The Annals of Mathematical Statistics
1969, Vol. 40, No. 5, 1833–1837

NOTES

RESULTS FROM THE RELATION BETWEEN TWO STATISTICS OF THE KOLMOGOROV-SMIRNOV TYPE

BY M. A. STEPHENS

McGill University and University of Nottingham

1. Introduction. In this paper we demonstrate the relation existing between the distributions of two statistics of the Kolmogorov-Smirnov type. They have been named K_n (Brunk, 1962) and V_n (Kuiper, 1960; Stephens, 1965). From the relation, given as a theorem in Section 3, new results are found for each statistic by using what is known about the other. Tables of percentage points are given for K_n, and it is shown how to adapt an existing table for K_n to give a table of probabilities for V_n.

2. The statistic V_n. This is a statistic of the Kolmogorov-Smirnov type, suitable for tests of goodness-of-fit. Suppose a random sample of size n is given, and let the values, in ascending order, be $x_1, x_2, \cdots, x_n$; let the sample or empirical distribution function be $F_n(x)$. It is required to test the null hypothesis H_0, that the sample comes from a continuous distribution $F(x)$; well-known test statistics are

$$D_n^+ = \sup_{-\infty<x<\infty} \{F_n(x) - F(x)\}$$

$$D_n^- = \sup_{-\infty<x<\infty} \{F(x) - F_n(x)\}$$

$$D_n = \max \{D_n^+, D_n^-\}.$$

V_n is given by $D_n^+ + D_n^-$. It was suggested by Kuiper (1960) for use with observations on a circle; the value of V_n does not depend on the choice of origin for x. This is a necessary property of a goodness-of-fit statistic for the circle, since otherwise the same data could, by a change of origin, yield different values of the test statistic. V_n may, of course, be used also for observations on a line. The asymptotic distribution of V_n was given by Kuiper, and the small-sample distribution, in the tails, by Stephens (1965); in the latter paper there are tables of upper and lower percentage points of V_n.

3. The statistic K_n. Suppose, following Brunk's (1962) notation, that $U_i = F(x_i)$, $i = 1, 2, \cdots, n$; put $U_0 = 0$, $U_{n+1} = 1$. Define statistics (the C-*class*):

$$C_n^+ = \max_{0 \leq i \leq n+1} (i/(n+1) - U_i),$$

$$C_n^- = \max_{0 \leq i \leq n+1} (U_i - i/(n+1)),$$

$$C_n = \max \{C_n^+, C_n^-\} \quad \text{and} \quad K_n = C_n^+ + C_n^-.$$

Received 24 April 1968.

On H_0, the U_i, $i = 1, \cdots, n$, should be uniformly distributed between 0 and 1; the C class is based on distances between the uniform order statistics and their expected values. They have some properties which make them more attractive, and in some ways might be regarded as more natural than the D statistics (Pyke (1959), Durbin (1967)). The statistic K_n is the subject of Brunk's paper, and he illustrates its use as a goodness-of-fit statistic.

It is convenient to define the D statistics also in terms of the U_i; then

$$D_n^+ = \max_{0 \leq i \leq n} (i/n - U_i);$$

$$D_n^- = \max_{0 \leq i \leq n} (U_i - (i-1)/n).$$

4. The connection between the null distributions of K_n and V_n. In this section we prove

THEOREM 1. $\Pr (K_n < z) = \Pr (V_{n+1} < z + 1/(n+1))$.

PROOF. Let $S = \{U_i; i = 1, 2, \cdots, n\}$ be the order statistics of a sample of n independent uniformly distributed random variables on [0, 1]. Imagine an extra observation U_0 added at the origin, and let T be the new set $U_0, U_1, \cdots, U_n$. Define $U_{n+1} = 1$ as before. The $n + 1$ spacings $U_i - U_{i-1}$, $i = 1, \cdots, n + 1$, will have the same joint distribution as the spacings between $n + 1$ observations independently sampled from a uniform distribution on a circle of unit circumference. Thus T may be regarded as such a set of observations on a circle.

