ON THE DERIVATION OF THE DISTRIBUTION OF THE KOLMOGOROV-SMIRNOV ONE-SAMPLE STATISTIC

ADEL MOHAMED ZAHER
Faculty of Economics and Political Science
Cairo University

1. INTRODUCTION

In several practical situations, the problem of obtaining information about the form of the population from which a sample is obtained is addressed. The compatibility of the observed values in a given sample with a specific distribution can be checked by a goodness-of-fit test. The null hypothesis in such a test is a statement about the form of the cumulative distribution function (cdf) of the parent population. The most widely used goodness-of-fit tests are the chi-square test, proposed by Karl Pearson in 1900, and the Kolmogorov-Smirnov (K-S) test which is the subject of this article.

The Kolmogorov-Smirnov test is one of several goodness-of-fit tests based on the empirical (sample) distribution function, denoted by $F_n(x)$ and defined for all real x as the proportion of the sample values not exceeding x. Specifically, the K-S test is based on the maximum deviation between the empirical distribution function and the distribution function specified by the null hypothesis.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with a continuous distribution function $F_X(x)$ and consider the testing problem : $H : F_X(x) = F_0(x)$ for all x against the alternative hypothesis $A : F_X(x) \neq F_0(x)$ for some x. For any value of x, the empirical distribution function, $F_n(x)$, provides a consistent point estimate for $F_X(x)$. Moreover, the Glivenko-Cantelli theorem states that $F_n(x)$ converges uniformly to $F_X(x)$; i.e., for any $\epsilon > 0$,

$$\lim_{n \to \infty} P\left[\sup_x \mid F_n(x) - F_X(x) \mid > \epsilon\right] = 0.$$

Therefore, for sufficiently large values of n, the deviation between the true cdf and its statistical image provided by the empirical distribution function should be small for all x except for sampling variation. This result suggests the use of the statistic

$$D_n = \sup_x \mid F_n(x) - F_0(x) \mid \tag{1}$$

for the testing problem mentioned above where H is rejected for large values of D_n (with $F_X(x)$ replaced by $F_0(x)$).

The statistic D_n ,called the K-S one-sample statistic, is very useful in nonparametric statistical inference because its sampling distribution does not depend on $F_X(x)$ as long as F_X is continuous (that is, it is distribution-free). Therefore, one can assume without loss of generality that $F_X(x)$ is the uniform distribution on (0,1). However, the derivation of the distribution of D_n is rather tedious (Gibbons, 1971, p. 77). For D_n as defined in (1) where $F_X(x)$ is any continuous function, we have

$$P\left(D_n < \frac{1}{2n} + \nu\right) = \begin{cases} 0 & \text{for } \nu \leq 0 \\ \displaystyle\int_{\frac{1}{2n}-\nu}^{\frac{1}{2n}+\nu} \int_{\frac{3}{2n}-\nu}^{\frac{3}{2n}+\nu} \cdots \int_{\frac{(2n-1)}{2n}-\nu}^{\frac{(2n-1)}{2n}+\nu} f(u_1, u_2, \cdots, u_n)\, du_n \cdots du_1 & \text{for } 0 < \nu \leq \frac{2n-1}{2n} \\ 1 & \text{for } \nu > \frac{2n-1}{2n} \end{cases} \tag{2}$$

where

$$f(u_1, u_2, \cdots, u_n) = \begin{cases} n! & \text{for } 0 < u_1 < u_2 < \cdots < u_n < 1 \\ 0 & \text{otherwise} \end{cases}$$

A proof of this result based on a number of properties of order statistics is given in Gibbons (1971, pp. 78-79). An alternative derivation of the distribution of D_n was obtained by Massey (1950).

