SOME SMIRNOV TYPE THEOREMS OF PROBABILITY THEORY[1]

By Miklós Csörgő

Princeton University and McGill University

1. Introduction. Let $\xi_{11}, \xi_{12}, \cdots, \xi_{1n}$ and $\xi_{21}, \xi_{22}, \cdots, \xi_{2m}$ be two samples of mutually independent random variables having a continuous distribution function $F(t)$. Let $F_{1n}(t)$ and $F_{2m}(t)$ be the corresponding empirical distribution functions. In 1939 Smirnov [10] proved the following two theorems:

$$(1.1) \qquad \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{-\infty<t<+\infty} (F_{1n}(t) - F_{2m}(t)) < y\} = 1 - e^{-2y^2},$$

if $y > 0$, zero otherwise, and

$$(1.2) \qquad \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{-\infty<t<+\infty} |F_{1n}(t) - F_{2m}(t)| < y\} = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2y^2},$$

if $y > 0$, zero otherwise.

In both cases $N = nm/(n + m)$, and $N \to \infty$ is to mean that $n \to \infty$, $m \to \infty$ so that $m/n \to \rho$, where ρ is a constant. (The problem of determining the exact distributions of the respective random variables for finite values of n and m was solved by Koroljuk [6] on the assumption that $m = np$ where p is an integer.)

Results (1.1) and (1.2) are used to test the statistical hypothesis that two random samples come from the same unknown population. Even if $F(t)$, the hypothetical distribution function of the two random samples in question, was assumed to have a specific form, we would not get more information out of these theorems, for they consider the supremum of the difference $(F_{1n}(t) - F_{2m}(t))$ and that of its absolute value with the same weight 1, regardless of the value of $F(t)$. Thus in this way the idea arises of considering the limit distribution of the quotients $\{F_{1n}(t) - F_{2m}(t)\}/F(t)$ and $|F_{1n}(t) - F_{2m}(t)|/F(t)$, with the natural limitation on $F(t)$ that we restrict ourselves to an interval $t_a \leqq t < +\infty$, where $F(t_a) = a > 0$, when taking the supremum of these random variables. The value of a can be arbitrarily close to zero.

2. Statement, discussion, and consequences of theorems. Using the notation and assumptions of Section 1 we are going to prove the following theorems (for the definition of the distribution functions $\Phi(\cdot)$, $L(\cdot)$, $N(\cdot)$ and $R(\cdot)$ of Theorems 1, 2, 3 and 4 we refer the reader to (3.4), (3.5), (3.6) and (3.7) of [8] respectively).

Theorem 1.

$$(2.1) \quad \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F(t)} (F_{1n}(t) - F_{2m}(t))/F(t) < y\} = \Phi(y\{a/[1 - a]\}^{\frac{1}{2}}),$$

if $y > 0$, $0 < a < 1$, *zero otherwise.*

Received 19 February 1965; revised 25 March 1965.

[1] This research was partially supported by the Canadian National Research Council at McGill University and by the Office of Naval Research 042-023, Contract Nonr 1858(05) to Princeton University.

THEOREM 2.

$$\lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F(t)} |F_{1n}(t) - F_{2m}(t)|/F(t) < y\} = L(y\{a/[1-a]\}^{\frac{1}{2}}), \tag{2.2}$$

if $y > 0$, $0 < a < 1$, *zero otherwise.*

THEOREM 3.

$$\lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F(t) \leqq b} (F_{1n}(t) - F_{2m}(t))/F(t) < y\} = N(y; a, b), \tag{2.3}$$

where $-\infty < y < +\infty$, $0 < a < b < 1$.

THEOREM 4.

$$\lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F(t) \leqq b} |F_{1n}(t) - F_{2m}(t)|/F(t) < y\} = R(y; a, b), \tag{2.4}$$

if $y > 0$, $0 < a < b < 1$, *zero otherwise.*

As a reminder we note that N of above theorems is defined as in Section 1 and $N \to \infty$ is also meant as explained there. The theorems themselves are Smirnov type analogies of Rényi's Kolmogorov-Smirnov type theorems in [8].

To prove these theorems we need only the assumption that $F(t)$ is continuous but the statistical application of them requires the assumption of a specific form for $F(t)$. Assuming then that $F(t)$ is a known continuous distribution function, the above theorems provide tests for verifying the statistical hypothesis that two random samples come from the same population with distribution function $F(t)$. The character of these tests is that, in case of Theorems 1 and 3, they give upper bounds below which the difference of two empirical distribution functions in question must lie with given probabilities and the width of this upper bound is proportional at all its points t to $F(t)$. In case of Theorems 2 and 4 we have proportional bands (two-sided confidence intervals) instead.

