Statistical Papers 40, 343–349 (1999)

Statistical Papers

A modified Kolmogorov-Smirnov test for a rectangular distribution with unknown parameters: Computation of the distribution of the test statistic

Helmut Schellhaas

Received: May 23, 1997; revised version: November 20, 1997

A recursive scheme for the calculation of the distribution of the test statistic of a modified Kolmogorov-Smirnov-test for a rectangular distribution with unknown parameters is given.

Key words: rectangular distribution, modified Kolmogorov-Smirnov one sample test, distribution of the test statistic, percentage points, recursive scheme for computation

1 Introduction

Recently, Lassahn (1996) gave an algorithm for the calculation of the distribution of the test statistic D_n^* of a modified Kolmogorov-Smirnov-test for a rectangular distribution with unknown parameters. His procedure is based on the calculation of the Steck-determinant (1971). In our paper we show a quite different approach. Based on procedures developed in Friedrich and Schellhaas (1996) for the Kolmogorov-Smirnov-test with completely specified distribution of the population we present a recursive scheme for the calculation of the distribution of D_n^*. With respect to numerical stability it is advantageous that all terms in the recursion are nonnegative. We give a table of percentage points for large n supplementing Lassahn's table.

2 Computation of the distribution of the test statistic

Assume $(X_1, X_2, \ldots, X_n)$, $n \geq 3$ is a random sample from a population with a rectangular distribution on $[a, b]$. Let F be the distribution function,

$$F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \leq x < b \\ 1, & x \geq b \end{cases} \tag{2.1}$$

where the parameters $a, b \in \mathbb{R}$, $a < b$, are both unknown. Let $(Y_1, Y_2, \ldots, Y_n)$ with $a \leq Y_1 \leq Y_2 \leq \cdots \leq Y_n \leq b$ be the corresponding sample of order statistics. The maximum likelihood estimators $\hat{a}$ and $\hat{b}$ of the parameters a and b are $\hat{a} = Y_1$ and $\hat{b} = Y_n$. Replacing a and b in (2.1) by these estimators we get the random function

$$F^*(x) = \begin{cases} 0, & x < Y_1 \\ \dfrac{x-Y_1}{Y_n - Y_1}, & Y_1 \leq x < Y_n \\ 1, & x \geq Y_n \end{cases} \tag{2.2}$$

Let S_n be the empirical distribution function of the sample with

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(X_i) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty, x]}(Y_i), \quad x \in \mathbb{R}$$

where $1_A(x) = 1$ for $x \in A$ and $1_A(x) = 0$ for $x \notin A$ is the indicator function of the set A. We investigate the distribution of the random variable

$$D_n^* = \sup_{x \in \mathbb{R}} \left| S_n(x) - F^*(x) \right|, \quad n \in \mathbb{N}, \quad n \geq 3 \tag{2.3}$$

D_n^* can serve as test statistic in a (modified) two-sided Kolmogorov-Smirnov-test for testing the hypothesis that a population has a rectangular distribution (with unknown parameters). For the application of the test the distribution (resp. the percentage points of the distribution) of D_n^* is needed. For $d \in \mathbb{R}$ we write

$$p_n(d) = P(D_n^* \leq d) \tag{2.4}$$

In theorem 2.1 we give a representation of $p_n(d)$. Note that $p_n(d)$ is independent of the parameters a and b as is seen by theorem 2.1. In the proof of theorem 2.1 we use the following lemma 2.1. The assertion of the lemma may be found in D'Agostino/Stephens (1986), chapter 8.16(a). A separate proof is given in Schellhaas (1996).

Lemma 2.1 *Let $(Y_1, Y_2, \ldots, Y_n)$ with $a \leq Y_1 \leq Y_2 \leq \cdots \leq Y_n \leq b$ be a sample of order statistics of size $n \geq 3$ from a rectangular distribution on $[a, b]$. Put*

$$Z_i = \frac{Y_i - Y_1}{Y_n - Y_1}, \qquad i = 2, 3, \ldots, n-1 \tag{2.5}$$

Then $(Z_2, Z_3, \ldots, Z_{n-1})$ *is distributed as a sample of order statistics of size* $n-2$ *from a rectangular distribution on* $[0,1]$.

