r \subset A_{2r}\), with \(r \leq k\) such that:

- \(U_r\) occurs if \(A_{1r}\) is the first event in the sequence (11), for \(r = 0, 1, \ldots, n\);
• \( V_r \) occurs if \( A_{2r} \) is the first event in the sequence (11), for \( r = 0, 1, \ldots, n \); therefore the event

\[
\left[ \bigcup_{r=0}^{n} U_r \right] \cup \left[ \bigcup_{r=0}^{n} V_r \right]
\]

is equivalent to the one in (14).

The events \( U_r \) and \( V_r \) are mutually exclusive, hence

\[
Pr\{D_n > D\} = \sum_{r=0}^{n} [Pr\{U_r\} + Pr\{V_r\}].
\]

From the definitions of \( A_{1k}, A_{2k}, U_r \) and \( V_r \) the following relations hold:

\[
\begin{align*}
Pr\{A_{1k}\} &= \sum_{r=0}^{k} [Pr\{U_r\} Pr\{A_{1k}|A_{1r}\} + Pr\{V_r\} Pr\{A_{1k}|A_{2r}\}] \\
Pr\{A_{2k}\} &= \sum_{r=0}^{k} [Pr\{U_r\} Pr\{A_{2k}|A_{1r}\} + Pr\{V_r\} Pr\{A_{2k}|A_{2r}\}]
\end{align*}
\]

(16)

where

• \( Pr\{A_{1k}\} \) for \( t = 1, 2 \) are the marginal probabilities, i.e. the probabilities of overtaking one of the two lines \( r_1 \) or \( r_2 \);

• \( Pr\{A_{1k}|A_{sr}\} \) for \( t = s = 1, 2 \) are the conditional probabilities, i.e. the probabilities of overtaking one of the two lines at level \( k \), conditionally on the same event at level \( r \), with \( r < k \);

• \( Pr\{U_r\} \) and \( Pr\{V_r\} \) are the probabilities that in the sequence (11) the first event to occur is \( A_{1r} \) or \( A_{2r} \), respectively.

The equation (16) defines a system of \( 2n + 2 \) linear equations for the \( 2n + 2 \) unknowns \( Pr\{U_r\} \) and \( Pr\{V_r\} \). After solving the system, and substituting into (15), we can obtain \( Pr\{D_n > D\} \).

4. MARGINAL AND CONDITIONAL PROBABILITIES

Now we have to compute the marginal and the conditional probabilities.

For the marginal probabilities from (13) we know that:

\[
C_{1k} = Pr\{A_{1k}\} \begin{cases} = 0, & \text{for } k = 0, 1, \ldots, m_1 - 1 \\ > 0, & \text{for } k = m_1, m_1 + 1, \ldots, n \end{cases}
\]

and

\[
C_{2k} = Pr\{A_{2k}\} \begin{cases} > 0, & \text{for } k = 0, 1, \ldots, m_2 \\ = 0, & \text{for } k = m_2 + 1, m_2 + 2, \ldots, n. \end{cases}
\]
In particular we see that \( C_{1k} \) is the probability that exactly \( k \) successes occur in \( n \) Binomial trials with probability

\[
p_{1k} = x_{1k} = F(x_{1k}) = \left( \frac{k}{n} - D \right),
\]

thus:

\[
C_{1k} = \frac{n!}{k!(n-k)!} \left( \frac{k-nD}{n} \right)^k \left( \frac{n-k+nD}{n} \right)^{n-k}
\]

for \( k = m_1, m_1 + 1, \ldots, n \).

Similarly \( C_{2k} \) is the probability that exactly \( k \) successes occur in \( n \) Binomial trials with probability

\[
p_{2k} = x_{2k} = F(x_{2k}) = \left( \frac{k}{n} + D \right),
\]

thus:

\[
C_{2k} = \frac{n!}{k!(n-k)!} \left( \frac{k+nD}{n} \right)^k \left( \frac{n-k-nD}{n} \right)^{n-k}
\]

for \( k = 0, 1, \ldots, m_2 \).

We observe that \( C_{1k} \) and \( C_{2k} \) depend only on \( k, n \) and \( D \).