From (1.1) it follows that

$$\lim_{N\to\infty} P\{\sup_{-\infty<t<+\infty} (F_{1n}(t) - F_{2m}(t)) < 0\} = 0, \tag{2.5}$$

that is, the probability of the event that one empirical distribution function does not exceed the other one all along the interval $-\infty < t < +\infty$, tends to zero as $N \to \infty$. It follows from Theorem 1 that the same is true for the interval $t_a < t < +\infty$, that is, we have:

COROLLARY 1.

$$\lim_{N\to\infty} P\{\sup_{t_a<t<+\infty} (F_{1n}(t) - F_{2m}(t)) < 0\} = 0. \tag{2.6}$$

On the other hand we have from Theorem 3:

COROLLARY 2.

$$\lim_{N\to\infty} P\{\sup_{t_a \leqq t \leqq t_b} (F_{1n}(t) - F_{2m}(t)) < 0\} = 1/\pi \arcsin \{a(1-b)/b(1-a)\}^{\frac{1}{2}}, \tag{2.7}$$

that is, the probability of the event that one empirical distribution function does not exceed the other one all along the interval $t_a \leqq t \leqq t_b$, remains positive in the limit.

A similar statement was proved by Rényi [8] and Gihman [5] concerning the asymptotic behavior of an empirical and theoretical distribution as follows:

$$\lim_{n\to\infty} P\{\sup_{t_a \leqq t \leqq t_b} (F_{1n}(t) - F(t)) < 0\} = 1/\pi \arcsin \{a(1 - b)/b(1 - a)\}^{\frac{1}{2}}. \tag{2.8}$$

(Note that the right sides of (2.7) and (2.8) are the same.)

It follows from Theorem 5 of [3] that we also have:

THEOREM 1*.

$$\begin{aligned} &\lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F_{1n}(t)} (F_{1n}(t) - F_{2m}(t))/F_{1n}(t) < y\} \\ &= \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F_{2m}(t)} (F_{1n}(t) - F_{2m}(t))/F_{2m}(t) < y\} \\ &= \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{t \varepsilon T} (F_{1n}(t) - F_{2m}(t))/F_{1n}(t) < y\} \\ &= \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{t \varepsilon T} (F_{1n}(t) - F_{2m}(t))/F_{2m}(t) < y\} \\ &= \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{t \varepsilon T} (F_{1n}(t) - F_{2m}(t))/F_{n+m}(t) < y\} \\ &= \lim_{N\to\infty} P\{N^{\frac{1}{2}} \sup_{a \leqq F_{n+m}(t)} (F_{1n}(t) - F_{2m}(t))/F_{n+m}(t) < y\} \\ &= \Phi(y\{a/[1 - a]\}^{\frac{1}{2}}), \quad \textit{if} \quad y > 0, \quad 0 < a < 1, \quad \textit{zero otherwise}, \end{aligned} \tag{2.9}$$

where $\Phi(\cdot)$ *is the same distribution function as that of* (2.1) *and where* $T = \{t: a \leqq F_{1n}(t)\} \cap \{t: a \leqq F_{2m}(t)\}$, *and* $F_{n+m}(t)$ *is the empirical distribution function of the random sample gained by pooling the two random samples of size n and m respectively.*

Similar starred versions of Theorems 2, 3 and 4 are also true. These starred versions can be used to construct confidence intervals for the difference of two empirical distribution functions, and can also be used to test the statistical hypothesis that two random samples come from the same population without requiring the assumption of a specific form for $F(t)$, while we had to do so when the unstarred versions of these theorems were used to test this statistical hypothesis. Thus starred versions of Theorems 1, 2, 3 and 4 retain the advantage of Theorems 1, 2, 3 and 4 that they measure the relative deviation of two sample distribution functions and regain the convenient property of (1.1) and (1.2) that, in applications too, nothing has to be assumed about the form of $F(t)$ beyond its continuity.