Let S^*_{n-2} be the empirical distribution function of the sample $(Z_2, Z_3, \ldots, Z_{n-1})$ defined in lemma 2.1. For $\frac{1}{n} \le d < 1 - \frac{1}{n}$ let $m = \mathrm{INT}(nd)$ be the integer part of nd, put

$$I = \{m+1, \ldots, n-1\}, \quad \text{and} \quad J = \{1, \ldots, n-m-1\} \tag{2.6}$$

$$u_i = \tfrac{i}{n} - d, \quad i \in I \quad \text{and} \quad v_j = \tfrac{j}{n} + d, \quad j \in J \tag{2.7}$$

Note that $I \neq \emptyset$, $J \neq \emptyset$. We then have

Theorem 2.1 *The distribution function of* D^*_n, $n \ge 3$, *has the representation*

$$p_n(d) = \begin{cases} 0, & d < \frac{1}{n} \\ P\Big(S^*_{n-2}(u_i) < \frac{i-1}{n-2},\ i \in I;\ \frac{j-1}{n-2} < S^*_{n-2}(v_j),\ j \in J\Big), & \frac{1}{n} \le d < 1 - \frac{1}{n} \\ 1, & d \ge 1 - \frac{1}{n} \end{cases}$$

Proof: With the convention $S_n(Y_0) = 0$ we have by (2.3)

$$\begin{aligned} p_n(d) &= P\Big(|S_n(Y_{i-1}) - F^*(Y_i)| \le d,\ |S_n(Y_i) - F^*(Y_i)| \le d,\ i = 1, 2, \ldots, n\Big) \\ &= P\Big(S_n(Y_i) - d \le F^*(Y_i) \le S_n(Y_{i-1}) + d,\ i = 1, 2, \ldots, n\Big) \end{aligned} \tag{2.8}$$

For $d < \frac{1}{n}$ we have $S_n(Y_1) - d = \frac{1}{n} - d > 0$ and $F^*(Y_1) = 0$ with probability one. Therefore $S_n(Y_1) - d > F^*(Y_1)$ and (2.8) (considering $i = 1$) gives $p_n(d) = 0$ for $d < \frac{1}{n}$.

For $d \ge \frac{1}{n}$ the conditions for $i = 1$ and $i = n$ in (2.8) are trivial, since $F^*(Y_1) = 0$, $F^*(Y_n) = 1$ and $S_n(Y_{n-1}) + d = 1 - \frac{1}{n} + d \ge 1$ with probability one. Therefore

$$p_n(d) = P\Big(S_n(Y_i) - d \le F^*(Y_i) \le S_n(Y_{i-1}) + d,\ i = 2, \ldots, n-1\Big) \tag{2.9}$$

For $d \ge 1 - \frac{1}{n}$ we have $S_n(Y_{n-1}) - d \le 0$ and $S_n(Y_1) + d \ge 1$ with probability one. Therefore $p_n(d) = 1$ for $d \ge 1 - \frac{1}{n}$ by (2.9).

For $\frac{1}{n} \le d < 1 - \frac{1}{n}$ we have by the definition of F^* in (2.2) and of Z_i in (2.5) using (2.9) and the continuity of the Z_i

$$\begin{aligned} p_n(d) &= P\Big(\tfrac{i}{n} - d < Z_i \le \tfrac{i-1}{n} + d,\ i = 2, 3, \ldots, n-1\Big) \\ &= P\Big(\tfrac{i}{n} - d < Z_i,\ i = 2, 3, \ldots, n-1;\ Z_{j+1} \le \tfrac{j}{n} + d,\ j = 1, 2, \ldots, n-2\Big) \\ &= P\Big(u_i < Z_i,\ i \in I;\ Z_{j+1} \le v_j,\ j \in J\Big) \end{aligned} \tag{2.10}$$

Considering (2.10) we can apply the following arguments using the monotonicity of S^*_{n-2} and the fact that S^*_{n-2} has a jump at $Z_i (Z_j, Z_{j+1})$. For $i \in I$

we have $u_i < Z_i$ if and only if $S^*_{n-2}(u_i) < S^*_{n-2}(Z_i) = \frac{i-1}{n-2}$. For $j \in J$ and $j \geq 2$ we have $Z_{j+1} \leq v_j$ if and only if $S^*_{n-2}(v_j) > S^*_{n-2}(Z_j) = \frac{j-1}{n-2}$. For $j = 1$ we have $Z_{j+1} \leq v_j$ if and only if $S^*_{n-2}(v_j) > 0 = \frac{j-1}{n-2}$, each with probability one.

