NOTES

JUSTIFICATION AND EXTENSION OF DOOB'S HEURISTIC APPROACH TO THE KOLMOGOROV-SMIRNOV THEOREMS[1]

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 $x_1, x_2, \cdots$ 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 $x_1, x_2, \cdots, x_n$ which are $\leqq \lambda$. 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

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

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 $0 \leqq \lambda \leqq 1$. Let[2]

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

One of Kolmogorov's theorems states

$$\text{(2.3)} \qquad \lim_{n\to\infty} P\{n^{\frac{1}{2}} D_n^+ \leqq \alpha\} = 1 - e^{-2\alpha^2},$$

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

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

[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,

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

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

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

$$\text{(b)}\qquad E\{x(t)\} = 0, \qquad 0 \leqq t \leqq 1,$$

$$(2.6)\qquad E\{[x(t) - x(s)]^2\} = (t - s)[1 - (t - s)], \qquad 0 \leqq s \leqq t \leqq 1;$$

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

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 $0 \leqq t_1 < \cdots < t_j \leqq 1$, and if $(\alpha_1, \alpha_2, \cdots, \alpha_j)$ is an arbitrary vector, we have from the central limit theorem

$$(2.7)\qquad \lim_{n\to\infty} P\{x_n(t_i) \leqq \alpha_i;\, i = 1, 2, \cdots, j\} = P\{x(t_i) \leqq \alpha_i;\, i = 1, 2, \cdots, j\}.$$

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

$$(2.8)\qquad \lim_{n\to\infty} P\{n^{\frac{1}{2}} D_n^+ \leqq \alpha\} = P\{D^+ \leqq \alpha\},$$

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

$$(2.9)\qquad \lim_{n\to\infty} P\left\{\operatorname*{l.u.b.}_{0\leqq t\leqq 1}\left[n^{\frac{1}{2}}\left(\frac{\nu_n(t)}{n} - t\right)\right] \leqq \alpha\right\} = P\left\{\max_{0\leqq t\leqq 1} x(t) \leqq \alpha\right\}.$$

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

$$(2.10)\qquad \lim_{n\to\infty} P\{E_n\} = P\{E\}.$$

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

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

and an exactly similar argument shows $\lim_{n\to\infty} P\{E''_n\} = P\{E\}$. 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 $x_1, x_2, x_3, \cdots$. We have, clearly,

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

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

$$P\{E'_n\} = P\{s_i \leqq \alpha n^{1/2}; \quad i = 1, 2, \cdots, n \mid s_n = 0\}. \tag{2.13}$$

Now,

$$1 - P\{E'_n\} = \sum_{r=1}^{n} P\{s_i \leqq \alpha n^{1/2}; i = 1, 2, \cdots, r-1, s_r > \alpha n^{1/2} \mid s_n = 0\}. \tag{2.14}$$

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

$$\begin{aligned} 1 - P\{E'_n\} = &\sum_{j=0}^{k-1} \sum_{n_j < r \leqq n_{j+1}} P\{s_i \leqq \alpha n^{1/2}; i = 1, 2, \cdots, r-1, \\ &\qquad s_r > \alpha n^{1/2}, |s_{n_{j+1}} - s_r| < \epsilon n^{1/2} \mid s_n = 0\} \\ &+ \sum_{j=0}^{k-1} \sum_{n_j < r \leqq n_{j+1}} P\{s_i \leqq \alpha n^{1/2}; i = 1, 2, \cdots, r-1, s_r > \alpha n^{1/2}, \\ &\qquad |s_{n_{j+1}} - s_r| \geqq \epsilon n^{1/2} \mid s_n = 0\}. \end{aligned} \tag{2.15}$$

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

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

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)

$$\begin{aligned} &\frac{n!\,e^n}{n^n} \sum_{j=0}^{k-1} \sum_{n_j < r \leqq n_{j+1}} P\{s_i \leqq \alpha n^{1/2}; i = 1, 2, \cdots, r-1, s_r > \alpha n^{1/2}, \\ &\qquad |s_{n_{j+1}} - s_r| \geqq \epsilon n^{1/2}, s_n = 0\} \\ &= \frac{n!\,e^n}{n^n} \Bigg[\sum_{j=0}^{k-2} \sum_{n_j < r \leqq n_{j+1}} P\{s_i \leqq \alpha n^{1/2}; i = 1, 2, \cdots, r-1, s_r > \alpha n^{1/2}\} \\ &\qquad \cdot \sum_{y} P\{|s_{n_{j+1}} - s_r| \geqq \epsilon n^{1/2}, s_{n_{j+1}} = y\} P\{s_n - s_{n_{j+1}} = -y\} \\ &\qquad + \sum_{n_{k-1} < r \leqq n_k} P\{s_i \leqq \alpha n^{1/2}; i = 1, 2, \cdots, r-1, s_r > \alpha n^{1/2}, \\ &\qquad\qquad |s_n - s_r| \geqq \epsilon n^{1/2}, s_n = 0\} \Bigg]. \end{aligned} \tag{2.17}$$

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

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

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

$$(2.18) \qquad P\{s_n - s_{n_{j+1}} = -y\} \leqq A_1 k^{\frac{1}{2}} n^{-\frac{1}{2}},$$

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

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

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

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

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)

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

This together with (2.16) gives us

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

From (2.7) we have

$$(2.22) \qquad \lim_{n \to \infty} P_{n,k}(\alpha) = P\left\{x\left(\frac{i}{k}\right) \leqq \alpha, \quad i = 1, 2, \cdots, k\right\}.$$

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

$$(2.23) \qquad \begin{aligned} P\left\{x\left(\frac{i}{k}\right) \leqq \alpha - \epsilon;\ i = 1, 2, \cdots, k\right\} - \frac{\sqrt{2\pi}\, A_3}{k^{\frac{1}{2}} \epsilon^2} &\leqq \varliminf_{n \to \infty} P\{E'_n\} \\ \leqq \varlimsup_{n \to \infty} P\{E'_n\} &\leqq P\left\{x\left(\frac{i}{k}\right) \leqq \alpha;\ i = 1, 2, \cdots, k\right\}. \end{aligned}$$

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

$$\begin{aligned} P\{x(t) \leqq \alpha - \epsilon, \quad t \in (0, 1)\} &\leqq \varliminf_{n \to \infty} P\{E'_n\} \leqq \varlimsup_{n \to \infty} P\{E'_n\} \\ &\leqq P\{x(t) \leqq \alpha, t \in (0, 1)\}. \end{aligned}$$

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

3. Extension. Having shown that

$$\lim_{n \to \infty} P\{\underset{0 \leqq t \leqq 1}{\text{l.u.b.}}\ x_n(t) \leqq \alpha\} = P\{\max_{0 \leqq t \leqq 1} x(t) \leqq \alpha\}, \tag{3.1}$$

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 $\max_{0 \leqq t \leqq 1} x(t)$ 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 $0 \leqq t \leqq 1$ 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,*

$$\lim_{n \to \infty} P\{F[x_n(t)] \leqq \alpha\} = P\{F[x(t)] \leqq \alpha\} \tag{3.2}$$

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 $f(u_1, u_2, \cdots, u_{2k})$, where $u_i = \sup g(t)$ for $(i-1)/k < t \leqq i/k$ and $u_{k+i} = \inf g(t)$ for $(i-1)/k < t \leqq i/k$, $i = 1, 2, \cdots, k$, where $f(u_1, u_2, \cdots, u_{2k})$ 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 $e^{itF(g)}$ 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 une 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).