

NOTES

JUSTIFICATION AND EXTENSION OF DOOB'S HEURISTIC APPROACH TO THE KOLMOGOROV-SMIRNOV THEOREMS¹

BY MONROE D. DONSKER

University of Minnesota

1. Introduction and summary. Doob [1] has given heuristically an appealing methodology for deriving asymptotic theorems on the difference between the empirical distribution function calculated from a sample and the actual distribution function of the population being sampled. In particular he has applied these methods to deriving the well known theorems of Kolmogorov [2] and Smirnov [3]. In this paper we give a justification of Doob's approach to these theorems and show that the method can be extended to a wide class of such asymptotic theorems.

2. The justification for Kolmogorov's theorem. Let $x_1, x_2, \dots$ be mutually independent, identically distributed random variables with distribution function $F(\lambda)$, and let $\nu_n(\lambda)$ be the number of x_i 's among $x_1, x_2, \dots, x_n$ which are $\leq \lambda$. In studying the difference between the empirical distribution function, $\nu_n(\lambda)/n$, and $F(\lambda)$, Kolmogorov showed that if $F(\lambda)$ is continuous, the distribution of

$$(2.1) \quad \text{l.u.b.}_{-\infty < \lambda < +\infty} \left(\frac{\nu_n(\lambda)}{n} - F(\lambda) \right)$$

is independent of $F(\lambda)$. For convenience, therefore, we will assume that the variables are uniformly distributed on $(0, 1)$, that is, $F(\lambda) = \lambda$ for $0 \leq \lambda \leq 1$. Let²

$$(2.2) \quad D_n^+ = \text{l.u.b.}_{0 \leq \lambda \leq 1} \left(\frac{\nu_n(\lambda)}{n} - \lambda \right).$$

One of Kolmogorov's theorems states

$$(2.3) \quad \lim_{n \rightarrow \infty} P\{n^{\frac{1}{2}} D_n^+ \leq \alpha\} = 1 - e^{-2\alpha^2},$$

and for our purposes it will be sufficient to justify Doob's method for this particular theorem since the justification of the method in general follows from it. Following Doob, define

$$(2.4) \quad x_n(t) = n^{\frac{1}{2}} \left(\frac{\nu_n(t)}{n} - t \right), \quad 0 \leq t \leq 1.$$

¹ Research begun while the writer was a member of an ONR sponsored project in probability at Cornell University.

² For ease of comparison, we are using Doob's notation wherever possible.

Clearly,

$$(2.5) \quad \begin{aligned} E\{x_n(t)\} &= 0, & 0 \leq t \leq 1, \\ E\{[x_n(t) - x_n(s)]^2\} &= (t-s)[1 - (t-s)], & 0 \leq s \leq t \leq 1. \end{aligned}$$

Let $\{x(t)\}$ be a one parameter family of random variables, $0 \leq t \leq 1$, with the properties:

(a) for each j , if $0 \leq t_1 < \cdots < t_j \leq 1$, the j -variate distribution of the variables $x(t_1), x(t_2), \cdots, x(t_j)$ is Gaussian;

$$(2.6) \quad \begin{aligned} (b) \quad E\{x(t)\} &= 0, & 0 \leq t \leq 1, \\ E\{[x(t) - x(s)]^2\} &= (t-s)[1 - (t-s)], & 0 \leq s \leq t \leq 1; \end{aligned}$$

$$(c) \quad P\{x(0) = 0\} = 1.$$

The $x(t)$ process can be selected so that with probability one it has continuous sample functions. Let Y be the space of these sample functions. The $x(t)$ process selected here is such that for any j , if $0 \leq t_1 < \cdots < t_j \leq 1$, and if $(\alpha_1, \alpha_2, \cdots, \alpha_j)$ is an arbitrary vector, we have from the central limit theorem

$$(2.7) \quad \lim_{n \rightarrow \infty} P\{x_n(t_1) \leq \alpha_1; i = 1, 2, \cdots, j\} = P\{x(t_i) \leq \alpha_i; i = 1, 2, \cdots, j\}.$$

