STUDIA MATHEMATICA, T. XVII. (1958)

A limit theorem for empirical distribution functions *

by

M. FISZ (Warszawa)

1. Let $(x_{11}, x_{12}, \ldots, x_{1n_1})$ and $(x_{21}, x_{22}, \ldots, x_{2n_2})$ be two independent simple samples drawn from a population with a continuous distribution function, *i. e.* let $x_{11}, x_{12}, \ldots, x_{1n_1}, x_{21}, x_{22}, \ldots, x_{2n_2}$ be independent observations of a random variable X having a continuous distribution function. Denote by $S_{1n_1}(x)$ and $S_{2n_2}(x)$ the empirical distribution functions of the two samples, *i. e.*, if $x'_{j1}, x'_{j2}, \ldots, x'_{jn_j}$ $(j = 1, 2)$ are the values of $x_{j1}, x_{j2}, \ldots, x_{jn_j}$ arranged in increasing magnitude, $S_{jn_j}(x)$ is given by the formula

$$S_{jn_j}(x) = \begin{cases} 0 & \text{for} \quad x \leqslant x'_{j1}, \\ k/n_j & \text{for} \quad x'_{jk} < x \leqslant x'_{j(k+1)} \quad (k = 1, 2, \ldots, n_j - 1), \\ 1 & \text{for} \quad x > x'_{jn_j}. \end{cases}$$

Define

$$D^+_{n_1n_2} = \max_x [S_{1n_1}(x) - S_{2n_2}(x)], \qquad D_{n_1n_2} = \max_x |S_{1n_1}(x) - S_{2n_2}(x)|.$$

Smirnov [5], [6] has shown that if $n_2/n_1 = \alpha > 0$ the relations

$$\lim_{n_1 \to \infty} P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D^+_{n_1n_2} < \lambda\right) = 1 - \exp(-2\lambda^2),$$

$$\lim_{n_1 \to \infty} P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D_{n_1n_2} < \lambda\right) = Q(\lambda)$$

hold for every $\lambda > 0$, where $Q(\lambda)$ is the function found by Kolmogorov [3] given by the formula

$$(*) \qquad Q(\lambda) = \sum_{r=-\infty}^{\infty} (-1)^r \exp(-2\lambda^2 r^2).$$

---

* The paper contains a proof of a theorem of the author published (Bull. Pol. Acad. Sci. 5 (1957), p. 699) without proof.

In our paper a limit theorem for 3 samples is given which can be used for statistical purposes similar to that of Smirnov's theorems.

2. Let us consider 3 simple samples drawn independently from a population in which the random variable X has a continuous distribution function. Let n_j $(j = 1, 2, 3)$ denote the number of elements of the j-th sample and let the relations

$$\lim_{n_1\to\infty} \frac{n_j}{n_1} = a_j \qquad (j = 2, 3) \tag{1}$$

hold. Further let $S_{jn_j}(x)$ $(j = 1, 2, 3)$ denote the empirical distribution function of the j-th sample and

$$n_{ij} = \frac{n_i n_j}{n_i + n_j} \qquad (i, j = 1, 2, 3;\ i \neq j). \tag{2}$$

We now define two stochastic processes in the following way:

$$Y_{n_1n_2n_3}(x) = \frac{(\sqrt{n_{12}} + \sqrt{n_{13}})S_{1n_1}(x) - \sqrt{n_{12}}S_{2n_2}(x) - \sqrt{n_{13}}S_{3n_3}(x)}{\sqrt{2}\sqrt{1 + \frac{1}{n_1}\sqrt{n_{12}n_{13}}}},$$

$$Z_{n_1n_2n_3}(x) = \frac{(\sqrt{n_{12}} - \sqrt{n_{13}})S_{1n_1}(x) - \sqrt{n_{12}}S_{2n_2}(x) + \sqrt{n_{13}}S_{3n_3}(x)}{\sqrt{2}\sqrt{1 - \frac{1}{n_1}\sqrt{n_{12}n_{13}}}}. \tag{3}$$

If $n_1 = n_2 = n_3 = n$ the formulae (2) and (3) are of the form

$$n_{ij} = \frac{n}{2} \qquad (i, j = 1, 2, 3;\ i \neq j), \tag{2'}$$

$$Y_n(x) = \sqrt{\frac{2}{3}n}\left[S_{1n}(x) - \frac{S_{2n}(x) + S_{3n}(x)}{2}\right],$$

$$Z_n(x) = \sqrt{\frac{1}{2}n}\,[S_{3n}(x) - S_{2n}(x)]. \tag{3'}$$

We consider the functionals

$$A^+_{n_1n_2n_3} = \max_x Y_{n_1n_2n_3}(x), \qquad A_{n_1n_2n_3} = \max_x |Y_{n_1n_2n_3}(x)|,$$

$$B^+_{n_1n_2n_3} = \max_x Z_{n_1n_2n_3}(x), \qquad B_{n_1n_2n_3} = \max_x |Z_{n_1n_2n_3}(x)|. \tag{4}$$

