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

Silvia Facchinetti[1]

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 A) 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:

$$\begin{cases} H_0 : F(x) = F_0(x) & \text{for every } x \\ H_1 : F(x) \neq F_0(x) & \text{for some } x \end{cases} \tag{1}$$

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:

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

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:

$$D_n = \sup_{-\infty < x < \infty} |S_n(x) - F_0(x)| \tag{3}$$

for which the critical region of size α 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_α depends only on α.

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:

$$\lim_{n \to \infty} Pr\left(D_n < \frac{d_\alpha}{\sqrt{n}} \right) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-k^2 d_\alpha^2} = L(d_\alpha). \tag{4}$$

Moreover, for $n \geq 35$, the approximation

$$Pr\left(D_n < \frac{d_\alpha}{\sqrt{n}} \right) \simeq 1 - 2e^{-d_\alpha^2} \tag{5}$$

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-

thor (Smirnov, 1939 A, 1944) suggested an asymptotic distribution of the one-sided statistic

$$D_n^+ = \sup_{-\infty<x<\infty} \{S_n(x) - F_0(x)\} \tag{6}$$

and another one regarding the maximum difference between the empirical distribution of two samples with the same cumulative distribution function. The proof is given in Smirnov (1939 B).

2. LITERATURE REVIEW

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 α and n = 1, 2, …, 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.

Frosini (1978) studied the several related statistics by examining the differences between the distribution curves, as the graduation curves; the Author presented an outline concerning inferential applications of goodness of fit statistics when the null hypothesis is composite and about comparison of powers of several tests.

D'Agostino & Stephens (1986), in chapter 4 (due to Stephens), presented a comprehensive coverage on the use of some statistics based on the empirical distribution function.

Regarding the development of computational procedures, Drew, Glen & Leemis (2000) 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 $P(D_n < D)$ 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.

More details on the application of the Kolmogorov-Smirnov test for discontinuous distribution functions are given in Kolmogorov (1941), Schmid (1958), Noether (1963), Conover (1972; 1999), Pettitt & Stephens (1977), Wood & Altavela (1978), Jalla (1979), Marvulli (1980), Facchinetti & Chiodini (2008), Facchinetti & Osmetti (2009).

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.

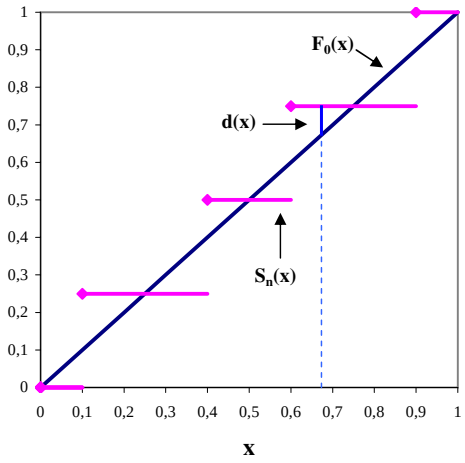

Figure 1: Hypothesized cumulative distribution function $F_0(x)$ and empirical cumulative distribution function $S_n(x)$, for a sample size $n = 4$

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:

$$\begin{cases} y = x + D & \text{upper line } r_1 \\ y = x - D & \text{lower line } r_2 \end{cases}$$

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 α 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

$$S_n(x) - F_0(x) > D \tag{7}$$

this inequality holds for all values of x in the interval ${}_1I_k = [x_{(k)}, x_{1k})$ (where x_{1k} is the point of intersection of $S_n(x)$ with r_1), at whose upper endpoint x_{1k} we have:

$$S_n(x_{1k}) - F_0(x_{1k}) = D. \tag{8}$$

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:

$$S_n(x) - F_0(x) < -D \tag{9}$$

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:

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

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:

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

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:

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

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

$$\{D_n > D\} \Longleftrightarrow \left\{ \left[\bigcup_{k=0}^{n} A_{1k} \right] \cup \left[\bigcup_{k=0}^{n} A_{2k} \right] \right\}. \tag{12}$$

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

Summarizing:

$$\begin{cases} 0 < x_{1k} < 1 & \Leftrightarrow k = m_1, m_1 + 1, \ldots, n \\ 0 < x_{2k} < 1 & \Leftrightarrow k = 0,1,\ldots,m_2 \end{cases}$$

with $m_1 + m_2 = n$.