Doob's heuristic argument consisted in assuming that in calculating asymptotic $x_n(t)$ process distributions when $n \rightarrow \infty$, one could replace the $x_n(t)$ process by the $x(t)$ process. In particular, with reference to (2.3), his assumption was that

$$(2.8) \quad \lim_{n \rightarrow \infty} P\{n^{\frac{1}{2}} D_n^+ \leq \alpha\} = P\{D^+ \leq \alpha\},$$

where $D^+ = \max_{0 \leq t \leq 1} x(t)$. What we wish to show, therefore, is that

$$(2.9) \quad \lim_{n \rightarrow \infty} P\left\{\text{l.u.b.}_{0 \leq t \leq 1} \left[n^{\frac{1}{2}} \left(\frac{\nu_n(t)}{n} - t\right)\right] \leq \alpha\right\} = P\left\{\max_{0 \leq t \leq 1} x(t) \leq \alpha\right\}.$$

Let E_n be the event that for all t in $(0, 1)$, $\nu_n(t) \leq \alpha n^{\frac{1}{2}} + nt$, and let E be the event that for all t in $(0, 1)$, $x(t) \leq \alpha$. We can write (2.9) as

$$(2.10) \quad \lim_{n \rightarrow \infty} P\{E_n\} = P\{E\}.$$

Let E'_n be the event that for all $i = 1, 2, \cdots, n$, $\nu_n(i/n) \leq \alpha n^{\frac{1}{2}} + i$, and let E''_n be the event that for all $i = 1, 2, \cdots, n$, $\nu_n(i/n) \leq \alpha n^{\frac{1}{2}} + i - 1$. We have, clearly, $E''_n \subset E_n \subset E'_n$. In what follows we will show that

$$(2.11) \quad \lim_{n \rightarrow \infty} P\{E'_n\} = P\{E\},$$

and an exactly similar argument shows $\lim_{n \rightarrow \infty} P\{E''_n\} = P\{E\}$. Hence, we will have shown (2.10).

To show (2.11), let N be a Poisson distributed random variable with mean n and independent of the random variables $x_1, x_2, x_3, \dots$. We have, clearly,

$$(2.12) \quad P\{E'_n\} = P\{\nu_N\left(\frac{i}{n}\right) \leq \alpha n^{\frac{1}{2}} + i; i = 1, 2, \dots, n \mid N = n\}.$$

Let $y_1 = \nu_N(1/n)$, $y_i = \nu_N(i/n) - \nu_N((i-1)/n)$, $i = 2, 3, \dots, n$. The variables $y_1, y_2, \dots, y_n$ are independent (cf. Kac [4]), are Poisson distributed with mean 1, and if we let $z_i = y_i - 1$, $i = 1, 2, \dots, n$, $s_m = z_1 + z_2 + \dots + z_m$, then s_m is a sum of independent variables and we can rewrite (2.12) as

$$(2.13) \quad P\{E'_n\} = P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, n \mid s_n = 0\}.$$

Now,

$$(2.14) \quad 1 - P\{E'_n\} = \sum_{r=1}^n P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, s_r > \alpha n^{\frac{1}{2}} \mid s_n = 0\}.$$

Let k be a fixed positive integer; define $n_j = [jn/k]$, $j = 0, 1, 2, \dots, k$, and let an $\epsilon > 0$ be given. From (2.14) we obtain

$$(2.15) \quad \begin{aligned} 1 - P\{E'_n\} &= \sum_{j=0}^{k-1} \sum_{n_j < r \leq n_{j+1}} P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, \\ &\quad s_r > \alpha n^{\frac{1}{2}}, |s_{n_{j+1}} - s_r| < \epsilon n^{\frac{1}{2}} \mid s_n = 0\} \\ &+ \sum_{j=0}^{k-1} \sum_{n_j < r \leq n_{j+1}} P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, s_r > \alpha n^{\frac{1}{2}}, \\ &\quad |s_{n_{j+1}} - s_r| \geq \epsilon n^{\frac{1}{2}} \mid s_n = 0\}. \end{aligned}$$

