Commun. Fac. Sci. Univ. Ank. Series A1
V. 52. (no.2) pp. 21-27 (2003)

THE KOLMOGOROV GOODNESS-OF-FIT TEST OF INDEPENDENCE BASED ON COPULAS

Bilgehan Güven

Department of Statistics, Middle East Technical University 06531 Ankara, Turkey

(Received: March, 31,2003; Revised: Nov.05, 2003; Accepted: Nov.11, 2003)

ABSTRACT

We present a method which reduces the Kolmogorov goodness-of-test of independence to the Kolmogorov-Smirnov one sample test. The null distribution of test statistic is the same as the Kolmogorov-Smirnov test statistic in this test of independence.

1. INTRODUCTION

The Kolmogorov type test of independence for a pair of random variables has been tackled by many authors. Distribution free tests of independence (Blumm et al. [1], Hoeffding [4]) and the Cramér-von Misses test for independence (De Wet [2], Deheuvels [3]) are the well known examples of Kolmogorov type test. In both the distribution free and the Cramér-von Misses test, the characteristic function of the limiting null distribution of the test statistic is obtained and then the corresponding upper quantiles of it are tabulated.

Saunders and Laud [5] showed that a test statistic in the multidimensional Kolmogorov goodness-of-fit test can be reduced to the Kolmogorov-Smirnov one sample statistic. Consequently, the distribution of the Kolmogorov-Smirnov statistic can be used as the exact null distribution of the test statistic in the multidimensional Kolmogorov goodness-of-fit test.

In this article, it is showed that the Distribution free test for independence, which is one of the two-dimensional Kolmogorov goodness-of fit tests, is reduced to the Kolmogorov-Smirnov one sample test, as the result of [5]. This reduction holds when the marginal distributions of a pair of continuous random variables are completely known. It provides us to have two things: the exact null distribution and the greatest lower bound (g.l.b.) of power of the test.

2. THE TEST

Let a pair of random variables (X,Y) have a bivariate distribution function (d.f.) $F_{X,Y}(x,y,\rho)$ with the fixed marginals $F_X(x)$ and $F_Y(y)$. Then the test of independence is the following

$$H_0 : F_{X,Y}(x,y,\rho) = F_X(x)F_Y(y),$$
$$H_1 : F_{X,Y}(x,y,\rho) \neq F_X(x)F_Y(y), \qquad (1)$$

where $F_{X,Y}(x,y,\rho)$ is a member of the family of bivariate distribution functions depending on the dependence parameter ρ such that

$$F_{X,Y}(x,y,\rho) = F_X(x)F_Y(y)$$

when $\rho = 0$. For example, the bivariate normal distribution with correlation coefficient ρ is a member of such family.

A function $C(u,v,\rho)$ on $1^2 = \{(u,v) : 0 \leq u,v \leq 1\}$ is called a parametrized copula if it satisfies the conditions: $C(u,v,\rho)$ is increasing in both u and v, $C(1,v,\rho) = v$ for every $v \in [0,1]$ and $C(u,1,\rho) = u$ for every $u \in [0,1]$. The functions

$$C(u,v,\rho) = \rho \max\{u+v-1,0\} + (1-\rho)\min\{u,v\}$$

and $C(u,v) = uv\exp(-\rho \ln u \ln v)$, where $(u,v) \in 1^2$ and $\rho \in [0,1]$ are the examples of a copula.

For any bivariate d.f. $F_{X,Y}(x,y,\rho)$ of a pair of continuous random variables, Sklar [6] proved that there exists a unique parametrized copula such that

$$F_{X,Y}(x,y,\rho) = C(u,v,\rho)$$

where $u = F_X(x)$ and $v = F_Y(y)$.

Under $H_0 : F_{X,Y}(x, y, \rho) = F_X(x)F_Y(y)$, the copula representation of $F_{X,Y}(x, y, \rho)$ is $\Pi(u, v) = uv$.