For varying \( k \), \( C_{1k} \) and \( C_{2k} \) become the elements of the two vectors \( C_1 \) and \( C_2 \) of order \((1 \times (n+1))\) which together define the vector \( C_{(1 \times (2n+2))} \) of marginal probabilities:

\[
C = \left[ \begin{array}{c} C_1 \\ C_2 \end{array} \right].
\]

Now we define the conditional events:

\[
\begin{align*}
A_{1k} | A_{1r}, & \quad \text{for } k = m_1, \ldots, n \text{ and } r = m_1, \ldots, n \\
A_{2k} | A_{1r}, & \quad \text{for } k = 0, \ldots, m_2 \text{ and } r = m_1, \ldots, n \\
A_{1k} | A_{2r}, & \quad \text{for } k = m_1, \ldots, n \text{ and } r = 0, \ldots, m_2 \\
A_{2k} | A_{2r}, & \quad \text{for } k = 0, \ldots, m_2 \text{ and } r = 0, \ldots, m_2.
\end{align*}
\]

(17)

In order to consider these events as consequent, the following relations can be verified simultaneously:

\[
\begin{align*}
& x_{1k} \geq x_{sr}, \quad \text{for } t, s = 1, 2 \\
& k \geq r.
\end{align*}
\]

Let us consider separately the four events:
1. $t = s = 1 \Rightarrow A_{1k}|A_{1r}$.
   The indexes $k$ and $r$ must verify the inequalities:
   \[ m_1 \leq r \leq k \leq n. \] (18)

2. $t = 2, s = 1 \Rightarrow A_{2k}|A_{1r}$.
   The indexes $k$ and $r$ must verify the inequalities:
   \[ m_1 \leq r \leq k \leq m_2. \] (19)

3. $t = 1, s = 2 \Rightarrow A_{1k}|A_{2r}$.
   The indexes $k$ and $r$ must verify the inequalities:
   \[ r + 2nD \leq k \leq n \] (20)

4. $t = s = 2 \Rightarrow A_{2k}|A_{2r}$.
   The indexes $k$ and $r$ must verify the inequalities:
   \[ 0 \leq r \leq k \leq m_2. \] (21)

Now we want to evaluate the probabilities of the consequent events $A_{ik}|A_{sr}$ for $(t, s = 1, 2)$.
In particular we see that also these probabilities are defined by a Binomial expression:
\[ tsb_{kr} = \Pr\{A_{ik}|A_{sr}\} = \frac{(n-r)!}{(k-r)!(k-n)!} \left( \frac{x_{ik} - x_{sr}}{1-x_{sr}} \right)^{k-r} \left( \frac{1-x_{ik}}{1-x_{sr}} \right)^{n-k} \] (22)
for $t, s = 1, 2$ and with respect to (17).
In particular:

1. $t = s = 1 \Rightarrow B_{11} = (11b_{kr})$.
   Replacing $t = s = 1$ in (22) we have that $\Pr\{A_{1k}|A_{1r}\}$ is:
   \[ 11b_{kr} = \frac{(n-r)!}{(k-r)!(n-k)!} \left( \frac{k-r}{n_1-r} \right)^{k-r} \left( \frac{n_1-k}{n_1-r} \right)^{n-k} \]
   for $m_1 \leq r \leq k \leq n$, with:
   \[ \begin{cases} n_1 = n(1+D) \\ n_2 = n(1-D). \end{cases} \]
As \( k \geq r \), we obtain a lower triangular matrix of order \((n + 1)\), and in particular, for \( k = r \) the diagonal terms are all equal to one.

From (18) the number of probabilities to be defined is

\[
\frac{(n - m_1)(n - m_1 + 1)}{2} = \frac{(m_2)(m_2 + 1)}{2}.
\]

Thus we have the matrix \( B_{11} \) having the following framework:

![Figure 2: Framework of the matrix \( B_{11} \)](image)

2. \( t = 2, s = 1 \Rightarrow B_{21} = (21bkr) \).

Replacing \( t = 2, s = 1 \) in (22) we have that \( Pr\{A_{2k}|A_{1r}\} \) is:

\[
21bkr = \frac{(n-r)!}{(k-r)!(n-k)!} \left( \frac{k-r+2nD}{n_1-r} \right)^{k-r} \left( \frac{n_2-k}{n_1-r} \right)^{n-k}
\]

for \( m_1 \leq r \leq k \leq m_2 = n - m_1 \).