Let $P_{n,k}(\alpha) = P\{s_{n_j} \leq \alpha n^{\frac{1}{2}}; j = 1, 2, \dots, k \mid s_n = 0\}$. Clearly,

$$(2.16) \quad P\{E'_n\} \leq P_{n,k}(\alpha),$$

and also the first sum on the right of (2.15) is less than $1 - P_{n,k}(\alpha - \epsilon)$. The second sum on the right of (2.15) can be written as (cf. Chung [5], pp. 39-41)

$$(2.17) \quad \begin{aligned} &\frac{n! e^n}{n^n} \sum_{j=0}^{k-1} \sum_{n_j < r \leq n_{j+1}} P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, s_r > \alpha n^{\frac{1}{2}}, \\ &\quad |s_{n_{j+1}} - s_r| \geq \epsilon n^{\frac{1}{2}}, s_n = 0\} \\ &= \frac{n! e^n}{n^n} \left[\sum_{j=0}^{k-2} \sum_{n_j < r \leq n_{j+1}} P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, s_r > \alpha n^{\frac{1}{2}}\} \right. \\ &\quad \cdot \sum_y P\{|s_{n_{j+1}} - s_r| \geq \epsilon n^{\frac{1}{2}}, s_{n_{j+1}} = y\} P\{s_n - s_{n_{j+1}} = -y\} \\ &\quad \left. + \sum_{n_{k-1} < r \leq n_k} P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, s_r > \alpha n^{\frac{1}{2}}, \right. \\ &\quad \left. |s_n - s_r| \geq \epsilon n^{\frac{1}{2}}, s_n = 0\} \right]. \end{aligned}$$

To estimate the first term in the brackets we note that since the z_i 's are distributed as follows:

$$P\{z_i = m - 1\} = \frac{e^{-1}}{m!}, \quad m = 0, 1, 2, \dots,$$

we have, noting the maximum term of the Poisson distribution,

$$(2.18) \quad P\{s_n - s_{n+1} = -y\} \leq A_1 k^{\frac{1}{2}} n^{-\frac{1}{2}},$$

where A_1 is an absolute constant. Also, from Tchebycheff's inequality we get

$$(2.19) \quad \sum_y P\{|s_{n+1} - s_r| \geq \epsilon n^{\frac{1}{2}}, s_{n+1} = y\} = P\{|s_{n+1} - s_r| \geq \epsilon n^{\frac{1}{2}}\} \leq \frac{1}{k\epsilon^2}.$$

The first term in the brackets on the right of (2.17) is therefore less than $A_1 k^{-\frac{1}{2}} n^{-\frac{1}{2}} \epsilon^{-2}$.

The second term in the brackets on the right of (2.17) is less than

$$\sum_{n_{k-1} < r \leq n} \sum_{y > \alpha n^{\frac{1}{2}}} P\{s_i \leq \alpha n^{\frac{1}{2}}; i = 1, 2, \dots, r-1, s_r = y\} P\{s_n - s_r = -y\},$$

and using similar estimates is shown to be less than $A_2 k^{-\frac{1}{2}} n^{-\frac{1}{2}}$, where A_2 is an absolute constant. Thus, we have from (2.15)

$$(2.20) \quad 1 - P\{E'_n\} \leq 1 - P_{n,k}(\alpha - \epsilon) + \frac{n!e^n}{n^n} \frac{A_3}{k^{\frac{1}{2}} n^{\frac{1}{2}} \epsilon^2}.$$

This together with (2.16) gives us

$$(2.21) \quad P_{n,k}(\alpha - \epsilon) - \frac{n!e^n}{n^n} \frac{A_3}{k^{\frac{1}{2}} n^{\frac{1}{2}} \epsilon^2} \leq P\{E'_n\} \leq P_{n,k}(\alpha).$$

From (2.7) we have

$$(2.22) \quad \lim_{n \rightarrow \infty} P_{n,k}(\alpha) = P\left\{x\left(\frac{i}{k}\right) \leq \alpha, i = 1, 2, \dots, k\right\}.$$

If in (2.21) we hold k and ϵ fixed and let $n \rightarrow \infty$, we get from (2.22) and Stirling's formula that

$$(2.23) \quad \begin{aligned} P\left\{x\left(\frac{i}{k}\right) \leq \alpha - \epsilon; i = 1, 2, \dots, k\right\} - \frac{\sqrt{2\pi} A_3}{k^{\frac{1}{2}} \epsilon^2} &\leq \lim_{n \rightarrow \infty} P\{E'_n\} \\ &\leq \overline{\lim}_{n \rightarrow \infty} P\{E'_n\} \leq P\left\{x\left(\frac{i}{k}\right) \leq \alpha; i = 1, 2, \dots, k\right\}. \end{aligned}$$