3. Sketch of proof of Theorems 1, 2, 3 and 4, by means of adapting the ideas of Rényi's proof of his Kolmogorov-Smirnov type theorems in [8]. Without loss of generality we may assume $F(t) = t$ with t uniformly distributed on [0, 1]. Let $F_{1n}(t)$, $F_{2m}(t)$ be two empirical distribution functions constructed by selecting two random samples of size n and m respectively from this uniform distribution on [0, 1]. Accordingly, we will have to derive the limit distribution of the random variable $\sup_{a \leqq t} (F_{1n}(t) - F_{2m}(t))/t$, which, in turn, can be shown (using the Glivenko-Cantelli theorem) to have the same limit distribution as $\sup_{a \leqq F_{2m}(t)} (F_{1n}(t) - F_{2m}(t))/t$. Using Theorem 5 of [3] it is seen that the later random variable has the same limit distribution as $\sup_{a \leqq F_{2m}(t)} (F_{1n}(t) - F_{2m}(t))/F_{2m}(t)$. Again using the Glivenko-Cantelli theorem one can easily show that the later random variable has the same limit distribution as

$$\sup_{t \varepsilon T'} [(F_{1n}(t)/t)/(F_{2m}(t)/t) - 1], \tag{3.1}$$

where $T' = \{t \colon a \leqq F_{1n}(t)\} \cap \{t \colon a \leqq F_{2m}(t)\} \cap \{t \colon a \leqq F_{n+m}(t)\}$ and where $F_{n+m}(t)$ is the empirical distribution function of the random sample gained by pooling the random samples of size n and m respectively.

Having got this far we remark here that assuming the validity of Theorems 1, 2, 3 and 4 for a moment and using Theorem 5 of [3] we can conclude that starred versions of these theorems, as specified in Section 2, are true.

Continuing, let $\eta_{11} < \eta_{12} < \cdots < \eta_{1n}$ and $\eta_{21} < \eta_{22} < \cdots < \eta_{2m}$ be the two order statistics corresponding to the two random samples of size n and m respectively, taken on the uniformly distributed $t \varepsilon [0, 1]$ and let $\eta_1 < \eta_2 < \cdots < \eta_{n+m}$ be the order statistics gained by pooling these random samples of size n and m. Now the only places where one or the other of the step functions $F_{1n}(t)$, $F_{2m}(t)$ change their value are at η_k, $k = 1, \cdots, n + m$. Also, $F_{1n}(t)$ and $F_{2m}(t)$ are constants in any interval $\eta_k \leqq t < \eta_{k+1}$, $k = 1, \cdots, n + m$. Thus, concerning the random variable of (3.1), we have the following inequality:

$$\max_S \left[\frac{F_{1n}(\eta_{k+1} - 0)/\eta_{k+1}}{F_{2m}(\eta_k + 0)/\eta_k} - 1\right] \leqq \sup_{t \varepsilon T'} \left[\frac{F_{1n}(t)/t}{F_{2m}(t)/t} - 1\right] \leqq \max_S \left[\frac{F_{1n}(\eta_k + 0)/\eta_k}{F_{2m}(\eta_{k+1} - 0)/\eta_{k+1}} - 1\right] \tag{3.2}$$

where $S = \{i \colon an \leqq i \leqq n\} \cap \{j \colon am \leqq j \leqq m\} \cap \{k \colon a(n + m) \leqq k \leqq n + m\}$, a finite set corresponding to T' of (3.1). We also have that

$$\text{right hand side of (3.3)} \leqq \max_S \left[\frac{\frac{i}{n} \Big/ \eta_{1i}}{\frac{j}{m} \Big/ \eta_{2j+1}} - 1\right] \tag{3.3}$$

and

$$\max_S \left[\frac{\frac{i}{n} \Big/ \eta_{1i+1}}{\frac{j}{m} \Big/ \eta_{2j}} - 1\right] \leqq \text{left hand side of (3.2)}, \tag{3.4}$$

for: (1) if $\eta_k = \eta_{1j}$, then we have that $\eta_{1i} \varepsilon (\eta_{2j}, \eta_{2j+1})$ for some j, $j = 0, 1, \cdots, m$, with $\eta_{20} = 0$ and $\eta_{2m+1} = 1$ and so $F_{1n}(\eta_k + 0) = F_{1n}(\eta_{k+1} - 0) = i/n$, $F_{2m}(\eta_k + 0) = F_{2m}(\eta_{k+1} - 0) = j/m$; (2) if $\eta_k = \eta_{2j}$, then $\eta_{2j} \varepsilon (\eta_{1i}, \eta_{1i+1})$ for some i, $i = 0, 1, \cdots, n$, with $\eta_{10} = 0$ and $\eta_{1n+1} = 1$ and so $F_{2m}(\eta_k + 0) = F_{2m}(\eta_{k+1} - 0) = j/m$, $F_{1n}(\eta_k + 0) = F_{1n}(\eta_{k+1} - 0) = i/n$.