We prove the following

THEOREM. *Let* $A^+_{n_1n_2n_3}$, $A_{n_1n_2n_3}$, $B^+_{n_1n_2n_3}$ *and* $B_{n_1n_2n_3}$ *be defined by formulae* (2)-(4) *and let the* n_j $(j = 2, 3)$ *satisfy* (1). *Then:*

(i) *For arbitrary positive* a *and* b *the following relations hold:*

$$\lim_{n_1\to\infty} P(A^+_{n_1n_2n_3} < a, B^+_{n_1n_2n_3} < b) = [1 - \exp(-2a^2)][1 - \exp(-2b^2)], \tag{5}$$

$$\lim_{n_1\to\infty} P(A_{n_1n_2n_3} < a, B_{n_1n_2n_3} < b) = Q(a)Q(b), \tag{5'}$$

where $Q(\lambda)$ *is given above by formula* (*).

(ii) *If we denote by* $\max(A, B)$ *the greatest of the numbers* A *and* B, *the following relations hold for every* $\lambda > 0$:

$$\lim_{n_1\to\infty} P(\max(A^+_{n_1n_2n_3}, B^+_{n_1n_2n_3}) < \lambda) = [1 - \exp(-2\lambda^2)]^2, \tag{6}$$

$$\lim_{n_1\to\infty} P(\max(A_{n_1n_2n_3}, B_{n_1n_2n_3}) < \lambda) = [Q(\lambda)]^2. \tag{6'}$$

Proof. The idea of the proof consists in showing that $A^+_{n_1n_2n_3}$, and $B^+_{n_1n_2n_3}$ (resp. $A_{n_1n_2n_3}$ and $B_{n_1n_2n_3}$) are asymptoticaly independent, as $n_1 \to \infty$ (thus in virtue of (1) $n_2 \to \infty$ and $n_3 \to \infty$). This is shown — following a fruitful idea of Doob [2] — by reducing the problem considered to that of finding the probability distributions of some functionals defined on a Gaussian stochastic process.

Without restricting the generality of our considerations we can assume — since the distribution function of X is supposed to be continuous — that X has a uniform distribution in the interval $[0,1]$. We easily observe that for every value of x, where $0 \leqslant x \leqslant 1$, we have

$$E[Y_{n_1n_2n_3}(x)] = E[Z_{n_1n_2n_3}(x)] = 0 \tag{7}$$

and for every pair (x_1, x_2), where $0 \leqslant x_1 \leqslant x_2 \leqslant 1$, we have

$$E[Y_{n_1n_2n_3}(x_1)Y_{n_1n_2n_3}(x_2)] = E[Z_{n_1n_2n_3}(x_1)Z_{n_1n_2n_3}(x_2)] = x_1(1 - x_2). \tag{8}$$

We now consider three independent, equally distributed Gaussian stochastic processes $\eta_1(x)$, $\eta_2(x)$ and $\eta_3(x)$, where $0 \leqslant x \leqslant 1$, whose means, variances and covariances are given by formulae (7) and (8). In other words, the vector $\{\eta_j(x_1), \ldots, \eta_j(x_m)\}$ $(j = 1, 2, 3)$ is normally distributed for $m = 1, 2, 3, \ldots$ and for arbitrary points $x_1, \ldots, x_m$, where $0 \leqslant x_1 \leqslant \ldots \leqslant x_m \leqslant 1$, and, moreover, the relations

$$E[\eta_j(x)] = 0 \qquad (0 \leqslant x \leqslant 1), \tag{9}$$

$$E[\eta_j(x_1)\eta_j(x_2)] = x_1(1 - x_2) \qquad (0 \leqslant x_1 \leqslant x_2 \leqslant 1) \tag{10}$$

hold for $j = 1, 2, 3$. From (9) and (10) follows $P(\eta_j(0) = 0) = 1$.