According to Gibbons, the result of Eq. (2) is troublesome to evaluate as it must be used with care. For example, when n=2, then for $\nu \in (0, \frac{3}{4}]$,

$$P(D_2 < \tfrac{1}{4} + \nu) = 2! \underset{0 < u_1 < u_2 < 1}{\int_{\frac{1}{4}-\nu}^{\frac{1}{4}+\nu} \int_{\frac{3}{4}-\nu}^{\frac{3}{4}+\nu}} du_2 \, du_1$$

The limits of this double integral can not be determined for the whole interval $(0, \frac{3}{4}]$. Instead, the two subinterval : $(0, \frac{1}{4})$ and $[\frac{1}{4}, \frac{3}{4}]$ must be considered separately. For all ν, the probability is given by :

$$P(D_2 < \tfrac{1}{4} + \nu) = \begin{cases} 0 & \text{for} \quad \nu \le 0 \\ 2(2\nu)^2 & \text{for} \quad 0 < \nu < \frac{1}{4} \\ -2\nu^2 + 3\nu - 0.125 & \text{for} \quad \frac{1}{4} \le \nu \le \frac{3}{4} \\ 1 & \text{for} \quad \nu > \frac{3}{4} \end{cases}$$

For any specific values of ν and n, one can evaluate $P(D_n < \frac{1}{2n} + \nu)$. The inverse procedure, which is more appropriate for inference, is to find that value $D_{n,\alpha}$ such that $P(D_n > D_{n,\alpha}) = \alpha$. The K-S one-sample goodness-of-fit test with significance level α is then to reject $H : F_X(x) = F_0(x)$ for all x when $D_n > D_{n,\alpha}$. Numerical values of $D_{n,\alpha}$ for $\alpha = 0.01$ and $\alpha = 0.05$ have been tabulated for some selected values of n [see, for example, Owen (1962)]. For large samples, Kolmogorov (1933) derived an approximation to the exact distribution of the test statistic D_n.

In the present article, the properties of the multiple integral that defines the probability $P\ (\ D_n < \frac{1}{2n} + \nu\ )$ are investigated; and a partitioning procedure is proposed for the evaluation of this probability. The procedure is then applied to the case n=3 for the sake of illustration.

2. SOME BASIC PROPERTIES OF THE MULTIPLE INTEGRAL DEFINING THE DISTRIBUTION OF D_n

As indicated in Section One, the multiple integral of equation (2) that defines the distribution of the K-S one-sample statistic, D_n, is troublesome to evaluate. In order to avoid limits overlapping, the integral can be written in the following more precise form:

$$P\ (\ D_n < \tfrac{1}{2n} + \nu\ ) = \int_{\max(0,\frac{1}{2n}-\nu)}^{\min(\frac{1}{2n}+\nu,1)} \int_{\max(u_1,\frac{3}{2n}-\nu)}^{\min(\frac{3}{2n}+\nu,1)} \cdots \int_{\max(u_{n-1},\frac{2n-1}{2n}-\nu)}^{\min(\frac{2n-1}{2n}+\nu,1)} n!\ du_n \cdots du_1 \quad \text{for} \quad 0 < \nu \le \tfrac{2n-1}{2n} \qquad (3)$$

The lower limit of the j-th integral, $\max(\ u_{j-1}\ , \frac{2j-1}{2n} - \nu)$, is free from u_{j-1} if $c_j = \frac{2j-1}{2n} - \nu$ is greater than the maximum value taken by u_{j-1} which is equal to $\min\left(\frac{(2j-3)}{2n} + \nu,\ 1\right)$ (the upper limit of the (j-1)st integral). Thus, the lower limits of the multiple integral of Eq. (3) are free from the $u_j's$ if

$$\frac{2j-1}{2n} - \nu \ge \min\left(\frac{(2j-3)}{2n} + \nu,\ 1\right)\ ,\quad j = 2, 3, \cdots, n \qquad (4)$$

For $\nu \le \frac{1}{2n}$, $\min\left(\frac{2j-1}{2n} + \nu,\ 1\right) = \frac{2j-1}{2n}$ for $j = 1, 2, \cdots, n$. Also condition (4) is satisfied. Therefore, for $\nu \le \frac{1}{2n}$, we have

$$P\ (D_n < \tfrac{1}{2n} + \nu\ ) = n! \int_{\frac{1}{2n}-\nu}^{\frac{1}{2n}+\nu} \int_{\frac{3}{2n}-\nu}^{\frac{3}{2n}+\nu} \cdots \int_{\frac{(2n-1)}{2n}-\nu}^{\frac{(2n-1)}{2n}+\nu} du_n \cdots du_1$$
$$= n! \prod_{j=1}^{n} \left(\int_{\frac{(2j-1)}{2n}-\nu}^{\frac{(2j-1)}{2n}+\nu} du_j \right)$$
$$= n!\,(2\,\nu)^n \qquad (5)$$