Relations (3.2), (3.3) and (3.4) imply that

$$P\left\{N^{\frac{1}{2}}\max_S\left[\frac{\frac{i}{n}\Big/\eta_{1i}}{\frac{j}{m}\Big/\eta_{2j+1}}-1\right]<y\right\}$$

$$\leqq P\left\{N^{\frac{1}{2}}\sup_{t\varepsilon T'}\left[\frac{F_{1n}(t)/t}{F_{2m}(t)/t}-1\right]<y\right\} \tag{3.5}$$

$$\leqq P\left\{N^{\frac{1}{2}}\max_S\left[\frac{\frac{i}{n}\Big/\eta_{1i+1}}{\frac{j}{m}\Big/\eta_{2j}}-1\right]<y\right\}.$$

What follows after this is a straightforward argument, adapting Rényi's method of proof of his theorems in [8], showing that, as $N \to \infty$, the right and left hand expressions of (3.5) tend to the same limit, namely to $\Phi(\cdot)$ of (2.1). This proves Theorem 1. The proof of Theorem 3 can be based on the same ideas, and the above kind argument leads to functions $L(\cdot)$ and $R(\cdot)$ in case of Theorems 2 and 4 as lower estimates for the respective statements.

4. Proof of Theorems 1, 2, 3 and 4, using Rényi's results by means of the invariance principle. This alternative and shorter way of proving Theorems 1, 3 and sharpening 2, 4 to their present form was discovered by the referee of this paper. Following his advice I attach this proof here and express my deep appreciation to him for it.

Let $C[0, 1]$ be the space of continuous functions on $[0, 1]$. In Appendix 1 of [7] Yu. V. Prokhorov studies the functions φ which are defined on $[0, 1]$ and have at every point, right and left limiting values. The distance function $d(\varphi_1, \varphi_2)$ ((2) of p. 206 of [7]) is introduced, in relation to which the totality of these functions becomes a complete separable metric space (Theorem 1, p. 206 of [7]). This space is called $D[0, 1]$. We define here $\tilde{D} = D[0, 1] \times D[0, 1]$ and let φ_i be the projections of $\tilde{D}$ onto its ith coordinate space ($i = 1, 2$). $\tilde{D}$ is then a complete separable metric space in the sense of [7], Theorem 1, p. 206. Let

$$\xi_{1n}(t) = n^{\frac{1}{2}}(F_{1n}(t) - t)$$

$$\xi_{2m}(t) = m^{\frac{1}{2}}(F_{2m}(t) - t).$$

We define $\Lambda_{n,m} = (\xi_{1n}(t), \xi_{2m}(t))$, which is a random variable with values in $\tilde{D}$, and for any n, m, $\varphi_1\Lambda_{n,m}$ and $\varphi_2\Lambda_{n,m}$ are independent random variables. According to Donsker's theorem ([7], p. 187, Theorem 2.4) we have

$$\xi_{1n}(t) \Rightarrow X_1 \tag{4.1}$$

$$\xi_{2m}(t) \Rightarrow X_2$$

where $\Rightarrow$ means weak convergence in the sense of [7], p. 164, of the respective distributions and X_i ($i = 1, 2$) are two copies of the Gaussian process on $[0, 1]$ with mean 0 and covariance $r(s, t) = s(1 - t)$, $0 \leqq s \leqq t \leqq 1$, considered as random variables with values in $D[0, 1]$ (in fact in $C[0, 1]$). Also X_1 and X_2 are independent random variables. It follows then from (4.1) that

$$\Lambda_{n,m} \Rightarrow (X_1, X_2)$$

if m and n both go to ∞, for consider sets $G_1 \times G_2$ with $G_i \subset D[0, 1]$ open, $P\{X_i \,\varepsilon\, \text{boundary of } G_i\} = 0$, $i = 1, 2$ and apply Theorem 1.9, p. 165 of [7].

