The Annals of Mathematical Statistics
1971, Vol. 42, No. 4, 1157–1166

ON THE COMPARISON OF TWO EMPIRICAL DISTRIBUTION FUNCTIONS[1]

BY LAJOS TAKÁCS

Case Western Reserve University

1. Introduction. Let $\xi_1, \xi_2, \cdots, \xi_m$ be mutually independent random variables having a common distribution function $F(x)$. Denote by $F_m(x)$ the empirical distribution function of the sample $(\xi_1, \xi_2, \cdots, \xi_m)$. The empirical distribution function $F_m(x)$ is defined as the number of variables $\xi_1, \xi_2, \cdots, \xi_m$ less than or equal to x divided by m.

Furthermore, let $\eta_1, \eta_2, \cdots, \eta_n$ be mutually independent random variables having a common distribution function $G(x)$, and denote by $G_n(x)$ the empirical distribution function of the sample $(\eta_1, \eta_2, \cdots, \eta_n)$.

For the purpose of testing the hypothesis that $F(x) \equiv G(x)$ in 1939 N. V. Smirnov [6] introduced the statistic

$$\delta^+(m, n) = \sup_{-\infty < x < \infty} [F_m(x) - G_n(x)] \tag{1}$$

and showed that if $F(x)$ and $G(x)$ are two identical continuous distribution functions, then the distribution of $\delta^+(m, n)$ does not depend on $F(x) \equiv G(x)$, and

$$\lim_{m \to \infty, n \to \infty} \mathbf{P}\left\{\left(\frac{mn}{m+n}\right)^{\frac{1}{2}} \delta^+(m, n) \leqq x\right\} = 1 - e^{-2x^2} \tag{2}$$

for $x \geqq 0$. In this case the distribution of the random variable $\delta^+(m, n)$ for $n = m$ was found in 1951 by B. V. Gnedenko and V. S. Korolyuk [2], and for $n = mp$ where p is a positive integer in 1955 by V. S. Korolyuk [3]. (See also [7] and [8].) Obviously $\delta^+(m, n)$ and $\delta^+(n, m)$ have the same distribution for all $m = 1, 2, \cdots$ and $n = 1, 2, \cdots$.

We can express $\delta^+(m, n)$ also in a simpler way. Denote by $\eta_1^*, \eta_2^*, \cdots, \eta_n^*$ the random variables $\eta_1, \eta_2, \cdots, \eta_n$ arranged in increasing order of magnitude. Then we can write that

$$\delta^+(m, n) = \max_{1 \leqq r \leqq n} [F_m(\eta_r^*) - G_n(\eta_r^* - 0)]. \tag{3}$$

Now let us introduce another statistic. For any a let us define $\eta_a(m, n)$ as the number of subscripts $r = 1, 2, \cdots, n$ for which

$$G_n(\eta_r^* - 0) \leqq F_m(\eta_r^*) + a/n < G_n(\eta_r^*). \tag{4}$$

Received May 25, 1970.

[1] This research was sponsored by the National Science Foundation under Contract No. GP-9629.