ON THE DISTRIBUTION OF THE KOLMOGOROV-SMIRNOV D-STATISTIC

By Pedro Egydio de Oliveira Carvalho[1]

University of São Paulo

Summary. Gnedenko and Korolyuk [1] have pointed out that the exact distribution of the Kolmogorov-Smirnov D-statistic can be obtained explicitly by solving a certain double-boundary random walk problem, which, in turn, is solved by the principle of reflection. This principle is employed here in what is believed to be a new way to derive Gnedenko's and Korolyuk's result.

A random walk problem. Let us consider a random walk on the half plane $(t > 0, s)$, starting from the origin, such that at every point (t, s) there are two possible steps to take, either to $(t + 1, s + 1)$ or to $(t + 1, s - 1)$, each with equal probability and for some positive integer n, consider the paths from the origin to the point $(2n, 0)$. Among these, let us denote for any non-negative integer $k \leqq n$, the set of all paths that have a point on the line $s = k$ by $C(d_y \geqq k)$, the set of all those paths that reach the line $s = -k$ by $C(d_x \geqq k)$, the set of all those that have a point on at least one of these two lines by $C(d \geqq k)$, and the set of all those that reach both $s = \alpha k$ and $s = -k$, but go to the $s = \alpha k$ line first, by $C(d_y \geqq \alpha k \rightarrow d_x \geqq k)$. Let the number of elements in $C(\ \ )$ be $C^*(\ \ )$.

While it is well known (p. 70, 2) that

$$C^*(d_x \geqq k) = \binom{2n}{n+k} \tag{1}$$

$C^*(d \geqq k)$ is more difficult to calculate.

Clearly, $C^*(d \geqq k) = C^*(d_x \geqq k) + C^*(d_y \geqq k)$ less the number of paths in $C(d_x \geqq k) \cap C(d_y \geqq k)$ or

$$\begin{aligned} C^*(d \geqq k) = C^*(d_x \geqq k) + C^*(d_y \geqq k) - C^*(d_x \geqq k \rightarrow d_y \geqq k) \\ - C^*(d_y \geqq k \rightarrow d_x \geqq k) \end{aligned} \tag{2}$$

and by symmetry

$$C^*(d \geqq k) = 2C^*(d_x \geqq k) - 2C^*(d_x \geqq k \rightarrow d_y \geqq k). \tag{3}$$

Because of (1) it remains to calculate the last term in (3). As the first step, we show that for $i = 2, 3, \cdots, [n/k]$,

$$C^*(d_x \geqq ik) - C^*(d_x \geqq ik \rightarrow d_y \geqq k) = C^*(d_x \geqq (i-1)k \rightarrow d_y \geqq k). \tag{4}$$

Received January 20, 1958; revised June 12, 1958.

[1] Posthumous note. Revisions were made and references to literature supplemented following the referee's suggestions by Agnes Berger and Ruth Gold, School of Public Health and Administrative Medicine, Columbia University.

A path counted on the left side of (4) is one of two types. The first type reaches $s = -ik$, but does not reach $s = k$, the second reaches both lines but reaches $s = k$ first.

Let P be a path of the first type. By definition, it has points on $s = -ik$. Let p be the first of these. There must be points of P on $s = -(i-1)k$ to the left and also to the right of p; let the closest one to the left be p_{1l} and to the right, p_{1r}. Replace the portion of p from p_{1l} to p_{1r} by its reflection about $s = -(i-1)k$. The new path P' contains the image of p, say p_2', falling on the line $s = -(i-2)k$. On $s = -(i-2)k$, let the points of P' nearest to p_{1r} be p_{2l}' on the left and p_{2r} on the right. Reflect P' between p_{2l}' and p_{2r} about $s = -(i-2)k$, to get a new path P''. On $s = -(i-3)k$, let the points of P'' nearest to p_{2r} be p_{3l}'' on the left and p_{3r} on the right. Continuing in this manner, i reflections will lead to a path $P^{(i)}$ that goes first to $s = -(i-1)k$ and then to $s = k$, but does not reach $s = -ik$ except possibly after reaching $s = k$. Thus

$$P^{(i)} \varepsilon\, C(d_x \geqq (i-1)k \rightarrow d_y \geqq k)$$

but

$$P^{(i)} \not\varepsilon\, C(d_x \geqq ik \rightarrow d_y \geqq k).$$

Conversely, let Q be any path such that

$$Q \,\varepsilon\, C(d_x \geqq (i-1)k \rightarrow d_y \geqq k)$$

but

$$Q \not\varepsilon\, C(d_x \geqq ik \rightarrow d_y \geqq k).$$

Q has points on $s = k$, let q be the first of these. On $s = 0$, let q_l and q_r be the nearest points of Q to the left and right of q, respectively. Let all the other points of Q on $s = 0$ to the right of q_r be $a_1, \cdots, a_m$, in order. Let us reflect the portions of Q between q_l and q_r and at the same time between all those points a_i, a_{i+1} between which Q reaches $s = k$ about the line $s = 0$. The new path Q' does not reach k. The reflection of q, say q', lies on $s = -k$. Next reflect Q' between q' and the nearest point to the left of it on $s = -k$. Continuing in the same manner, the ith reflection will produce a path that reaches $s = -ik$ but never reaches $s = k$.