We obtain a lower triangular matrix of order \((n + 1)\), and for (19) the number of probabilities to be defined is

\[
\frac{(m_2 - m_1 + 1)(m_2 - m_1 + 2)}{2} = \frac{(n - 2m_1 + 1)(n - 2m_1 + 2)}{2}.
\]

Thus we have the matrix \( B_{21} \) having the following framework:
3. \( t = 1, s = 2 \Rightarrow B_{12} = (12b_{kr}) \).

Replacing \( t = 1, s = 2 \) in (22) we have that \( Pr \{A_{1k}|A_{2r}\} \) is:

\[
12b_{kr} = \frac{(n-r)!}{(k-r)!(n-k)!} \left( \frac{k-r-2nD}{n_2-r} \right)^{k-r} \left( \frac{n_1-k}{n_2-r} \right)^{n-k}
\]

for \( 0 \leq r \leq n - 2nD \) and \( r + 2nD \leq k \leq n \).

We obtain a lower triangular matrix of order \((n+1)\), and for (20) the number of probabilities to be defined is

\[
\frac{(n-l_1+1)(n-l_1+2)}{2},
\]

where \( l_1 = \text{int}(2nD+1) \).

Thus we have the matrix \( B_{12} \) having the following framework:
4. \( t = s = 2 \Rightarrow B_{22} = (22b_{kr}) \).

Replacing \( t = s = 2 \) in (22) we have that \( Pr \{ A_{2k} | A_{2r} \} \) is:

\[
22b_{kr} = \frac{(n-r)!}{(k-r)!(n-k)!} \left( \frac{k-r}{n2-r} \right)^{k-r} \left( \frac{n2-k}{n2-r} \right)^{n-k}
\]

for \( 0 \leq r \leq k \leq m_2 \).

As \( k \geq r \), we obtain a lower triangular matrix of order \((n+1)\), and in particular, for \( k = r \) the diagonal terms are all equal to one. For (21) the number of probabilities to be defined is

\[
\frac{(m_2)(m_2 + 1)}{2}.
\]

Thus we have the matrix \( B_{22} \) having the following framework:

![Figure 5: Framework of the matrix \( B_{22} \)](image)

Combining the previous four matrices we have that the matrix \( B \) of the conditional probabilities is the square block matrix of order \((2n + 2)\):

\[
B = \begin{bmatrix}
B_{11} & B_{12} \\
B_{21} & B_{22}
\end{bmatrix}
\]

with the following framework:
5. DISTRIBUTION FUNCTION OF $D_n$ STATISTIC

From above, (16) is a set of $2n + 2$ linear equations for $2n + 2$ unknowns:

$$W_r = Pr\{U_r\}; Y_r = Pr\{V_r\}.$$  

For varying $r$, $W_r$ and $Y_r$ constitute the elements of the two vectors:

$$W = \{W_r\}$$

and

$$Y = \{Y_r\}$$

which together define the vector

$$Z = \begin{bmatrix} W \\ Y \end{bmatrix}$$

of order $(1 \times (2n + 2))$.

Consequently we can rewrite the system (16) as follows:

$$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \cdot \begin{bmatrix} W \\ Y \end{bmatrix}$$
or:

\[ C = B \cdot Z. \]

As the matrix \( B \) is singular, we cannot calculate its inverse \( B^{-1} \), thus the system is indeterminate. To solve this problem we calculate the Moore-Penrose pseudo-inverse matrix \( B^+ \) instead of \( B^{-1} \) (Gentle, 2007).

In this way we calculate the probabilities \( W_r \) and \( Y_r \) such that:

\[
Pr\{D_n > D\} = \sum_{r=0}^{n} [Pr\{U_r\} + Pr\{V_r\}].
\] (23)

From the previous equation we obtain the values of the distribution function of the Kolmogorov-Smirnov statistic \( D_n \):

\[
F_{D_n}(D) = Pr\{D_n \leq D\}.
\] (24)

The cumulative distribution function of \( D_n \) is shown for different values of \( n \) in Figure 7.

![Figure 7: Cumulative distribution function of \( D_n \) statistic](image)
For a fixed significance level $\alpha$, from (24) we calculate the critical values $D_{\alpha,n}^*$ of the Kolmogorov-Smirnov test. Table 1 gives many critical values for various sample sizes and significance levels.

