Note on the Kolmogorov Statistic in the Discrete Case[1])

By G. E. NOETHER, Boston[2])

Zusammenfassung: Ein einfacher Beweis des konservativen Charakters des Kolmogoroffschen Tests bei diskreten Verteilungen wird erbracht.

Summary: A simple demonstration of the conservative character of the Kolmogorov test in the case of discrete distributions is given.

The statistic

$$D_n = \text{l. u. b.}_u \left| F_n(u) - F(u) \right|,$$

where $F_n(u)$ is the empirical distribution function of a sample of size n from $F(u)$, was introduced by KOLMOGOROV in 1933. If $F(u)$ is continuous, D_n is distribution free and its usefulness in testing a simple hypothesis about $F(u)$ or finding a confidence band for $F(u)$ is well known.

Let $H_n(d)$ be the distribution function of D_n when $F(u)$ is continuous. In a 1941 paper [1], KOLMOGOROV remarks that quite generally $P(D_n \leqq d) \geqq H_n(d)$. Thus the use of D_n with discrete populations produces conservative results. Mention of this fact is found in various places in the statistical literature. On the other hand, several standard statistics texts do not refer to it and at least one recent text states that the Chi-square test has the advantage that it is applicable in the discrete case while the KOLMOGOROV test is not. For these reasons the following extremely simple demonstration should be of some interest.

Let the discrete random variable Y have distribution function $G(u)$. For simplicity of notation, it will be assumed that $G(u)$ has only a finite number of discontinuities occurring at $u_1 < \ldots < u_m$ with associated probabilities $p_1, \ldots, p_m$. The case of infinitely many discontinuities can be handled in the same way. Let X be a continuous random variable with distribution $F(u)$ that assigns probability p_i to the interval (u_{i-1}, u_i), $i = 1, \ldots, m$, where $u_0 < u_1$. Then

$$G(u_i) = F(u_i). \tag{1}$$

[1]) The study was supported by the United States Air Force Office of Scientific Research.

[2]) Boston University, Boston, Massachusetts, USA.

Any random sample $y_1, \ldots, y_n$ from $G(u)$ can be thought of as having arisen from a random sample $x_1, \ldots, x_n$ from $F(u)$ by setting $y_k = u_i$ if $u_{i-1} < x_k \leqq u_i$, $k = 1, \ldots, n$; $i = 1, \ldots, m$. If $F_n(u)$ is the empirical distribution of the x-sample and $G_n(u)$ that of the y-sample,

$$G_n(u_i) = F_n(u_i). \tag{2}$$

It follows from (1) and (2) that

$$D'_n = \max_{u_i} |G_n(u_i) - G(u_i)| \leqq \text{l. u. b.}_u |F_n(u) - F(u)| = D_n,$$

so that $P(D'_n \leqq d) \geqq H_n(d)$.

The same argument can be used to show that the KOLMOGOROV-SMIRNOV two-sample test produces conservative results when applied to discrete populations.

Reference

[1] KOLMOGOROV, Andrei N. (1941): Confidence limits for an unknown distribution function. Ann. Math. Statistics 12, 461—463.