THE EGYPTIAN STATISTICAL JOURNAL
RSSR, CAIRO UNIV., VOL., 45, NO.1, 2001

Kolmogorov-Smirnov Statistic and a Markov Chain Property

Ali S. Barakat

Department of Statistics, An-Najah National University, Nablus, West Bank, Palestine. Fax. +(972)9-2387982, E-mail: barakat@najah.edu

Abstract. Kolmogorov-Smironov statistic is rewritten using nearest neighbors techniques and a characterization in terms of a Markov chain is established.

Keywords. Nearest neighbors, Kolmogrov-Smirnov, Markov chain

1. Introduction

Let $X_1, X_2, \ldots\ldots, X_n$ and $Y_1, Y2, \ldots\ldots, Y_m$ be independent random samples in R^d from distributions F(x) and G(x), respectively, with corresponding continuous densities f(x) and g(x).

Construct the combined sample $Z_1, \ldots, Z_N$, where N = n+m, such that

$$Z_i = \begin{cases} x_i & , \quad i = 1,2,\ldots,n \\ y_{i-n} & , \quad i = n+1, n+2, \ldots\ldots, N \end{cases}$$

Let $\|\bullet\|$ be the Euclidean norm, and define the K^{th} nearest neighbor to Z_i as that point Z_j satisfying $\|Z_{j'} - Z_i\| < \|Z_j - Z_i\|$ for exactly (k-1) values of $j'(1 \le j' \le N, \ j' \ne i, j)$. Ties are neglected, since they occur with probability zero. Define also

$$h(i,j)=\begin{cases}1 \ , & \text{if } Z_i \text{ and its } k^{th} \text{ nearest neighbor } ,Z_j, \text{are from different samples} \\ 0 \ , & \text{otherwise}\end{cases}$$

and for i = 1,…, n and k = 1,…., N-1, define

$$T_{i,k}=\sum_{j=1}^{k}h(i,j)$$

Let us define the d-variate Kolmogorov-Smirnov statistic for the i^{th} observation as

$$D_i = \max_k |D_{i,k}|$$

where

$$D_{i,k}=\frac{S_k}{n-1}-\frac{r_k}{m}$$

s_k = number of X observations for which the rank is ≤ k, and r_k = number of Y observations for which the rank is ≤ k where the rank is with respect to distance from the i^{th} observation.

2. A representation for $D_{i,k}$

To write $D_{i,k}$ in terms of the value of $T_{i,k}$, note that in order to satisfy the condition $T_{i,k} = r_k$, the first k nearest neighbors in the combined ordered arrangement of the two samples must include r_k Y's and $(k-r_k)$ X's. Since k is the rank of the k^{th} nearest neighbor in the combined ordered arrangement of the two samples, then in the first k values we have r_k Y's with rank ≤ k and $(k-r_k)$ X's with rank ≤ k..

Therefore,

$$D_{i,k}=\frac{S_k}{n-1}-\frac{r_k}{m}$$

$$=\frac{k-r_k}{n-1}-\frac{r_k}{m}$$

$$= \frac{1}{m(n-1)}[mk - r_k(N-1)]$$

$$= r_k^{'}$$

So,

D_i can be written as

$$D_i = \frac{1}{m(n-1)} \max_k |mk - r_k(N-1)|$$

3. A Markovian Property of $D_{i,k}$

Theorem:

For every $i(1 \leq i \leq n)$, the sequence $\{D_{i,k};\ k = 1,\ldots,N-1\}$ is a Markov chain, i.e., for every $k \leq N-1$

$$P(D_{i,k+1} = r_{k+1}^{'} / D_{i,j} = r_j^{'};\ j \leq k) = P(D_{i,k+1} = r_{k+1}^{'} / D_{i,k} = r_k^{'})$$

Proof:

Let P be the set of all permutations of $(\{1,\ldots,N\} - \{i\})$ satisfying the condition $\{D_{i,1} = r_1^{'}, \ldots\ldots\ldots, D_{i,k} = r_k^{'}\}$. It is clear that for any $p \in P$, $D_{i,k+1}$ can only assume the values $\left(r_k^{'} + \frac{1}{n-1}\right)$ and $\left(r_k^{'} - \frac{1}{m}\right)$. If $\{\alpha_1, \ldots\ldots, \alpha_N\} \in$ P, then the set $\{\alpha_1, \ldots\ldots, \alpha_k\}$ has r_k elements of the set $\{n+1, \ldots, N\}$ and $(k-r_k)$ elements of the set $\{1,\ldots,n\} - \{i\}$.

Then we may have either of the following:

i. $k+1 \in \{n+1,\ldots, N\}$. This happens with the (conditional) probability

$$\frac{m - r_k}{N-1-k}$$

ii. $k+1 \notin \{n+1,\ldots,N\}$. This happens with the (conditional) probability

$$1-\frac{m-r_k}{N-1-k}=\frac{n-1-k+r_k}{N-1-k}$$

In case (i), $D_{i,k+1}$ can assume only the value $\frac{k-r_k}{n-1}-\frac{r_k+1}{m}$ which is equal to $\left(r_k'-\frac{1}{m}\right)$ with probability $\frac{m-r_k}{N-1-k}$, while in case (ii), $D_{i,k+1}$ can assume only the value $\frac{k-r_k+1}{n-1}-\frac{r_k}{m}$ which is equal to $\left(r_k'+\frac{1}{n-1}\right)$ with probability $\frac{n-1-k+r_k}{N-1-k}$.

Thus, the assumable values of $D_{i,k+1}$(viz. $r_k'-\frac{1}{m}$, $r_k'+\frac{1}{n-1}$) and their respective (conditional) probabilities (given the Dij, $j \leq k$) depend only on the value r_k' assumed by $D_{i,k}$.

Note that the distribution of $D_{i,k+1}$ given $D_{i,k}$ can be written in the form:

$$P(D_{i,k+1}=s/D_{i,k}=r)=\begin{cases}\frac{m-r}{N-1-k} & , \quad s=r-\frac{1}{m}\\ \frac{n-1-k+r}{N-1-k} & , \quad s=r+\frac{1}{n-1}\\ 0 & , \quad otherwise\end{cases}$$

Hence, from the distribution of $\{D_{i,k+1} \mid D_{i,k}\}$ we have

$$E(D_{i,k+1}/D_{i,k})=\left(r-\frac{1}{m}\right)\left(\frac{m-r}{N-1-k}\right)+\left(r+\frac{1}{n-1}\right)\left(\frac{n-1-k+r}{N-1-k}\right)$$

$$=\left[1+\frac{N-1}{m(n-1)(N-1-k)}\right]D_{ik}-\frac{k}{(n-1)(N-1-k)}$$

where, k = 1,...., N-1

Note: For i = n+1,...., N, the same resents are obtained but n and m are interchanged.

References:

Bailey, N.T.J. (1964) The Elements of Stochastic Processes. Wiley, New York.

Barakat, A.S., Quade, D., and Salama, I.A. (1996) "Multivariate homogeneity testing using an extended concept of nearest neighbors. "Biometrical Journal, 38, 5, 605-612.

Bishop, Y.M.N., Fienberg, s.E., and Holland, P.W. (1976) Discrete Multivariate Analysis: Theory and Practice. MIT Press, Massachusetts.

Gibbons, J.D. (1985) Nonparametric Statistical Inference. Marcel Dekker, New York.

Randless, R.H. and Wolfe, D.A. (1979) Introduction to the Theory of Nonparametric Statistics. Wiley, New York.

Sen, P.K. and Salama, I.A. (1983) "The Spearman footrule and a Markov chain property". Statistics and Probability Letters, 1, 285-289.