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

$$\begin{cases} A_{1k} & \text{for } k = m_1, m_1 + 1, \ldots, n \\ A_{2k} & \text{for } k = 0,1,\ldots,m_2. \end{cases} \tag{13}$$

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

$$\{D_n > D\} \Longleftrightarrow \left\{ \left[\bigcup_{k=0}^{n} A_{1k} \right] \cup \left[\bigcup_{k=0}^{n} A_{2k} \right] \right\} \Longleftrightarrow \left\{ \left[\bigcup_{k=m_1}^{n} A_{1k} \right] \cup \left[\bigcup_{k=0}^{m_2} A_{2k} \right] \right\} \tag{14}$$

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

$$Pr\{D_n > D\} = \sum_{r=0}^{n} \left[Pr\{U_r\} + Pr\{V_r\}\right]. \tag{15}$$

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

$$\begin{cases} Pr\{A_{1k}\} = \sum_{r=0}^{k} \left[Pr\{U_r\}\, Pr\{A_{1k}|A_{1r}\} + Pr\{V_r\}\, Pr\{A_{1k}|A_{2r}\}\right] \\ Pr\{A_{2k}\} = \sum_{r=0}^{k} \left[Pr\{U_r\}\, Pr\{A_{2k}|A_{1r}\} + Pr\{V_r\}\, Pr\{A_{2k}|A_{2r}\}\right] \end{cases} \tag{16}$$

where

- $Pr\{A_{tk}\}$ for $t = 1, 2$ are the marginal probabilities, i.e. the probabilities of overtaking one of the two lines r_1 or r_2;
- $Pr\{A_{tk}|A_{sr}\}$ 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 $\underline{C}_1$ and $\underline{C}_2$ of order $(1 \times (n+1))$ which together define the vector $\underline{C}_{(1\times(2n+2))}$ of marginal probabilities:

$$\underline{C} = \begin{bmatrix} \underline{C}_1 \\ \underline{C}_2 \end{bmatrix}.$$

Now we define the conditional events:

$$\begin{cases} A_{1k}|A_{1r}, & \text{for } k = m_1, \ldots, n \text{ and } r = m_1, \ldots, n \\ A_{2k}|A_{1r}, & \text{for } k = 0, \ldots, m_2 \text{ and } r = m_1, \ldots, n \\ A_{1k}|A_{2r}, & \text{for } k = m_1, \ldots, n \text{ and } r = 0, \ldots, m_2 \\ A_{2k}|A_{2r}, & \text{for } k = 0, \ldots, m_2 \text{ and } r = 0, \ldots, m_2. \end{cases} \tag{17}$$

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

$$\begin{cases} x_{tk} \geq x_{sr}, & \text{for } t, s = 1, 2 \\ k \geq r. \end{cases}$$

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:

$$m_1 \leq r \leq k \leq n. \tag{18}$$

2. $t = 2, s = 1 \Rightarrow A_{2k}|A_{1r}$.
   The indexes k and r must verify the inequalities:

$$m_1 \leq r \leq k \leq m_2. \tag{19}$$

3. $t = 1, s = 2 \Rightarrow A_{1k}|A_{2r}$.
   The indexes k and r must verify the inequalities:

$$r + 2nD \leq k \leq n \tag{20}$$

4. $t = s = 2 \Rightarrow A_{2k}|A_{2r}$.
   The indexes k and r must verify the inequalities:

$$0 \leq r \leq k \leq m_2. \tag{21}$$

Now we want to evaluate the probabilities of the consequent events $A_{tk}|A_{sr}$ for $(t, s = 1, 2)$.

In particular we see that also these probabilities are defined by a Binomial expression:

$$_{ts}b_{kr} = Pr\{A_{tk}|A_{sr}\} = \frac{(n-r)!}{(k-r)!(k-n)!}\left(\frac{x_{tk}-x_{sr}}{1-x_{sr}}\right)^{k-r}\left(\frac{1-x_{tk}}{1-x_{sr}}\right)^{n-k} \tag{22}$$

for $t, s = 1, 2$ and with respect to (17).