For $\nu > \frac{1}{2n}$, the lower limits of the multiple integral defining the probability $P\ (D_n < \frac{1}{2n}+\nu\ )$ are not free from the u_j's. Consider the region of integration $R = \left\{\nu : \frac{1}{2n} < \nu \leq \frac{2n-1}{2n}\right\}$ and define the additional order statistic $u_0 = 0$. Partition R into the subregions $R_1,\ R_2, \cdots,\ R_{n-1}$, where

$$R_i = \left\{\nu : \frac{2i-1}{2n} < \nu \leq \frac{2i+1}{2n}\right\} \quad , \quad i = 1, 2, \cdots, n-1 \qquad (6)$$

For the k-th subregion, R_k,the limits of the multiple integral of Eq. (3) have the following properties:

1. The first k lower limits $L_1,\ L_2, \cdots, L_k$ are equal to $u_0,\ u_1, \cdots, u_{k-1}$, respectively.

$$L_j = \max\ \left(u_{j-1},\ \tfrac{2j-1}{2n} - \nu\right) \ , \qquad j = 1, 2, \cdots, n$$

$$\tfrac{2j-1}{2n} - \nu < \tfrac{2j-1}{2n} - \tfrac{2k-1}{2n} \qquad (\text{ since } \nu > \tfrac{2k-1}{2n}\ )$$
$$< \tfrac{(j-k)}{n}$$

As $\frac{(j-k)}{n} \leq 0$ for $j \leq k$, then $\max(\, u_{j-1}\ , \frac{2j-1}{2n} - \nu) = u_{j-1}$, for $j = 1, 2, \cdots, k$.

2. The first n-k upper limits $M_1,\ M_2, \ldots, M_{n-k}$ are equal to $\frac{1}{2n} + \nu,\ \frac{3}{2n} + \nu, \ldots, \frac{2(n-k)-1}{2n} + \nu$, respectively.

$$M_j = \min\ \left(\ \tfrac{2j-1}{2n} + \nu\ ,\ 1\right) \ , \qquad j = 1, 2, \cdots, n$$

$$\tfrac{2j-1}{2n} + \nu \leq \tfrac{2j-1}{2n} + \tfrac{2k+1}{2n} \qquad (\text{ since } \nu \leq \tfrac{2k+1}{2n}\ )$$
$$\leq \tfrac{j+k}{n}$$

Since $\frac{j+k}{n} \le 1$ for $j \le n-k$, it follows that $\min\left(\frac{2j-1}{2n} + \nu ,\ 1\right) = \frac{2j-1}{2n} + \nu$ for $j = 1, 2, \cdots, n-k$.

3. Each of the last k upper limits is equal to one

$$\begin{aligned} \tfrac{2j-1}{2n} + \nu &> \tfrac{2j-1}{2n} + \tfrac{2k-1}{2n} \qquad (\text{ since } \nu > \tfrac{2k-1}{2n}\ ) \\ &> \tfrac{j+k-1}{n} \\ &\ge 1 \quad \text{for} \quad j = n-k+1,\ n-k+2, \cdots, n. \end{aligned}$$

Hence, $\min\left(\frac{2j-1}{2n} + \nu ,\ 1\right) = 1$ for $j = n-k+1,\ n-k+2, \cdots, n$.

Using the above properties, the multiple integral defining the probability $P\ (\ D_n < \frac{1}{2n} + \nu\ )$ for $\nu \in R_k$ can be written as:

$$P\ (\ D_n < \tfrac{1}{2n} + \nu\ ) = \int_{u_0}^{M_1(\nu)} \int_{u_1}^{M_2(\nu)} \cdots \int_{u_{k-1}}^{M_k(\nu)} \int_{\max(u_k,\, \frac{2k+1}{2n} - \nu)}^{M_{k+1}(\nu)} \cdots \int_{\max(u_{n-1},\, \frac{2n-1}{2n} - \nu)}^{M_n(\nu)} n!\, du_n \cdots du_1 \quad \text{for } \nu \in R_k \qquad (7)$$