Let $\Pi = UV$ be a random variable produced by the transformations $U = F_X(X)$ and $V = F_Y(Y)$. Define the collection of subsets of 1^2 by $A_p = \{(u,v) : uv \le p\}$ and $0 \le p \le 1$. A_p is measurable and $A_{p1} \subset A_{p2}$ if $p_1 < p_2$. Then there exists a unique $c(p)$ for a given p such that $P(A_{c(p)}) = p$ where $P(A_{c(p)}) = F_\Pi(c(p))$ and $F_\Pi(.)$ is the d.f. of Π. It follows that, for a given p, the critical region can uniquely be determined by $F_\Pi(.)$, when the transformed observation $u_i v_i = F_X(x_i)F_Y(y_i)$ is used instead of $(x_i, y_i) i = 1,2..., n$. It should be noted that the transformation

$$T : (x_i y_i) \rightarrow F_X(x_i)F_Y(y_i)$$

is not one-to-one, however, as it is explained above, the critical region is uniquely determined.

When continuous random variables X and Y are independent and whose marginal distribution functions $F_X(x), F_Y(y)$ are completely known, the d.f. of Π is the d.f. of the product of two independent uniform random variables on [0,1] and is given by

$$F_0(w) = w(1 - \ln w), \qquad 0 \le w \le 1. \qquad (2)$$

Let $F_\Pi(w)$ denote the d.f. of Π. Then, the hypotheses in (1) can be rewritten as:

$$H_0 : F_\Pi(w) = F_0(w),$$
$$H_1 : F_\Pi(w) \ne F_0(w). \qquad (3)$$

Thus, the testing problem in (1) is reduced to the Kolmogorov-Smirnov one sample test.

For the hypotheses in (3), the Kolmogorov-Smirnov one sample test statistic D_n, based on a sample of size n, is given by

$$D_n = \sup_{0 \le w \le 1} \left| S_n(w) - w(1 - \ln w) \right|,$$

where

$$S_n(w) = \frac{1}{n} \sum_{i=1}^{n} \delta(w - U_i V_i),$$

and $\delta(t)$ is the d.f. of a point mass at the origin; $\delta(t) = 0$, if $t < 0$ and $\delta(t) = 1$, if $t \ge 0$ and $u_i = F_X(x_i), v_i = F_Y(y_i)$.

We reject $H_0 : F_\Pi(w) = F_0(w)$, if $D_n > d_{n,\alpha}$. Numerical values of the percentage point $d_{n,\alpha}$ of the distribution of D_n have been tabulated for selected values of n when $\alpha = 0.01$ and $\alpha = 0.05$ and can be found in any book of the statistical tables.

The power P of the Kolmogorov-Smirnov test for the hypotheses in (3) is:

$$P = P\left(\sup_{0 \le w \le 1} \left| S_n(w) - w(1 - \ln w) \right| > d_{n,\alpha} \middle| H_1 \right), \tag{4}$$

where $S_n(w)$ and $d_{n,\alpha}$ are defined before.

3. THE BOUND FOR THE POWER OF THE TEST

In this section, the greatest lower bound (g.l.b.) of the power of the test is obtained and tabulated, when $F_{X,Y}(x, y, \rho)$ is a member of the Farlie-Gumbel-Morgenstern family and it is defined as:

$$F_{X,Y}(x, y, \rho) = F_X(x)F_Y[1 + 3\rho(1 - F_X(x))(1 - F_Y(y))], \tag{5}$$

with $\rho \in [-1/3, 1/3]$.

The copula representation of (5) is :

$$C(u, v, \rho) = uv[1 + 3\rho(1-u)(1-v)] \quad 0 \le u, v \le 1. \tag{6}$$

Lemma : Let the distributions of random variables X and Y be a member of the Farlie-Gumbel-Morgenstern family. Then the g.l.b. of the power of the test is:

$$P > 1 - P(0.78 - d_{n,\alpha} \le S_n(w) \le 0.78 + d_{n,\alpha}), \tag{7}$$

where $S_n(w) \sim$ Binomial $(n, 0.78 - 0.181\rho)$. and ρ is the dependence parameter of the Farlie-Gumbel-Morgenstern distribution.