Let $D^+(S)$ be the value of D_n^+ for the set S, the sample size subscript being dropped; similarly $D^+(T)$, $C^+(S)$, etc. Calculate the V statistic for the set T, with origin at U_0; as noted above, this does not affect the value of V. Then

$$V(T) = \max_{0 \leq i < n+1} \{(i+1)/(n+1) - U_i\} + \max_{0 \leq i < n+1} \{U_i - i/(n+1)\}$$

$$= C^+(S) + 1/(n+1) + C^-(S) = K(S) + 1/(n+1).$$

Thus a value of K_n produces, by a one-to-one correspondence, a value of V_{n+1} given by $V_{n+1} = K_n + 1/(n+1)$, and the theorem follows.

Brunk continues by deriving $\Pr (K_n < t/(n+1))$, for t an integer, and gives a table of probabilities for given n and t. Stephens continues in a different way to obtain the upper and lower tails of the distribution of V_n, i.e. to give explicit formulae for $\Pr (V_n < z)$. These are used to give percentage points of V_n. The theorem above now makes it possible to extend the results for V_n by using those for K_n, and vice versa.

5. New results for V_n. In particular, we may adapt Brunk's Table 2.1 as follows: add one to the values of n in the horizontal heading to the table, and add one to the values of t in the left-hand column. The table entries, with the new labelling, now give $\Pr (V_n < t/n)$, for $n = 2, 3, \cdots, 21$, and for $t = 2, 3, \cdots, T$, where $T = 8$ or 12. For example, the value .6713 appears, in Brunk's table, at the intersection of $t = 2$ and $n = 5$; after relabelling it occurs at the intersection $t = 3$, $n = 6$, so that $\Pr (V_6 < 3/6) = .6713$. The new table will not include $t = 1$, and this value is not necessary, since $\Pr (V_n < 1/n) = 0$. (Stephens, 1965).

TABLE 1

Upper tail percentage points for K_n

r	Significance levels as percentages						
	15.0	10.0	5.0	2.5	1.0	0.5	0.1
4	0.419	0.452	0.500	0.540	0.589	0.622	0.681
5	0.404	0.434	0.479	0.520	0.565	0.595	0.657
6	0.389	0.418	0.461	0.498	0.543	0.573	0.632
7	0.376	0.403	0.444	0.480	0.522	0.551	0.609
8	0.364	0.389	0.428	0.463	0.503	0.531	0.588
9	0.352	0.377	0.414	0.447	0.486	0.513	0.568
10	0.341	0.365	0.401	0.433	0.471	0.496	0.550
12	0.323	0.345	0.377	0.408	0.444	0.468	0.519
14	0.307	0.329	0.359	0.388	0.421	0.444	—
16	0.294	0.314	0.344	0.370	0.401	0.424	—
18	0.282	0.302	0.329	0.354	0.384	0.405	—
20	0.271	0.291	0.316	0.340	0.370	0.389	—
25	0.249	0.267	0.290	0.313	0.339	0.356	—
30	0.233	0.248	0.270	0.291	0.314	0.331	—
35	0.219	0.233	0.254	0.273	0.295	0.311	—
40	0.208	0.221	0.241	0.258	0.279	0.294	—
45	0.198	0.210	0.229	0.245	0.266	0.279	—
50	0.189	0.201	0.219	0.234	0.254	0.267	—
60	0.175	0.186	0.202	0.217	0.234	0.245	—
80	0.155	0.165	0.178	0.190	0.205	0.216	—
100	0.140	0.148	0.161	0.172	0.186	0.195	—
∞	0.000	0.000	0.000	0.000	0.000	0.000	0.000

6. New results for K_n. (a) Tables 1 and 2 give upper and lower percentage points for K_n; they have been derived by interpolation in Stephens' (1965) tables for V_n and may therefore contain an inaccuracy in the third decimal place. Exact formulae for the distribution in the tails of K_n may be deduced from the V_n formulae, given in Stephens (1965).