Table 1: Exact critical values of Kolmogorov-Smirnov statistic obtained by the proposed procedure

<table>
<thead>
<tr>
<th>n</th>
<th>0.001</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.97764</td>
<td>0.92930</td>
<td>0.84189</td>
<td>0.77639</td>
<td>0.72614</td>
<td>0.68377</td>
</tr>
<tr>
<td>3</td>
<td>0.92063</td>
<td>0.82900</td>
<td>0.70760</td>
<td>0.63604</td>
<td>0.59582</td>
<td>0.56481</td>
</tr>
<tr>
<td>4</td>
<td>0.85046</td>
<td>0.73421</td>
<td>0.62394</td>
<td>0.56522</td>
<td>0.52476</td>
<td>0.49265</td>
</tr>
<tr>
<td>5</td>
<td>0.78137</td>
<td>0.66855</td>
<td>0.56327</td>
<td>0.50945</td>
<td>0.47439</td>
<td>0.44697</td>
</tr>
<tr>
<td>6</td>
<td>0.72479</td>
<td>0.61660</td>
<td>0.51926</td>
<td>0.46799</td>
<td>0.43526</td>
<td>0.41035</td>
</tr>
<tr>
<td>7</td>
<td>0.67930</td>
<td>0.57580</td>
<td>0.48343</td>
<td>0.43607</td>
<td>0.40497</td>
<td>0.38145</td>
</tr>
<tr>
<td>8</td>
<td>0.64098</td>
<td>0.54180</td>
<td>0.45427</td>
<td>0.40962</td>
<td>0.38062</td>
<td>0.35828</td>
</tr>
<tr>
<td>9</td>
<td>0.60846</td>
<td>0.51330</td>
<td>0.43001</td>
<td>0.38746</td>
<td>0.36006</td>
<td>0.33907</td>
</tr>
<tr>
<td>10</td>
<td>0.58042</td>
<td>0.48895</td>
<td>0.40925</td>
<td>0.36866</td>
<td>0.34250</td>
<td>0.32257</td>
</tr>
<tr>
<td>11</td>
<td>0.55588</td>
<td>0.46770</td>
<td>0.39122</td>
<td>0.35242</td>
<td>0.32734</td>
<td>0.30826</td>
</tr>
<tr>
<td>12</td>
<td>0.53422</td>
<td>0.44905</td>
<td>0.37543</td>
<td>0.33815</td>
<td>0.31408</td>
<td>0.29573</td>
</tr>
<tr>
<td>13</td>
<td>0.51490</td>
<td>0.43246</td>
<td>0.36143</td>
<td>0.32548</td>
<td>0.30233</td>
<td>0.28466</td>
</tr>
<tr>
<td>14</td>
<td>0.49753</td>
<td>0.41760</td>
<td>0.34890</td>
<td>0.31417</td>
<td>0.29181</td>
<td>0.27477</td>
</tr>
<tr>
<td>15</td>
<td>0.48182</td>
<td>0.40420</td>
<td>0.33760</td>
<td>0.30397</td>
<td>0.28233</td>
<td>0.26585</td>
</tr>
<tr>
<td>16</td>
<td>0.46750</td>
<td>0.39200</td>
<td>0.32733</td>
<td>0.29471</td>
<td>0.27372</td>
<td>0.25774</td>
</tr>
<tr>
<td>17</td>
<td>0.45440</td>
<td>0.38085</td>
<td>0.31796</td>
<td>0.28627</td>
<td>0.26587</td>
<td>0.25035</td>
</tr>
<tr>
<td>18</td>
<td>0.44234</td>
<td>0.37063</td>
<td>0.30936</td>
<td>0.27851</td>
<td>0.25867</td>
<td>0.24356</td>
</tr>
<tr>
<td>19</td>
<td>0.43119</td>
<td>0.36116</td>
<td>0.30142</td>
<td>0.27135</td>
<td>0.25202</td>
<td>0.23731</td>
</tr>
<tr>
<td>20</td>
<td>0.42085</td>
<td>0.35240</td>
<td>0.29407</td>
<td>0.26473</td>
<td>0.24587</td>
<td>0.23152</td>
</tr>
<tr>
<td>25</td>
<td>0.37843</td>
<td>0.31656</td>
<td>0.26404</td>
<td>0.23767</td>
<td>0.22074</td>
<td>0.20786</td>
</tr>
<tr>
<td>30</td>
<td>0.34672</td>
<td>0.28988</td>
<td>0.24170</td>
<td>0.21756</td>
<td>0.20207</td>
<td>0.19029</td>
</tr>
<tr>
<td>35</td>
<td>0.32187</td>
<td>0.26898</td>
<td>0.22424</td>
<td>0.20184</td>
<td>0.18748</td>
<td>0.17655</td>
</tr>
</tbody>
</table>