In (2.23), if we hold ϵ fixed and let $k \rightarrow \infty$ we get from the continuity of the $x(t)$ process that

$$\begin{aligned} P\{x(t) \leq \alpha - \epsilon, t \in (0, 1)\} &\leq \varliminf_{n \rightarrow \infty} P\{E'_n\} \leq \overline{\lim}_{n \rightarrow \infty} P\{E'_n\} \\ &\leq P\{x(t) \leq \alpha, t \in (0, 1)\}. \end{aligned}$$

Now finally, using the fact that the distribution function of $\max_{0 \leq t \leq 1} x(t)$ is continuous, and letting $\epsilon \rightarrow 0$ we obtain the desired statement (2.11).

3. Extension. Having shown that

$$(3.1) \quad \lim_{n \rightarrow \infty} P\{\text{l.u.b. } x_n(t) \leq \alpha\} = P\{\max_{0 \leq t \leq 1} x(t) \leq \alpha\},$$

it is possible, using methods identical to those used by the writer in a recent paper (Donsker [6]), to obtain a general theorem like (3.1), but where the functional $\max_{0 \leq t \leq 1} x(t)$ is replaced by an arbitrary functional $F[x(t)]$ subject to certain restrictions. Indeed, we can obtain the following theorem.

THEOREM. *Let R be the space of real, single-valued functions $g(t)$ which are continuous on $0 \leq t \leq 1$ except for at most a finite number of finite jumps. Let $F[g]$ be a functional defined on R and continuous in the uniform topology at almost all points of Y^3 . Then,*

$$(3.2) \quad \lim_{n \rightarrow \infty} P\{F[x_n(t)] \leq \alpha\} = P\{F[x(t)] \leq \alpha\}$$

at all points of continuity of the distribution function on the right.

This theorem is proved (precisely as is the main theorem in [6]) by first obtaining (3.2) for functionals of the form $f(u_1, u_2, \dots, u_{2k})$, where $u_i = \sup g(t)$ for $(i-1)/k < t \leq i/k$ and $u_{k+i} = \inf g(t)$ for $(i-1)/k < t \leq i/k$, $i = 1, 2, \dots, k$, where $f(u_1, u_2, \dots, u_{2k})$ as a function of its $2k$ variables is bounded on the whole space, Borel measurable and Riemann integrable on every finite $2k$ -dimensional interval. Such a theorem is obtainable from (3.1), and moreover these functionals can be used to approximate functionals $F[g]$ which are bounded on R and continuous in the uniform topology at almost all points of Y . The approximation is such that (3.2) can be obtained for this latter class of functionals. Finally, the assumption that $F(g)$ be bounded on R may be removed, and hence we can obtain the theorem stated above, by considering the functional $e^{uF(g)}$ and using the continuity theorem for characteristic functions.

REFERENCES

- [1] J. L. DOOB, "Heuristic approach to the Kolmogorov-Smirnov theorems," *Annals of Math. Stat.*, Vol. 20 (1949), pp. 393-403.
- [2] A. KOLMOGOROV, "Sulla determinazione empirica di una legge di distribuzione," *Giorn. Ist. Ital. Attuari*, Vol. 4 (1933), pp. 83-91.
- [3] N. SMIRNOV, "Sur les écarts de la courbe de distribution empirique," *Rec. Math. (Mat. Sbornik) (NS)*, Vol. 6 (1939), pp. 3-26.
- [4] M. KAC, "On deviations between theoretical and empirical distribution functions," *Proc. Nat. Acad. Sci.*, Vol. 35 (1949), pp. 252-257.
- [5] K. L. CHUNG, "An estimate concerning the Kolmogorov limit distribution," *Trans. Am. Math. Soc.*, Vol. 67 (1949), pp. 36-50.
- [6] M. D. DONSKER, "An invariance principle for certain probability limit theorems," *Memoirs Am. Math. Soc.*, No. 6, 1951, 12 pp.

³ The space Y is defined above just after (2.6).