3. A PROPOSED PROCEDURE FOR THE EVALUATION OF THE PROBABILITY $P(D_n < \frac{1}{2n} + \nu\ )$

As noted earlier, for values of $\nu > \frac{1}{2n}$, the lower limits of the multiple integral defining the probability $P\ (\ D_n < \frac{1}{2n} + \nu\ )$ are not free from the u_j's. In this case, the determination of these limits is a troublesome task. Dealing individually with each subregion $R_k = \left\{ \nu : \frac{2k-1}{2n} < \nu \le \frac{2k+1}{2n} \right\}$, $k = 1, 2, \cdots, n-1$; and the extra partitioning of R_k (if needed), however, would facilitate the accomplishment of this task.

According to Eq. (7), the problem of evaluating the probability $P\ (\ D_n < \frac{1}{2n} + \nu\ )$ for $\nu \in R_k$ is reduced to the problem of determining each of $\max(\ u_k\ ,\ c_{k+1})$, $\max(\ u_{k+1}\ ,\ c_{k+2}), \ldots,$ and $\max(\ u_{n-1}\ ,\ c_n)$ where $c_j = \frac{2j-1}{2n} - \nu$,

$j = k+1, k+2, \ldots, n$. For $\nu \in R_k$ we have

$$P\,(\,D_n < \tfrac{1}{2n} + \nu\,) = I_k = \int \cdots \int_{L_1}^{L_2} \int_{\max(u_k\,,c_{k+1})}^{M_{k+1}(\nu)} \cdots \int_{\max(u_{n-1},\ c_n)}^{M_n(\nu)} n!\, du_n \cdots du_1 \qquad \text{for } \nu \in R_k \qquad (8)$$

where L_1 and L_2 are the limits of the k-th integral corresponding to the order statistic u_k (L_1 and L_2 depend on k. They are not written here as $L_1(k)$ and $L_2(k)$ just for the sake of simplicity).

In order to determine $\max(u_k\,, c_{k+1})$, the following procedure is used :

1. $\text{Max}(u_k\,, c_{k+1})$ can be directly simplified to either u_k or c_{k+1} under the following conditions:

(a) $L_1 \leq L_2$ or $L_2 \leq L_1$ for all values of $\nu \in R_k$, and

(b) $$\max(u_k\,, c_{k+1}) = \begin{cases} c_{k+1} & \text{if } a_1 \leq c_{k+1} \qquad \text{for all } \nu \in R_k \\ u_k & \text{if } a_2 \geq c_{k+1} \qquad \text{for all } \nu \in R_k \end{cases}$$

where $a_1 = \min(\,L_1,\ L_2)$ and $a_2 = \max(\,L_1,\ L_2)$

2. If $\max(u_k\,, c_{k+1})$ is not determined in step (1), then the interval $[\,L_1,\ L_2]$ is divided into the two intervals : $[\,L_1,\ c_{k+1})$ and $(\,c_{k+1},\ L_2]$ and the integral I_k can then be written as:

$$I_k = \int \cdots \int_{L_1}^{c_{k+1}} \int_{\max(u_k\,,c_{k+1})}^{M_{k+1}(\nu)} \cdots \int_{\max(u_{n-1},\ c_n)}^{M_n(\nu)} n!\, du_n \cdots du_1$$

$$+ \int \cdots \int_{c_{k+1}}^{L2} \int_{\max(u_k\,,c_{k+1})}^{M_{k+1}(\nu)} \cdots \int_{\max(u_{n-1},\ c_n)}^{M_n(\nu)} n!\, du_n \cdots du_1 \qquad (9)$$

Conditions (a) and (b) of step (1) are then checked. If $\max(u_k\,, c_{k+1})$ is still not determined in either integral of I_k, then another partition is needed; in which

The probability: $P\ (D_3 < \frac{1}{6} + \nu\ )$ for $\nu \in R_1$ is finally given by :

$$P\ (D_3 < \tfrac{1}{6} + \nu\ ) = \begin{cases} 8\nu^2 - 12\nu^3 + \nu - \frac{1}{9} & \text{for } \ \frac{1}{6} < \nu \leq \frac{1}{3} \\ \frac{11}{3}\nu - 4\nu^3 - \frac{11}{27} & \text{for } \ \frac{1}{3} < \nu \leq \frac{1}{2} \end{cases}$$