Proof: The g.l.b. of P in (4) is :

$$P \ge P|S_n(w) - w_\Delta(1 - \ln w_\Delta)| \ge d_{n,\alpha}), \tag{8}$$

where a point w_Δ maximizes the function $\Delta(w)$ defined as:

$$\Delta(w) = |F_\Pi(w) - w(1 - \ln w)|, \tag{9}$$

and $S_n(w) \sim$ Binomial $(n, F_\Pi(w_\Delta))$ (Stuart et al., [7]). Here $F_\Pi(w_\Delta)$ the d.f. of $\Pi = UV$, where (U, V) has the d.f. given in (6) with the standard uniform marginal distribution functions.

To find w_Δ , we first obtain $F_\Pi(w_\Delta)$. The probability density function of Π is uniquely determined by the inversion integral of the Mellin transform $M(s_1, s_2)$ of $\partial^2 C(u, v, \rho) / \partial u \partial v$. It is given by

$$M(s_1,s_2)=\int_0^1\int_0^1 u^{s_1-1}v^{s_2-1}\frac{\partial^2 C(u,v,\rho)}{\partial u\partial v}dudv$$
$$=\frac{1}{s_1s_2}+3\rho\frac{(1-s_1)(1-s_2)}{s_1s_2(s_1+1)(s_2+1)}. \qquad (10)$$

The variables s_1 and s_2 in (10) are replaced by s since U and V are dependent. The inversion integral of $M(s,s)$ is:

$$F_\Pi(w)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} w^{-s}M(s,s)ds$$
$$=-\ln w-3\rho(\ln w+4w\ln w-4w+4),$$

where $0\le w\le 1$.

Thus, the d.f. of Π is:

$$F_\Pi(w)=\int_0^w f_\Pi(t)dt=F_0(w)-g(w,\rho), \qquad (11)$$

where $F_0(w)$ is given in (2) and

$$g(w,\rho)=3\rho(2w^2\ln w+w\ln w-3w^2+3w). \qquad (12)$$

When a pair of random variables (X,Y) has the Fairle-Gumbel-Morgenstern distribution, the function $\Delta(w)$ in (9) is equal to $g(w,\rho)$ in (12), which is maximized at the point $w_\Delta=0.41553$. From (11), we get

$$F_\Pi(w_\Delta)=0.78-0.181\rho. \qquad (13)$$

Thus, (8) is equal to (7), when $w_{\Delta} = 0.41553$.

From (7), the g.l.b. of the power P of the test is tabulated for $\alpha = 0.05$, and some given n and p where $S_n(w)$ is distributed as the binomial distribution with the parameters n and p given in (13). The results are the following table.

Table 1
The g.l.b. of the power of the 0.05-size test depending on n and ρ

ρ	$n = 3$	$n = 5$	$n = 7$
-0.3	0.931	0.996	0.999
-0.2	0.917	0.995	0.999
-0.1	0.902	0.993	0.999
0.1	0.870	0.987	0.998
0.2	0.853	0.984	0.998
0.3	0.836	0.979	0.997

REFERENCES

[1] Blum, J.R., Kiefer, J. Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. Ann. Math. Stat. **32**, 485-498.
[2] De Wet, T. (1980) Cramér-von Misses test for independence. Journal of Multivariate Analysis, **10**, 38-50.
[3] Deheuvels, P. (1981) An asymptotic decomposition for multivariate distribution free test of independence. Journal of Multivariate Analysis, **11**, 102-113.
[4] Hoeffding, W. (1948). A non-parametric test of independence. Ann. Math. Stat. **19**, 546-557.
[5] Saunders, R., Laud, P. (1980) The multidimensional Kolmogorov goodness-of-fit test. Biometrika **67**, 237.
[6] Sklar, A. (1959). Fonctionas de repartition á n dimensions et leura merges. Inst. Statist.Univ. Paris Pupl. **8**, 229-231.
[7] Stuart, A., Ord, J.K., (1999) Arnold, S. Kendall's Advanced Theory of Statistics, Arnold, London.