In particular:

1. $t = s = 1 \Rightarrow \underline{B}_{11} = (_{11}b_{kr})$.
   Replacing $t = s = 1$ in (22) we have that $Pr\{A_{1k}|A_{1r}\}$ is:

$$_{11}b_{kr} = \frac{(n-r)!}{(k-r)!(n-k)!}\left(\frac{k-r}{n_1-r}\right)^{k-r}\left(\frac{n_1-k}{n_1-r}\right)^{n-k}$$

   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 $\underline{B}_{11}$ having the following framework:

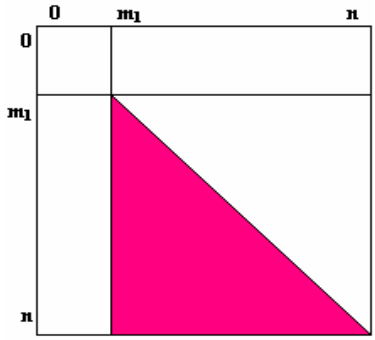

Figure 2: Framework of the matrix B_{11}

2. $t = 2, s = 1 \Rightarrow \underline{B}_{21} = ({}_{21}b_{kr})$.
Replacing $t = 2, s = 1$ in (22) we have that $Pr\{A_{2k}|A_{1r}\}$ is:

$${}_{21}b_{kr} = \frac{(n-r)!}{(k-r)!(n-k)!}\left(\frac{k-r+2nD}{n_1-r}\right)^{k-r}\left(\frac{n_2-k}{n_1-r}\right)^{n-k}$$

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 $\underline{B}_{21}$ having the following framework:

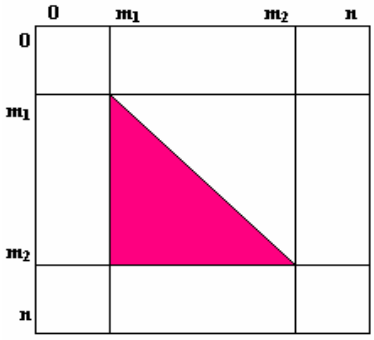

Figure 3: Framework of the matrix B_{21}

3. $t = 1, s = 2 \Rightarrow \underline{B}_{12} = ({}_{12}b_{kr})$.
   Replacing $t = 1, s = 2$ in (22) we have that $Pr\{A_{1k}|A_{2r}\}$ is:

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

   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 $l_1 = int(2nD+1)$.
   Thus we have the matrix $\underline{B}_{12}$ having the following framework:

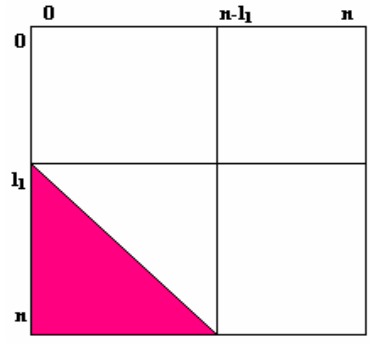

Figure 4: Framework of the matrix B_{12}

4. $t = s = 2 \Rightarrow \underline{B}_{22} = ({}_{22}b_{kr})$.
   Replacing $t = s = 2$ in (22) we have that $Pr\{A_{2k}|A_{2r}\}$ is:

$$ {}_{22}b_{kr} = \frac{(n-r)!}{(k-r)!(n-k)!} \left(\frac{k-r}{n_2-r}\right)^{k-r} \left(\frac{n_2-k}{n_2-r}\right)^{n-k} $$

   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 $\underline{B}_{22}$ having the following framework:

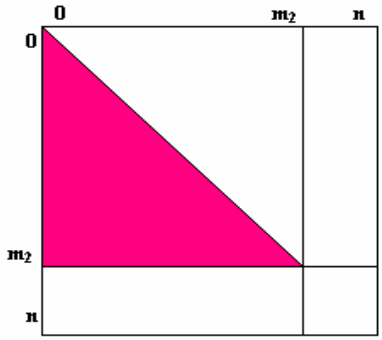

Figure 5: Framework of the matrix B_{22}

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

$$ \underline{B} = \begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix} $$

with the following framework:

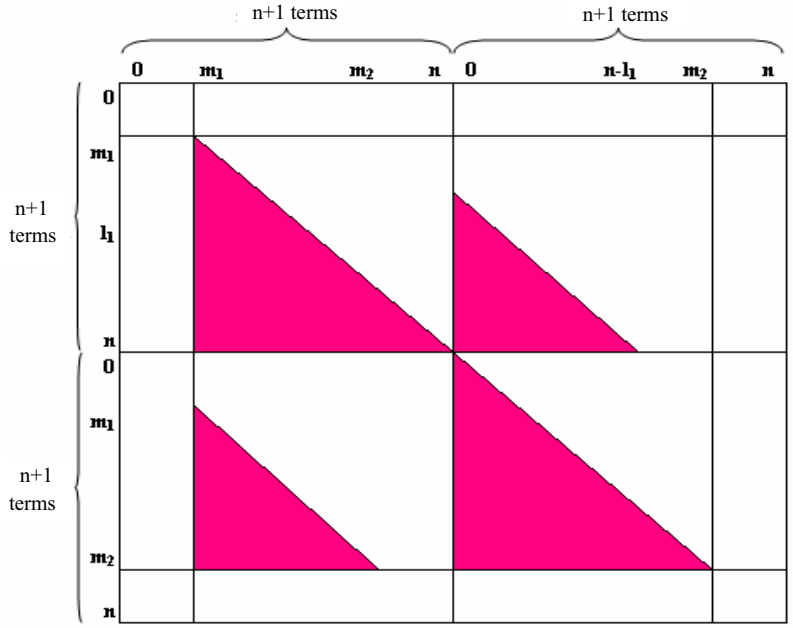

Figure 6: Framework of the matrix *B*

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:

$$\underline{W} = \{W_r\}$$

and

$$\underline{Y} = \{Y_r\}$$

which together define the vector

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

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

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

$$\begin{bmatrix} \underline{C}_1 \\ \underline{C}_2 \end{bmatrix} = \begin{bmatrix} \underline{B}_{11} & \underline{B}_{12} \\ \underline{B}_{21} & \underline{B}_{22} \end{bmatrix} \cdot \begin{bmatrix} \underline{W} \\ \underline{Y} \end{bmatrix}$$

or:

$$\underline{C} = \underline{B} \cdot \underline{Z}.$$

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

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

$$Pr\{D_n > D\} = \sum_{r=0}^{n} \left[Pr\{U_r\} + Pr\{V_r\} \right]. \tag{23}$$

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

$$F_{D_n}(D) = Pr\{D_n \leq D\}. \tag{24}$$

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

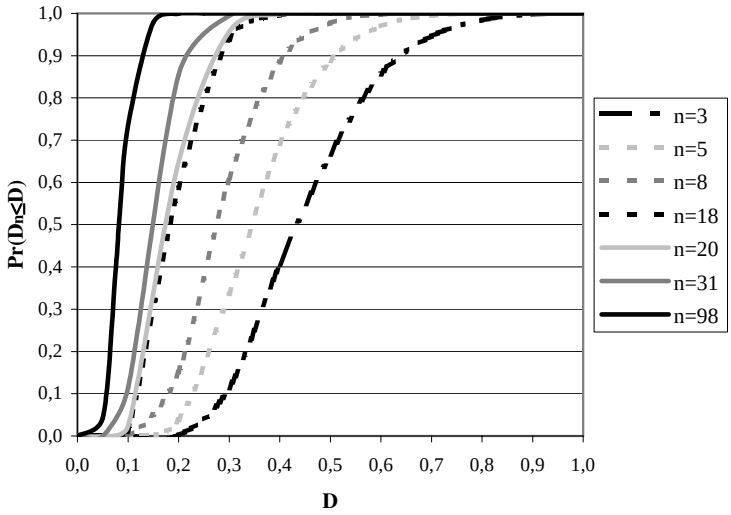

Figure 7: Cumulative distribution function of D_n statistic

For a fixed significance level α, 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

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

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.

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

Table 2: Critical values of Kolmogorov-Smirnov statistic given by Massey (1951) and integrated by Birnbaum (1952)

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

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 3: Percentage differences $\frac{D_{\alpha,n}-d_\alpha(n)}{d_\alpha(n)}$ between critical values in Table 1 and Table 2

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

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_α with $\sqrt{n}$, and are reported in Table 4.

Table 4: Asymptotic critical values d_α $(n > 35)$ of Kolmogorov-Smirnov statistic given by Smirnov (1948)

	Significance level (α)				
n	**0 .01**	**0 .05**	**0 .10**	**0 .15**	**0 .20**
> 35	$1.63/\sqrt{n}$	$1.36/\sqrt{n}$	$1.22/\sqrt{n}$	$1.14/\sqrt{n}$	$1.07/\sqrt{n}$

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

	Significance level (α)				
n	**0 .01**	**0 .05**	**0 .10**	**0 .15**	**0 .20**
50	1 .59834	1 .33014	1 .19918	1 .11391	1 .04913
80	1 .60532	1 .33806	1 .20453	1 .11902	1 .05408
100	1 .60808	1 .34028	1 .20663	1 .12105	1 .05600

In Tables 4 and 5 we observe that the differences between the values are from the second decimal place on, and for $n \to \infty$ 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))*
```

```
(((k-r+2*nD)/(n1-r))^(k-r))*(((n2-k)/(n1-r))^(n-k));
end
end
% B12 matrix
for r=0:n-l1
for k=l1+r:n
B(k+1,r+n+2)=factorial(n-r)/(factorial(k-r)*factorial(n-k))*
(((k-r-2*nD)/(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

Birnbaum, Z.W. (1952). Numerical tabulation of the distribution of Kolmogorov statistic for finite sample size. *Journal of the American Statistical Association.* (47): 425-441.