For $\nu \in R_2$, we have

$$P\ (\ D_3 < \tfrac{1}{6} + \nu\ ) = I_2 = \int_0^{\frac{1}{6}+\nu} \int_{u_1}^{1} \int_{\max(u_2,\ \frac{5}{6}-\nu)}^{1} 3!\ du_3\ du_2\, du_1$$

$L_1 =\ \ u_1$, $L_2 = 1$, and $c_3 = \frac{5}{6} - \nu$. $\quad L_1 \leq L_2$ for all $\nu \in R_2$

$a_1 =\ \ u_1$ and $a_2 = 1$

$a_2 \leq\ \ c_3$ if $\nu \leq -\frac{1}{6}$ (false in R_2)

$a_1 \geq\ \ c_3$ if $\nu \geq \frac{5}{6}$ (false in R_2 except at $\nu = \frac{5}{6}$)

Hence, I_2 is partitioned as follows :

$$\begin{aligned} I_2 &= \int_0^{\frac{1}{6}+\nu} \int_{u_1}^{\frac{5}{6}-\nu} \int_{\max(u_2,\ \frac{5}{6}-\nu)}^{1} 3!\ du_3\ du_2\, du_1 \\ &+ \int_0^{\frac{1}{6}+\nu} \int_{\frac{5}{6}-\nu}^{1} \int_{\max(u_2,\ \frac{5}{6}-\nu)}^{1} 3!\ du_3\ du_2\, du_1 \\ &= I_{21} + I_{22} \end{aligned}$$

I_{21} has to be partitioned into I_{211} and I_{212} for which $\max(u_2,\ \frac{5}{6} - \nu)$ is $\frac{5}{6} - \nu$ and u_2, respectively. No partitioning is needed for I_{22} for which $\max(u_2,\ \frac{5}{6} - \nu)$ is u_2.

Collecting the results for all ν, the probability $P\ (D_3 < \frac{1}{6} + \nu\ )$ is given by :

$$P\ (D_3 < \tfrac{1}{6} + \nu\ ) = \begin{cases} 0 & \text{for} \quad \nu \le 0 \\ 48\,\nu^3 & \text{for} \quad 0 < \nu \le \frac{1}{6} \\ 8\nu^2 - 12\nu^3 + \nu - \frac{1}{9} & \text{for} \quad \frac{1}{6} < \nu \le \frac{1}{3} \\ \frac{11}{3}\nu - 4\nu^3 - \frac{11}{27} & \text{for} \quad \frac{1}{3} < \nu \le \frac{1}{2} \\ 2\,\nu^3 - 5\,\nu^2 + \frac{25}{6}\nu - \frac{17}{108} & \text{for} \quad \frac{1}{2} < \nu \le \frac{5}{6} \\ 1 & \text{for} \quad \nu > \frac{5}{6} \end{cases}$$

REFERENCES

[1] Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. JASA, **47**, 425-441.

[2] Bradley, J. V. (1968). Distribution-Free Statistical Tests. New Jersy, Prentice-Hall.

[3] Capon, J. (1965). On the asymptotic efficiency of the Kolmogorov-Smirnov test. JASA, **60**, 843-853.

[4] Gibbons, J. D. (1971). Nonparametric Statistical Inference. New York, McGraw-Hill.

[5] Goodman, L. A. (1954). Kolmogorov-Smirnov tests for psychological research. Psychological Bulletin, **51**, 160-168.

[6] Massey, F. J. (1950). A note on the estimation of a distribution function by confidence limits. The Annals of Mathematical Statistics, **21**, 116-119.

[7] Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. JASA, **46**, 68-78.

[8] Noether, G. E. (1976). Introduction to Statistics: A Nonparametric Approach (2nd ed.). Boston, Houghton Mifflin.

[9] Owen, D. B. (1962). Handbook of Statistical Tables. Reading, Mass., Addison-Wesley.

[10] Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics, **19**, 279-281.