Let U be a path of the second type, i.e., one that reaches $s = k$ first and then reaches $s = -ik$. Let $p = (t, 0)$ be the first return of U to $s = 0$ after having reached $s = -ik$. Let the portion of U between $(0, 0)$ and $(t, 0)$ be represented by the ordered sequence $\epsilon_1, \epsilon_2, \cdots, \epsilon_t$, where ϵ_j is a vector of length $\sqrt{2}$ and slope $+1$ or -1. Let U' be a path such that from $(0, 0)$ to $(t, 0)$ it is given by the reversed sequence $\epsilon_t, \epsilon_{t-1}, \cdots, \epsilon_1$ and coincides with U from $(t, 0)$ to $(2n, 0)$. U' is clearly in $C(d_x \geqq ik \rightarrow d_y \geqq k)$ and therefore also in $C(d_x \geqq (i-1)k \rightarrow d_y \geqq k)$.

Conversely, let V be a path such that

$$V \,\varepsilon\, C(d_x \geqq (i-1)k \rightarrow d_y \geqq k)$$

and

$$V \varepsilon C(d_x \geqq ik \to d_y \geqq k)$$

and let $q_0' = (t', 0)$ be the first return of V to $s = 0$ after having reached $s = k$. Reversing the steps between $(0, 0)$ and q_0' uniquely determines a path of type II, completing the proof of (4).

Note that (4) has the structure

$$A_i - B_i = B_{i-1}$$

where $A_i = C^*(d_x \geqq ik)$ is known by (1). Thus knowing B_i for any i implies knowing all B_j for $j < i$. But for $i = [n/k]$, (4) gives

$$\binom{2n}{n + \left[\frac{n}{k}\right]} - 0 = C^*\left(d_x \geqq \left(\left[\frac{n}{k}\right] - 1\right)k \to d_y \geqq k\right) = 1 = B_{[n/k]-1}.$$

Carrying out the substitutions gives

$$C^*(d_x \geqq k \to d_y \geqq k) = \sum_{i=2}^{[n/k]} \binom{2n}{n + ik} (-1)^i$$

and from (3)

$$C^*(d \geqq k) = 2 \sum_{i=1}^{[n/k]} \binom{2n}{n + ik} (-1)^{i+1}$$

Application to the Kolmogorov-Smirnov problem. Let $X = (x_1 < x_2 < \cdots < x_n)$ and $Y = (y_1 < y_2 < \cdots < y_n)$ be two independent samples of ordered independent observations having the same continuous cumulative distribution function. Suppose $x_i \neq y_j$, $(i, j = 1, 2, \cdots, n)$ and let the two samples be combined and arranged in increasing order of magnitude, say $Z = (z_1 < z_2 < \cdots < z_{2n})$. Let $S_n(x)$ be the number of observed values x_i which are less than or equal to x and $S_n'(x)$ the number of observed y_j's less than or equal to x.

Let

$$D^+ = \max_x (S_n(x) - S_n'(x))$$

and

$$D = \max_x \left| S_n(x) - S_n'(x) \right|.$$

The limiting distribution of D was found by Kolmogorov [3, 4] and Smirnov [5, 6, see also 7] and an iterative method for its exact distribution has been given by Massey [8]. Gnedenko and Korolyuk recognized that a one to one correspondence exists between the set of all Z and all paths from $(0, 0)$ to $(2n, 0)$ in the above discussed random walk: Starting from $(0, 0)$, we move to $(1, 1)$ if z_1 is from Y, to $(1, -1)$, if z_1 is from X and so on. In particular, samples Z for which $D \geqq k$ correspond to paths in $C(d \geqq k)$ and vice versa. Thus we get Gnedenko's and Korolyuk's result

$$P\{D < k\} = 1 - \frac{C^*(d \geqq k)}{\binom{2n}{n}}.$$

REFERENCES

[1] B. V. Gnedenko and V. S. Korolyuk, "On the maximum discrepancy between two empirical distributions". *Doklady Akad. Nauk. SSSR (N.S.)*, Vol. 80 (1951), pp. 525–528. Reviewed by W. Feller in *Mathematical Reviews*, Vol. 13 (1952), pp. 570–571.

[2] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, John Wiley and Sons, Inc., New York, 1957.

[3] A. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione", *Inst. Ital. Attuari, Giorn.*, Vol. 4 (1933), pp. 1–11.

[4] A. Kolmogorov, "Über die Grenzwertsätze der Wahrscheinlichkeitsrechnung", *Bulletin [Izvestija] Academie des Sciences URSS*, (1933), pp. 363–372.

[5] N. Smirnov, "Ob uklonenijah empiričeskoi krivoi raspredelenija", *Recueil Mathematique (Matematičeskii Sbornik)*, N.S. Vol. 6 (48) (1939), pp. 3–26.

[6] N. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples", *Bulletin Mathématique de l'Université de Moscou*, Vol. 2 (1939), fasc. 2.

[7] W. Feller, "On the Kolmogorov-Smirnov limit theorems for empirical distributions", *Ann. Math. Stat.*, Vol. XIX, No. 2 (1948), pp. 177–189.

[8] F. Massey, Jr., "The distribution of the maximum deviation between two sample cumulative step functions", *Ann. Math. Stat.*, Vol. 22, No. 1 (1951), pp. 125–131.