For example, in 10% of the random samples of size 15, the maximum absolute deviation between the empirical distribution function and the theoretical distribution function will be at least 0.30397.
A Procedure to Find Exact Critical Values of Kolmogorov-Smirnov Test

Table 2 gives the critical values $d_{\alpha}(n)$ tabulated by Massey (1951) and integrated by Birnbaum (1952).

<table>
<thead>
<tr>
<th>Significance level ($\alpha$)</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.929</td>
<td>0.842</td>
<td>0.776</td>
<td>0.726</td>
<td>0.684</td>
</tr>
<tr>
<td>3</td>
<td>0.829</td>
<td>0.708</td>
<td>0.642</td>
<td>0.597</td>
<td>0.565</td>
</tr>
<tr>
<td>4</td>
<td>0.734</td>
<td>0.624</td>
<td>0.564</td>
<td>0.525</td>
<td>0.494</td>
</tr>
<tr>
<td>5</td>
<td>0.669</td>
<td>0.563</td>
<td>0.510</td>
<td>0.474</td>
<td>0.446</td>
</tr>
<tr>
<td>6</td>
<td>0.618</td>
<td>0.521</td>
<td>0.470</td>
<td>0.436</td>
<td>0.410</td>
</tr>
<tr>
<td>7</td>
<td>0.577</td>
<td>0.486</td>
<td>0.438</td>
<td>0.405</td>
<td>0.381</td>
</tr>
<tr>
<td>8</td>
<td>0.543</td>
<td>0.457</td>
<td>0.411</td>
<td>0.381</td>
<td>0.358</td>
</tr>
<tr>
<td>9</td>
<td>0.514</td>
<td>0.432</td>
<td>0.388</td>
<td>0.360</td>
<td>0.339</td>
</tr>
<tr>
<td>10</td>
<td>0.486</td>
<td>0.409</td>
<td>0.368</td>
<td>0.342</td>
<td>0.322</td>
</tr>
<tr>
<td>11</td>
<td>0.468</td>
<td>0.391</td>
<td>0.352</td>
<td>0.326</td>
<td>0.307</td>
</tr>
<tr>
<td>12</td>
<td>0.450</td>
<td>0.375</td>
<td>0.338</td>
<td>0.313</td>
<td>0.295</td>
</tr>
<tr>
<td>13</td>
<td>0.433</td>
<td>0.361</td>
<td>0.325</td>
<td>0.302</td>
<td>0.284</td>
</tr>
<tr>
<td>14</td>
<td>0.418</td>
<td>0.349</td>
<td>0.314</td>
<td>0.292</td>
<td>0.274</td>
</tr>
<tr>
<td>15</td>
<td>0.404</td>
<td>0.338</td>
<td>0.304</td>
<td>0.283</td>
<td>0.266</td>
</tr>
<tr>
<td>16</td>
<td>0.391</td>
<td>0.328</td>
<td>0.295</td>
<td>0.274</td>
<td>0.258</td>
</tr>
<tr>
<td>17</td>
<td>0.380</td>
<td>0.318</td>
<td>0.286</td>
<td>0.266</td>
<td>0.250</td>
</tr>
<tr>
<td>18</td>
<td>0.370</td>
<td>0.309</td>
<td>0.278</td>
<td>0.259</td>
<td>0.244</td>
</tr>
<tr>
<td>19</td>
<td>0.361</td>
<td>0.301</td>
<td>0.272</td>
<td>0.252</td>
<td>0.237</td>
</tr>
<tr>
<td>20</td>
<td>0.352</td>
<td>0.294</td>
<td>0.264</td>
<td>0.246</td>
<td>0.231</td>
</tr>
<tr>
<td>25</td>
<td>0.320</td>
<td>0.264</td>
<td>0.240</td>
<td>0.220</td>
<td>0.210</td>
</tr>
<tr>
<td>30</td>
<td>0.290</td>
<td>0.242</td>
<td>0.220</td>
<td>0.200</td>
<td>0.190</td>
</tr>
<tr>
<td>35</td>
<td>0.270</td>
<td>0.230</td>
<td>0.210</td>
<td>0.190</td>
<td>0.180</td>
</tr>
</tbody>
</table>