Let us define the map

$$\Psi_\alpha = \alpha^{\frac{1}{2}}\varphi_1 + (1 - \alpha)^{\frac{1}{2}}\varphi_2 \tag{4.2}$$

(where $0 \leqq \alpha \leqq 1$) from $\tilde{D}$ to $D[0, 1]$. Then Ψ_α is continuous at all points of $C[0, 1] \times C[0, 1]$ and

$$P\{(X_1, X_2) \,\varepsilon\, C[0, 1] \times C[0, 1]\} = 1.$$

Thus, we can apply Theorem 1.8 of [7] and conclude that

$$\Psi_\alpha\Lambda_{n,m} \Rightarrow \Psi_\alpha(X_1, X_2) = X, \tag{4.3}$$

where X is again a copy of the Gaussian process with mean 0 and covariance $r(s, t) = s(1 - t)$. Now

$$N^{\frac{1}{2}}(F_{1n}(t)\; F_{2m}(t)) = (m/(n + m))^{\frac{1}{2}}\xi_{1n}(t) - (n/(n + m))^{\frac{1}{2}}\xi_{2m}(t)$$
$$= (\rho/(1 + \rho))^{\frac{1}{2}}\xi_{1n}(t) - (1/(1 + \rho))^{\frac{1}{2}}\xi_{2m}(t) + \Gamma_{n,m} = \Psi_{\rho/(1+\rho)}\Lambda_{n,m} + \Gamma_{n,m},$$

where $\Gamma_{n,m} \to 0$ if $m/n \to \rho$ and $N = nm/(n + m) \to \infty$. Therefore, by (4.2) and (4.3), we have

$$N^{\frac{1}{2}}(F_{1n}(t) - F_{2m}(t)) \Rightarrow X.$$

Again by Theorem 1.8 of [7] it follows for $0 < a < b \leqq 1$ that

$$\sup_{t\varepsilon[a,b]} N^{\frac{1}{2}}(F_{1n}(t) - F_{2m}(t))/t \Rightarrow \sup_{t\varepsilon[a,b]} X(t)/t = I$$
$$\sup_{t\varepsilon[a,b]} N^{\frac{1}{2}}\, |F_{1n}(t) - F_{2m}(t)|/t \Rightarrow \sup_{t\varepsilon[a,b]}\, |X(t)|/t = II.$$

Similarly, (4.1) and Theorem 1.8 of [7] imply

$$\sup_{t\varepsilon[a,b]}\, \xi_{1n}(t)/t \Rightarrow \sup_{t\varepsilon[a,b]}\, X_1(t)/t = I'$$
$$\sup_{t\varepsilon[a,b]}\, |\xi_{1n}(t)|/t \Rightarrow \sup_{t\varepsilon[a,b]}\, |X_1(t)|/t = II'.$$

In [8], Theorems 5–8, Rényi determines the distributions of I' and II' which, by above argument, coincide with the distributions of I and II (the processes X_1, X_2 and X have the same distribution). This proves Theorems 1, 2, 3 and 4. Using Theorem 5 of [3] and above argument, the starred versions can also be obtained the same way or, as we have already observed earlier in Section 3, they (the starred versions of Theorems 1, 2, 3, and 4) can also be proved by using Theorems 1, 2, 3, and 4 of this paper and Theorem 5 of [3].

In [1] Anderson and Darling derive the limit distribution of the one-sample Kolmogorov-Smirnov statistics using a generalized nonnegative weight function $\Psi(t)$, $0 \leqq t \leqq 1$. We remark here that, using the above argument, their theorems can also be extended to the two-sample relations of this paper.

5. Acknowledgments. This paper is based partially on the author's doctoral dissertation, written under the supervision of Professor W. A. O'N. Waugh (formerly at McGill University, presently at the University of Hull, England) at McGill University while the author was supported by a Canadian National Research Council Studentship.

REFERENCES

[1] Anderson, T. W., and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Statist.* **23** 193–212.

[2] Csörgő, M. (1963). On some Kolmogorov-Smirnov-Rényi type theorems of probability. Unpublished doctoral dissertation. McGill University.

[3] Csörgő, M. (1965). Some Rényi type limit theorems for empirical distribution functions. *Ann. Math. Statist.* **36** 322–326.

[4] Erdös, P., and Kac, M. (1946). On certain limit theorems of the theory of probability. *Bull. Amer. Math. Soc.* **52** 292–302.

[5] Gihman, T. (1952). On the empirical distribution function in the case of grouping of data (in Russian). *Dokl. Akad. Nauk SSSR* **82** 837–840.

[6] Koroljuk, V. S. (1963). On the discrepancy of empiric distributions for the case of two independent samples. *Select. Transl. Math. Statist. Prob.* **4** 105–121.

[7] Prokhorov, Yu. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theor. Prob. Appl.* **1** 157–214.

[8] Rényi, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hungar.* **4** 191–231.

[9] Rényi, A. (1954). *Probability Theory* (in Hungarian). Tankönyvkiadó, Budapest.

[10] Smirnov, N. (1939). A digression on the empirical distribution curve (in French). *Rec. Maths. N. S.* **6**(98) 3–26.