Let us now define two stochastic processes by the formulae

$$Y(x) = \frac{\left(\sqrt{\frac{a_2}{1+a_2}} + \sqrt{\frac{a_3}{1+a_3}}\right)\eta_1(x) - \frac{1}{\sqrt{1+a_2}}\eta_2(x) - \frac{1}{\sqrt{1+a_3}}\eta_3(x)}{\sqrt{2}\sqrt{1+\sqrt{\frac{a_2 a_3}{(1+a_2)(1+a_3)}}}},$$

(11)

$$Z(x) = \frac{\left(\sqrt{\frac{a_2}{1+a_2}} - \sqrt{\frac{a_3}{1+a_3}}\right)\eta_1(x) - \frac{1}{\sqrt{1+a_2}}\eta_2(x) + \frac{1}{\sqrt{1+a_3}}\eta_3(x)}{\sqrt{2}\sqrt{1-\sqrt{\frac{a_2 a_3}{(1+a_2)(1+a_3)}}}},$$

where the a_j are given by (1). It is easily found that $Y(x)$ and $Z(x)$ are also Gaussian processes with the same means, variances and covariances as the $\eta(x)$ processes and that for every pair of points (x_1, x_2) the equality $E[Y(x_1)Z(x_2)] = 0$ holds. The last equality implies that the processes $Y(x)$ and $Z(x)$ are independent.

Let us now rewrite formulae (3) in the following way:

$$Y_{n_1n_2n_3}(x) = \frac{(\sqrt{n_{12}}+\sqrt{n_{13}})[S_{1n_1}(x)-x] - \sqrt{n_{12}}[S_{2n_2}(x)-x] - \sqrt{n_{13}}[S_{3n_3}(x)-x]}{\sqrt{2}\sqrt{1+\frac{1}{n_1}\sqrt{n_{12}n_{13}}}},$$

$$Z_{n_1n_2n_3}(x) = \frac{(\sqrt{n_{12}}-\sqrt{n_{13}})[S_{1n_1}(x)-x] - \sqrt{n_{12}}[S_{2n_2}(x)-x] + \sqrt{n_{13}}[S_{3n_3}(x)-x]}{\sqrt{2}\sqrt{1-\frac{1}{n_1}\sqrt{n_{12}n_{13}}}}.$$

Let us now remark that for $m = 1, 2, 3, \ldots$ and for arbitrary points $x_1, \ldots, x_m$, where $0 \leqslant x_1 \leqslant \ldots \leqslant x_m \leqslant 1$, the central limit theorem implies the convergence, as $n_1 \to \infty$, of the probability function of the vector

$$\{Y_{n_1n_2n_3}(x_1), \ldots, Y_{n_1n_2n_3}(x_m), Z_{n_1n_2n_3}(x_1), \ldots, Z_{n_1n_2n_3}(x_m)\}$$

to that of the vector

$$\{Y(x_1), \ldots, Y(x_m), Z(x_1), \ldots, Z(x_m)\}.$$

Let us now consider the functionals

$$A^+ = \max_x Y(x); \qquad B^+ = \max_x Z(x);$$

$$A = \max_x |Y(x)|; \qquad B = \max_x |Z(x)|.$$

(We can write max in these equalities since the realizations of the processes $Y(x)$ and $Z(x)$ are continuous with probability 1). We are now aimed to obtain for arbitrary positive numbers a and b the relations

$$\lim_{n_1\to\infty} P(A^+_{n_1n_2n_3} < a, B^+_{n_1n_2n_3} < b) = P(A^+ < a)P(B^+ < b), \tag{12}$$

$$\lim_{n_1\to\infty} P(A_{n_1n_2n_3} < a, B_{n_1n_2n_3} < b) = P(A < a)P(B < b). \tag{12'}$$

Let P_{n_j} and P_j $(j = 1, 2, 3)$ denote the probability measures generated by the processes $\sqrt{n_j}[S_{jn_j}(x)-x]$ and $\eta_j(x)$ respectively in the space $D[0,1]$ of functions $\varphi(x)$ defined on the interval $[0,1]$ having left- and right-hand limits and continuous at least from one side at each point ([4], p. 227-229). Donsker [1] has shown that

$$P_{n_j} \Rightarrow P_j. \tag{13}$$

Consider now the Cartesian product-space

$$\mathcal{D} = D_1[0,1]\times D_2[0,1]\times D_3[0,1].$$

Let $Q_{n_1n_2n_3}$ and Q denote the probability measures generated in $\mathcal{D}$ by the vector-processes $\{\sqrt{n_1}[S_{1n_1}(x)-x], \ldots, \sqrt{n_3}[S_{3n_3}(x)-x]\}$ and $\{\eta_1(x), \eta_2(x), \eta_3(x)\}$ respectively. The independence of $S_{jn_j}(x)$ $(j = 1, 2, 3)$ and relation (13) imply

$$Q_{n_1n_2n_3} = P_{n_1}\times P_{n_2}\times P_{n_3} \Rightarrow P_1\times P_2\times P_3 = Q. \tag{14}$$