Comparing Table 1 and Table 2 we observe the closeness of the values obtained by the proposed procedure with those given by Massey and Birnbaum.
6. CONCLUSIONS

To allow a synthetic comparison between the critical values in Tables 1 and 2, Table 3 gives the percentage differences

\[
\frac{D_{\alpha,n} - d_{\alpha}(n)}{d_{\alpha}(n)}
\]

based on the values given by Massey and Birnbaum.

<table>
<thead>
<tr>
<th>n</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.05</td>
<td>0.02</td>
<td>-0.03</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>-0.06</td>
<td>-0.93</td>
<td>-0.20</td>
<td>-0.03</td>
</tr>
<tr>
<td>4</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.22</td>
<td>-0.05</td>
<td>-0.27</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>0.05</td>
<td>-0.11</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>6</td>
<td>-0.23</td>
<td>-0.33</td>
<td>-0.43</td>
<td>-0.17</td>
<td>0.09</td>
</tr>
<tr>
<td>7</td>
<td>-0.21</td>
<td>-0.53</td>
<td>-0.44</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>-0.22</td>
<td>-0.60</td>
<td>-0.34</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>-0.14</td>
<td>-0.46</td>
<td>-0.14</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>0.61</td>
<td>0.06</td>
<td>0.18</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td>11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>12</td>
<td>-0.21</td>
<td>0.00</td>
<td>0.00</td>
<td>0.35</td>
<td>0.25</td>
</tr>
<tr>
<td>13</td>
<td>-0.12</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.23</td>
</tr>
<tr>
<td>14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.28</td>
</tr>
<tr>
<td>15</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.24</td>
<td>0.00</td>
</tr>
<tr>
<td>16</td>
<td>0.26</td>
<td>-0.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>17</td>
<td>0.22</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>18</td>
<td>0.17</td>
<td>0.00</td>
<td>0.18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>19</td>
<td>0.04</td>
<td>0.00</td>
<td>-0.24</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>20</td>
<td>0.11</td>
<td>0.00</td>
<td>0.28</td>
<td>0.00</td>
<td>0.23</td>
</tr>
<tr>
<td>25</td>
<td>-1.08</td>
<td>0.02</td>
<td>-0.97</td>
<td>0.34</td>
<td>-1.02</td>
</tr>
<tr>
<td>30</td>
<td>0.00</td>
<td>-0.12</td>
<td>-1.11</td>
<td>1.03</td>
<td>0.00</td>
</tr>
<tr>
<td>35</td>
<td>-0.38</td>
<td>-2.50</td>
<td>-3.89</td>
<td>-1.33</td>
<td>-1.92</td>
</tr>
</tbody>
</table>

From Table 3 we observe that the minimum and the maximum percentage differences are respectively \(-3.88571\%\) (in Table we see the value \(-3.89\%\) approximated to two decimal places), and \(1.03500\%\) (in Table we see the value \(1.03\%\) approximated to two decimal places).
Being these percentage differences less of four percentage points, we confirm that there is no difference in the use of both methodologies for calculating the critical values of the test.

The values in Table 1 were computed for small sample sizes \(n \leq 35\). Those for \(n > 35\) are obtained from Smirnov’s table (Smirnov, 1948) by relating the values \(d_\alpha\) with \(\sqrt{n}\), and are reported in Table 4.

### Table 4: Asymptotic critical values \(d_\alpha\) \((n > 35)\) of Kolmogorov-Smirnov statistic given by Smirnov (1948)

<table>
<thead>
<tr>
<th>Significance level ((\alpha))</th>
<th>(n)</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 35</td>
<td>(\sqrt{n})</td>
<td>(1.63/\sqrt{n})</td>
<td>(1.36/\sqrt{n})</td>
<td>(1.22/\sqrt{n})</td>
<td>(1.14/\sqrt{n})</td>
<td>(1.07/\sqrt{n})</td>
</tr>
</tbody>
</table>

Table 5 gives the critical values \(\sqrt{n}D_{\alpha,n}^*\) for large sample sizes \(n = 50; 80; 100\) obtained by the proposed procedure.