(b) *The asymptotic distribution of K_n.* Brunk suggests that, as n becomes large, $\Pr\{K_n < t/(n+1)\}$ might be approximated by $\Pr\{(n+1)^{\frac{1}{2}}V_n < (t+1)/(n+1)^{\frac{1}{2}}\}$ and gives an argument to support this. The second probability would then be in turn approximated by an expansion due to Kuiper (1960). However, when he writes down the expression, Brunk actually gives Kuiper's expansion for $\Pr\{(n+1)^{\frac{1}{2}}V_{n+1} < (t+1)/(n+1)^{\frac{1}{2}}\}$ as the approximation for $\Pr\{K_n < t/(n+1)\}$. The theorem above now shows that in fact these two probabilities are *identically* equal for *all* n; thus Brunk's series approximation is justified. He demonstrates its accuracy with a table of exact and approximate probabilities, for $n = 19$.

(c) It is clear that the limiting expressions, as $n \to \infty$, for the characteristic functions of nK_n^2 and nV_n^2 will be the same, i.e. $\phi(t) = \prod_{j=1}^{\infty}(1 - it/2j^2)^{-2}$. In Stephens (1965, Section 5), where $\phi(t)$ is given, the limit operation has been

TABLE 2

Lower tail percentage points for K_n

r	Significance levels as percentages						
	15.0	10.0	5.0	2.5	1.0	0.5	0.1
4	0.188	0.170	0.143	0.120	0.096	0.080	0.054
5	0.189	0.170	0.147	0.128	0.106	0.093	0.067
6	0.190	0.172	0.147	0.130	0.112	0.100	0.076
7	0.188	0.171	0.149	0.131	0.114	0.103	0.082
8	0.185	0.170	0.148	0.132	0.114	0.104	0.085
9	0.182	0.167	0.147	0.131	0.114	0.104	0.087
10	0.179	0.165	0.146	0.130	0.114	0.104	0.087
12	0.172	0.159	0.142	0.128	0.113	0.104	0.087
14	0.167	0.155	0.137	0.125	0.112	0.103	0.087
16	0.162	0.150	0.134	0.122	0.109	0.102	0.086
18	0.157	0.145	0.131	0.119	0.107	0.099	0.086
20	0.152	0.141	0.127	0.117	0.105	0.097	—
25	0.142	0.133	0.119	0.110	0.099	0.092	—
30	0.133	0.125	0.113	0.104	0.094	0.088	—
35	0.125	0.119	0.108	0.099	0.090	0.085	—
40	0.123	0.113	0.103	0.095	0.086	0.081	—
45	0.123	0.108	0.099	0.091	0.083	0.077	—
50	0.119	0.104	0.095	0.087	0.081	0.075	—
60	0.103	0.098	0.089	0.082	0.076	0.071	—
80	0.092	0.087	0.079	0.073	0.068	0.064	—
100	0.085	0.080	0.073	0.068	0.063	0.059	—
∞	0.000	0.000	0.000	0.000	0.000	0.000	0.000

wrongly attached to the statistic $n^{\frac{1}{2}}V_n$ instead of to the characteristic function (and similarly with other statistics mentioned in that section).

(d) The introduction of the artificial U_0, U_{n+1} values at 0, 1 ensures that neither C_n^+ nor C_n^- will ever become negative; for observations on a circle, as the origin is changed, the position of these values obviously alters relative to the genuine observations, and the value of K_n will depend on the choice of origin for such observations. Thus K_n may not be used for goodness-of-fit tests on the circle. Note that the original definitions of C_n^+, C_n^- given by Brunk (1962, page 526) are in error, since the maximization does not include U_{n+1} ; this is later corrected.

Acknowledgments. The author is grateful to the referee for pointing out the error just noted above: also for suggesting the present proof of Theorem 1, which replaces a longer version. This work was partly supported by the National Research Council of Canada.

REFERENCES

[1] Brunk, H. D. (1962). On the range of the difference between hypothetical distribution function and Pyke's modified empirical distribution function. *Ann. Math. Statist.* **33** 525–532.

[2] DURBIN, J. (1967). Tests of serial independence based on the cumulated periodogram. To be published in *Bull. Inst. Internat. Statist.* **42**.

[3] KUIPER, N. H. (1960). Tests concerning random points on a circle. *Proc. Nederl. Akad. Wetensch. Ser. A* **63** 38–47.

[4] PYKE, RONALD. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Statist.* **30** 568–576.

[5] STEPHENS, M. A. (1965). The goodness-of-fit statistic V_N : distribution and significance points. *Biometrika* **52** 309–321.