Cantelli, F.P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari.* (4): 421-424.

Conover, W.J. (1972). A Kolmogorov goodness of fit test for discontinuous distributions. *Journal of the American Statistical Association.* (67): 591-596.

Conover, W.J. (1999). *Practical Nonparametric Statistics.* John Wiley e Sons, New York.

D'Agostino, R.B. and Stephens, M.A (1986). *Goodness of Fit Techniques.* Marcell Dekker, New York.

Darling, D.A. (1957) . The Kolmogorov-Smirnov, Cramér-von Mises tests. *The Annals of Mathematical Statistics.* (28): 823-838.

Doob, J.L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics.* (20): 393-403.

Drew, J.H., Glen, A. G. and Leemis, L.M. (2000). Computing the cumulative distribution function of the Kolmogorov-Smirnov statistics. *Computational Statistics and Data Analysis.* (34): 1-15.

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function.* Society for Industrial and Applied Mathematics, Philadelphia.

Facchinetti, S. and Chiodini, P.M. (2008). Exact and approximate critical values of Kolmogorov-Smirnov test for discrete random variables. *XLIV Riunione scientifica SIS, Arcavacata di Rende (CS).* 1-2 CD.

Facchinetti, S. and Osmetti, S.A. (2009). The Kolmogorov-Smirnov goodness of fit test for discrete extreme value distributions. *Classification and Data Analysis Conference, Catania,* 485-488.

Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics.* (19): 177-189.

Frosini, B.V. (1978). A survey of class of goodness of fit statistics. *Metron.* XXXVI: 1-49.

Gentle, G.E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics.* Springer, New York.

Glivenko, V.I. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari.* (4): 92-99.

Jalla, E. (1979). Il test di Kolmogorov nel caso di distribuzione discreta. *Istituto di Statistica dell'Università degli Studi di Torino.* (4): 1-16.

Kendall, M.G. and Stuart, A. (1967). *The Advanced Theory of Statistics.* Griffin, London.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari.* (4): 83-91.

Kolmogorov, A. (1941). Confidence limits for an unknown distribution function. *The Annals of Mathematical Statistics.* (4): 461-463.

Marsaglia, G., Tsang, W.W. and Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software.* (84): 1-4.

Marvulli, R. (1980). Tabelle per l'uso del test di Kolmogorov nel caso discreto. *Istituto di Statistica dell'Università degli Studi di Torino.* (6): 1-84.

Massey, F.J. (1950). A note on the estimation of the distribution function by confidence limits. *The Annals of Mathematical Statistics.* (21): 116-119.

Massey, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association.* (46): 68-78.

Miller, L.H. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association.* (51): 111-121.

Noether, G.E. (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika.* (7): 115-116.

Pettitt, A.N. and Stephens, M.A. (1977). The Kolmogorov-Smirnov goodness of fit statistic with discrete and grouped data. *Journal of the American Statistical Association.* (19): 205-210.

Schmid, P. (1958). On the Kolmogorov and Smirnov limit theorems for discontinuous distribution functions. *The Annals of Mathematical Statistics.* (29): 1011-1027.

Smirnov, N. (1939 A). Sur les écarts de la courbe de distribution empirique. *Recueil Mathématique.* (6): 3-26.

Smirnov, N. (1939 B). On the estimation of the discrepancy between critical curves of distribution of two independent samples. *Bulletin Mathématique de l'Université de Moscou.* (2): 1-16.

Smirnov, N. (1944). Approximate laws of distribution of random variables from empirical data. *Uspehi Matem. Nauk.* (10): 179-206.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics.* (19): 279-281.

Stephens, M.A. (1970). Use of the Kolmogorov-Smirnov, Cramér-von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society B.* (32): 115-122.

Wood, C.L., Altavela, M. M. (1978). Large-sample results for Kolmogorov-Smirnov test for discrete distributions. *Biometrika*. (65): 23-239.