### Table 5: Asymptotic critical values \(\sqrt{n}D_{\alpha,n}^*\) of Kolmogorov-Smirnov statistic obtained by the proposed procedure

<table>
<thead>
<tr>
<th>Significance level ((\alpha))</th>
<th>(n)</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>(1.59834)</td>
<td>(1.33014)</td>
<td>(1.19918)</td>
<td>(1.11391)</td>
<td>(1.04913)</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>(1.60532)</td>
<td>(1.33806)</td>
<td>(1.20453)</td>
<td>(1.11902)</td>
<td>(1.05408)</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>(1.60808)</td>
<td>(1.34028)</td>
<td>(1.20663)</td>
<td>(1.12105)</td>
<td>(1.05600)</td>
<td></td>
</tr>
</tbody>
</table>

In Tables 4 and 5 we observe that the differences between the values are from the second decimal place on, and for \(n \rightarrow \infty\) the calculated values tend to approach Smirnov values. Reasonably these differences are due to the different type of approximation considered.

### 7. APPENDIX: A MATLAB PROGRAM FOR \(P(D_n \leq D)\)

The following Matlab program contains a procedure that provides the values of the cumulative distribution function of \(D_n\) statistic \(P(D_n \leq D)\), given the values of \(n\) and \(D\). The program is implemented following the procedure described in this paper as the solution of a linear system of equations whose coefficients are proper marginal and conditional probabilities.
clear all
% insert values n’ for n and D’ for D
n=n’;
D=D’;
% parameters definition
nD=n*D;
m1=round(n*D+0.5);
m2=round(n-n*D-0.5);
l1=round(2*n*D+0.5);
n1=n*(1+D);
n2=n*(1-D);
% B matrix
B=zeros(2*(n+1));
% B11 matrix
for k=m1+1:n+1
  B(k,k)=1;
end
for r=m1:n-1
  for k=r+1:n
    B(k+1,r+1)=factorial(n-r)/(factorial(k-r)*factorial(n-k))*(((k-r)/(n1-r))^(k-r))*(((n1-k)/(n1-r))^(n-k));
  end
end
% B22 matrix
for k=n+2:n+2+m2
  B(k,k)=1;
end
for r=0:m2-1
  for k=r+1:m2
    B(k+n+2,r+n+2)=factorial(n-r)/(factorial(k-r)*factorial(n-k))*(((k-r)/(n2-r))^(k-r))*(((n2-k)/(n2-r))^(n-k));
  end
end
% B21 matrix
for r=m1:m2
  for k=r:m2
    B(k+n+2,r+1)=factorial(n-r)/(factorial(k-r)*factorial(n-k))*
A Procedure to Find Exact Critical Values of Kolmogorov-Smirnov Test

\[
(((k-r+2nD)/(n1-r))^{(k-r)})*(((n2-k)/(n1-r))^{(n-k)});
\]
end
dead
% B12 matrix
for r=0:n-1
for k=r:n+1
B(k+1,r+n+2)=factorial(n-r)/(factorial(k-r)*factorial(n-k))*
(((k-r-2nD)/(n2-r))^{(k-r)})*(((n1-k)/(n2-r))^{(n-k)});
end
end
% C vector
C=zeros(2*(n+1),1);
% C1 vector
for k=m1:n
C(k+1)=factorial(n)/(factorial(k)*factorial(n-k))*
(((k-nD)/n)^k)*(((n1-k)/n)^{(n-k)});
end
% C2 vector
for k=0:m2
C(n+2+k)=factorial(n)/(factorial(k)*factorial(n-k))*
(((k+nD)/n)^k)*(((n2-k)/n)^{(n-k)});
end
% system solution
Binv=pinv(B);
Z=Binv*C;
alpha=sum(Z);
cdf=1-alpha;

Acknowledgements: The author wishes to thank Prof. B.V. Frosini and Prof. U. Magagnoli for the supervision of this work.

References


A Procedure to Find Exact Critical Values of Kolmogorov-Smirnov Test