Therefore by (2.10)

$$p_n(d) = P\Big(S^*_{n-2}(u_i) < \frac{i-1}{n-2},\ i \in I;\ \frac{j-1}{n-2} < S^*_{n-2}(v_j),\ j \in J\Big)$$

for $\frac{1}{n} \leq d < 1 - \frac{1}{n}$. This proves the theorem. □

To compute $p_n(d)$ for $\frac{1}{n} \leq d < 1 - \frac{1}{n}$ let

$$\mathbb{U} = \{u_i,\ i \in I\}, \quad \mathbb{V} = \{v_j,\ j \in J\}, \quad \mathbb{W} = \mathbb{U} \cup \mathbb{V}$$

Let

$$\mathbb{W} = \{w_1, w_2, \ldots, w_\sigma\}$$

with $w_1 < w_2 < \cdots < w_\sigma$.

Put

$$\Sigma = \{1, 2, \ldots, \sigma\}$$

and $h(u_i) = \dfrac{i-1}{n-2}$ for $i \in I$ and $g(v_j) = \dfrac{j-1}{n-2}$ for $j \in J$.

For $k \in \Sigma$ define the intervals

$$gh(k) = \begin{cases} \{x : x \in \mathbb{R},\ 0 \leq x < h(w_k)\} & \text{if} \quad w_k \in \mathbb{U} \cap \overline{\mathbb{V}} \\ \{x : x \in \mathbb{R},\ g(w_k) < x \leq 1\} & \text{if} \quad w_k \in \overline{\mathbb{U}} \cap \mathbb{V} \\ \{x : x \in \mathbb{R},\ g(w_k) < x < h(w_k)\} & \text{if} \quad w_k \in \mathbb{U} \cap \mathbb{V} \end{cases} \tag{2.11}$$

where $\overline{\mathbb{U}}$ ($\overline{\mathbb{V}}$) is the complement of $\mathbb{U}$ ($\mathbb{V}$) with respect to $\mathbb{W}$.

Then

$$p_n(d) = P(S^*_{n-2}(w_k) \in gh(k),\ k \in \Sigma) \tag{2.12}$$

Realize that the representation of $p_n(d)$ in (2.12) is of the type appearing in theorem 2.1 of the paper Friedrich and Schellhaas (1996). In that paper two recursive schemes for the calculation of such probabilities are given. Both schemes are easily transferred for the calculation of $p_n(d)$ in (2.12). The proof in that paper only needs slight modifications. We therefore omit the proof and merely present the corresponding algorithm of the second scheme .

For $k \in \Sigma$ put

$$\mathcal{L}_k = \{\ell :\ \ell \in \{0, 1, \ldots, n-2\},\ \frac{\ell}{n-2} \in gh(k)\}$$

Then $p_n(d)$ may be calculated recursively by the

algorithm:

(1) For $k = 1;\ i \in \mathcal{L}_1$

$$T_{i,1} = w_1^i$$

(2) For $k = 2, 3, \ldots, \sigma;\ i \in \mathcal{L}_k$

$$T_{i,k} = \sum_{\substack{j \in \mathcal{L}_{k-1} \\ j \le i}} \binom{i}{j} (w_k - w_{k-1})^{i-j} T_{j,k-1}$$

(3)

$$p_n(d) = \sum_{i \in \mathcal{L}_\sigma} \binom{n-2}{i} (1 - w_\sigma)^{n-2-i} T_{i,\sigma}$$

With respect to numerical stability of the algorithm it is advantageous that all terms in the recursion are nonnegative. Numerical stability may be expected even for large n.

To apply the modified Kolmogorov-Smirnov-test the percentage points $d_{n,p}$ of the distribution of D_n^+ are needed, i.e. the solution $d_{n,p}$ of

$$p_n(d) = p, \quad p \in (0, 1)$$

Krumbholz, Schader, Schmid (1996) gave percentage points obtained by Monte Carlo simulation. Lassahn (1996) tabulated percentage points for $n = 1(1)100$ and $p = 0.80,\ 0.85,\ 0.90,\ 0.95,\ 0.975,\ 0.99,\ 0.995$ to four decimal places using the Steck-determinant. Our calculation with the above algorithm confirmed Lassahn 's results for $n \le 87$. For $88 \le n \le 100$ and $p = 0.99,\ p = 0.995$ occasionally differences of at most two units in the fourth decimal appeared. In table 2 we present our results for $n = 88(1)100$, in table 1 we supplement Lassahn's table presenting percentage points for $n = 105(5)150$.

