A Nonparametric Test of Independence between Two Sets of Random Variables

임 태 진

한국 원자력 연구소 원자력 안전 연구부

ABSTRACT

A test of independence between two sets of random variables is developed for a simple case: between a random variable and a set of two random variables. This test is developed in a purely nonparametric framework. We do neither impose any restriction on the underlying distributions nor employ a new measure of independence (or correlation).

Suppose we have tri-variate observations $(x_{11},x_{21},y_1), \ldots, (x_{1n},x_{2n},y_n)$ from (X_1,X_2,Y) with unknown joint or marginal distributions. We test whether (X_1,X_2) is independent of Y or not.

First we select almost equally spaced order statistics from X_1 as strip boundaries for X_1-space. Then almost equally spaced conditional order statistics from X_2 for each strip are selected as cell boundaries for (X_1,X_2)-space, so that the set of two random variables (X_1,X_2) is divided into cells with almost equal number of observations.

Second we also select almost equally spaced order staistics from Y as cell boundaries for Y-space. Then each cell of (X_1,X_2)-space is divided into subcells by these selected order staistics from Y. The number of observations falling in each cell is counted and used as statistics for this test.

The number of cells is allowed to depend on the number of available observations, n. That is, it may increase as n increases. This means that we may not have a definite form of

limiting distribution of the cell frequences for this model. For this reason, we do not deal with the weak convergence (or convergence in distribution) of the distribution of cell frequences but deal with the asymptotic equivallence of it to a definite distribution for each n. We prove that the standardized cell frequences are asymptotically equivallent to a set of slightly rounded-off joint normal random variables. From that, we derive a Pearson χ^2-type distribution-free test statistic.

Monte Carlo simulation is employed to evaluate the performance of the proposed test. For some sample sizes, critical values are simulated. Some examples are provided to demonstrate characteristics of this test.