Denote by π and π_{n_1} the transformations of the space $\mathcal{D}$ into the space $\mathcal{D}'$ given respectively by the system of linear equations (11) and by a modified system (11) with a_j replaced by n_j/n_1. The sequence $\{\pi_{n_1}\}$ converges uniformly to π on every compact set in $\mathcal{D}$. Consequently from (14) and a theorem of Prohorov ([4], Theorem 1.10) as well as from the normality and independence of $Y(x)$ and $Z(x)$ the relation

$$Q^{\pi_{n_1}}_{n_1n_2n_3} \Rightarrow Q^{\pi} = Q'_1\times Q'_2 \tag{15}$$

follows, where $Q^{\pi_{n_1}}_{n_1n_2n_3}$ and Q^{π} are probability measures generated in $\mathcal{D}'$ by the vector-processes $\{Y_{n_1n_2n_3}(x), Z_{n_1n_2n_3}(x)\}$ and $\{Y(x), Z(x)\}$ respectively, whereas Q'_1 and Q'_2 are probability measures corresponding to $Y(x)$ and $Z(x)$ respectively, concentrated at a subset of continuous functions in $D[0,1]$. Since the transformations of the space $\mathcal{D}'$ into two-dimensional Euclidean spaces given by the correspondences

$$(\varphi_1, \varphi_2) \to (\sup_x \varphi_1, \sup_x \varphi_2)$$

$$(\varphi_1, \varphi_2) \to (\sup_x |\varphi_1|, \sup_x |\varphi_2|)$$

are almost everywhere (Q^n) continuous in $\mathcal{D}'$, we obtain from (15) — using again a theorem of Prohorov ([4], Theorem 1.8) — the relations (12) and (12′).

As Doob [2] has shown the equalities

$$P(A^+ < \lambda) = P(B^+ < \lambda) = 1-\exp(-2\lambda^2), \tag{16}$$

$$P(A < \lambda) = P(B < \lambda) = Q(\lambda) \tag{16′}$$

hold for every positive λ. We obtain formula (5) from formulae (12) and (16) and formula (5′) from formulae (12′) and (16′). Assertion (i) of our theorem is thus proved.

Assertion (ii) of our theorem immediately follows from the assertion (i). Indeed, formulae (12) and (12′) imply the relations

$$\lim_{n_1\to\infty} P\{\max(A^+_{n_1n_2n_3}, B^+_{n_1n_2n_3}) < \lambda\} = P\{\max(A^+, B^+) < \lambda\}, \tag{17}$$

$$\lim_{n_1\to\infty} P\{\max(A_{n_1n_2n_3}, B_{n_1n_2n_3}) < \lambda\} = P\{\max(A, B) < \lambda\}, \tag{17′}$$

respectively. Taking into account the independence of A^+ and B^+, we obtain at once relation (6) from relations (17) and (16). Relation (6′) follows from (17′) and (16′).

Assertion (ii) is thus also proved.

The theorem proved here can be used in an obvious way for statistical purposes. We can verify the hypothesis that three simple samples have been drawn from populations with the same continuous distribution function, which we do not specify. This hypothesis will be rejected for large n_1, n_2, n_3 if, for instance, $\max(A_{n_1n_2n_3}, B_{n_1n_2n_3}) > \lambda_\alpha$, where α is the significance level and $[Q(\lambda_\alpha)]^2 = 1-\alpha$.

We remark that our theorem remains true if (1) is replaced by the assumption that n_j/n_1 are bounded and $n_1, n_2, n_3 \to \infty$. Indeed, one can then choose subsequences of the indices n satisfying (1) for some values (which may be different) of α. Since the right sides of (5)-(6′) do not depend on α, the assertions of theorem remain valid.

References

[1] M. D. Donsker, *Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems*, Annals of Math. Stat. 23 (1952), p. 277-281.

[2] J. L. Doob, *Heuristic approach to the Kolmogorov-Smirnov theorems*, Annals of Mathematical Statistics 20 (1949), pp. 393-403.

[3] A. N. Kolmogorov, *Sulla determinazione empirico di una legge di distribuzione*, Giornale dell'Istituto Italiano d.Attuari 4 (1933), pp. 83-91.

[4] Ю. В. Прохоров, *Сходимость случайных процессов и предельные теоремы теории вероятностей*, Теория вероятностей-Применения 1 (1956), p. 177-238.

[5] Н.В. Смирнов, *Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках*, Бюлл. Московского Государственного Университета, сер. А, II вып. 2 (1939), p. 3-14.

[6] Н.В. Смирнов, *Приближение законов распредсления случайных величин по эмпирическим данным*, Успехи Математических Наук 10 (1944), pp. 176-206.

INSTYTUT MATEMATYCZNY POLSKIEJ AKADEMII NAUK
MATHEMATICAL INSTITUTE OF THE POLISH ACADEMY OF SCIENCES

Reçu par la Rédaction le 5. 4. 1957