Additional note: By a recalculation Lassahn got our results for $n = 88, 89$. He stated that the condition of the matrix of the Steck-determinant gets worse with increasing n presumably resulting in difficulties for $n \ge 90$ (private communication by R. Lassahn).

	p						
n	.80	.85	.90	.95	.975	.99	.995
105	.1031	.1094	.1177	.1306	.1424	.1567	.1666
110	.1007	.1069	.1150	.1277	.1392	.1531	.1629
115	.0985	.1046	.1125	.1249	.1362	.1498	.1593
120	.0965	.1024	.1102	.1223	.1334	.1467	.1560
125	.0946	.1004	.1080	.1199	.1307	.1438	.1529
130	.0928	.0984	.1059	.1176	.1282	.1410	.1500
135	.0911	.0966	.1040	.1154	.1259	.1384	.1472
140	.0894	.0949	.1021	.1134	.1236	.1360	.1446
145	.0879	.0933	.1004	.1114	.1215	.1337	.1422
150	.0864	.0917	.0987	.1096	.1195	.1314	.1398

Table 1: percentage points $d_{n,p}$ for $n = 105(5)150$

	n						
p	88	89	90	91	92	93	94
0.99	.1708	.1699	.1690	.1681	.1672	.1663	.1654
0.995	.1817	.1807	.1797	.1787	.1778	.1768	.1759

p	95	96	97	98	99	100
0.99	.1646	.1637	.1629	1621	.1613	.1605
0.995	.1750	.1741	.1732	.1724	.1715	.1707

Table 2: percentage points $d_{n,p}$ for $n = 88(1)100$

Note that for $n = 150$ the percentage points tabulated differ from the corresponding percentage points in case of a completely specified distribution, as considered and tabulated in Friedrich and Schellhaas (1996), in at most two units in the fourth decimal. This demonstrates the fact that $\sqrt{n}D_n^*$ has the classical Kolmogorov distribution as limit distribution, see Krumbholz, Schader, Schmid (1996).

Remark: The procedure is easily modified if only one of the parameters a, b is unknown. If for example b is unknown then

$$F^*(x) = \begin{cases} 0, & x \leq a \\ \dfrac{x-a}{Y_n - a}, & a < x < Y_n \\ 1, & x \geq Y_n \end{cases}$$

is used in (2.3) instead of (2.2) and a lemma analogous lemma 2.1 is used stating that $(z_1, z_2, \ldots, z_{n-1})$ with

$$Z_i = \frac{Y_i - a}{Y_n - a}, \quad i = 1, \ldots, n-1$$

is distributed as a sample of order statistics of size $n-1$ from a rectangular distribution on $[0,1]$. The derivation of the representation of the distribution of D_n^* corresponding to theorem 2.1 can be done analogously to the proof of theorem 2.1.

References

D'Agostino, R.B.; Stephens, M.A. (1986). Goodness-of-fit techniques. Marcel Dekker, Inc. New York

Friedrich, Th.; Schellhaas, H. (1996). Computation of the percentage points and the power for the two sided Kolmogorov-Smirnov one sample test.
Preprint-Nr. 1835, Fachbereich Mathematik, Technische Hochschule Darmstadt.
To appear in Statistical Papers

Krumbholz, W.; Schader, M.; Schmid, F. (1996). Modifications of the Kolmogorov, Cramer- von Mises, and Watson tests for testing uniformity with unknown limits. Commun.Statist.-Simula. 25, 1093-1104

Lassahn, R. (1996). Die exakte Berechnung der Quantile des Kolmogoroffschen Anpassungstestes auf Gleichverteilung mit Hilfe der Steck-Determinante.
Diskussionsbeiträge zur Statistik und Qualitativen Ökonomik Nr. 70,
Universität der Bundeswehr Hamburg

Schellhaas, H. (1996). The Kolmogorov-Smirnov-test for a rectangular distribution with unknown parameters: computation of the distribution of the test statistic. Preprint Nr. 1863. Fachbereich Mathematik, Technische Hochschule Darmstadt

Steck, G.P. (1971). Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions. Ann. Math. Statist. 42, 1-11

Helmut Schellhaas
Fachbereich Mathematik
Technische Universität Darmstadt
Schloßgartenstraße 7
D-64289 Darmstadt, Germany