The Annals of Mathematical Statistics
1971, Vol. 42, No. 6, 1962–1969

TESTS OF COORDINATE INDEPENDENCE FOR A BIVARIATE SAMPLE ON A TORUS

BY EDWARD D. ROTHMAN

The University of Michigan

1. Introduction and summary. This paper studies the problem of testing the independence of two random variables X, Y from a random sample, (X_1, Y_1), $(X_2, Y_2) \cdots (X_n, Y_n)$, of size n, where X and Y are angular variates (i.e., reals modulo 1). In the standard case where X and Y are ordinary real variables, the following approach has been useful. Suppose (X, Y) has the continuous distribution function $F(x, y)$ with marginal distribution functions $F_1(x)$ and $F_2(y)$, respectively. It is desired to test H_0: $F(x, y) = F_1(x)F_2(y)$ against the alternative H_A: $F(x, y) \neq F_1(x)F_2(y)$. Let $F_n(x, y)$ denote the sample distribution function of the random bivariate sample, i.e., if $H(x)$ denotes the left continuous Heaviside function then

$$F_n(x, y) = \frac{1}{n} \sum_{j=1}^{n} H(x - X_j)H(y - Y_j). \tag{1}$$

Also, let $F_{n1}(x)$ and $F_{n2}(y)$ denote the sample distribution functions associated with the first and second components of the random sample vector.

In terms of $H(x)$

$$F_{n1}(x) = \frac{1}{n} \sum_{j=1}^{n} H(x - X_j) \tag{2}$$

and

$$F_{n2}(y) = \frac{1}{n} \sum_{j=1}^{n} H(y - Y_j). \tag{3}$$

Blum, Kiefer and Rosenblatt [1] studied the following distribution free tests of independence based on the sample distribution function. Reject for large values of

$$A_n = \sup_{x,y} |T_n(x, y)| \tag{4}$$

or

$$B_n = n \int\int [T_n(x, y)]^2 \, dF_n(x, y), \tag{5}$$

where

$$T_n(x, y) = F_n(x, y) - F_{n1}(x)F_{n2}(y). \tag{6}$$

Received November 24, 1969; revised May 17, 1971.

[1] Research supported by the Office of Naval Research under Contract NONR 4010(09) awarded to the Department of Statistics, the Johns Hopkins University. This paper in whole or in part may be reproduced for any purpose of the United States Government.

The first statistic, constructed in the spirit of Kolomorov-Smirnov statistics, has good power properties, cf. [1], Section 4, but its asymptotic distribution is unknown. The statistic B_n is analogous to the Cramér-von Mises statistic and is also equivalent to a statistic originally proposed by Hoeffding [3]. The characteristic function of the null asymptotic distribution of B_n is (cf. [1] and [3])

$$E\,e^{izB} = \prod_{j,k=1}^{\infty}\left(1-\frac{2iz}{\pi^4 j^2 k^2}\right)^{-\frac{1}{2}}. \tag{7}$$

Asymptotic power properties are given in [1].

As in Rothman [9] the difficulty in modifying these tests in the toroidal case is that there is no natural origin for the distribution functions on a circle. Moreover, different arbitrary starting points give the test statistics A_n and B_n different values. In this paper we propose that the statistic C_n be used for our problem, with the surface of the torus replacing the plane.

$$C_n = n\int\int Z_n^2(x, y)\,dF_n(x, y) \tag{8}$$

where

$$Z_n(x, y) = T_n(x, y)+\int\int T_n(x, y)\,dF_n(x, y)-\int T_n(x, y)\,dF_{n1}(x)-\int T_n(x, y)\,dF_{n2}(y).$$

We note that C_n may be rewritten in the following form which we shall refer to in Section 3:

$$C_n = 1/n^2 \textstyle\sum_{j=1}^{n} \left[\sum_{a=1}^{n}\{T_n(X_j, Y_j)+T_n(X_a, Y_a)-T_n(X_a, Y_j)-T_n(X_j, Y_a)\}\right]^2. \tag{9}$$

We shall also have occasion to use the random variable,

$$D_n = n\int\int Z_n^{*2}(x, y)\,dF_1(x)\,dF_2(y),$$

where

$$Z_n^*(x, y) = T_n(x, y)+\int\int T_n(x, y)\,dF_1(x)\,dF_2(y)-\int T_n(x, y)\,dF_1(x)-\int T_n(x, y)\,dF_2(y).$$

An outline of the paper follows: In Section 2 it is shown that when H_0 is true, $C_n - D_n \to 0$ in probability. The invariance of C_n under changes of origin is proved in Section 3. Finally the asymptotic distribution of C_n under the null hypothesis is obtained in Section 4.

2. Asymptotic equivalence of tests. In this section it is shown that $C_n - D_n \to 0$ in probability when the null hypothesis of independence is true. Use will be made of the following result due to Kiefer and Wolfowitz [6]:

THEOREM 2.1. [Kiefer and Wolfowitz]. *Let*

$$K_n = \sup_{-\infty \leqq x,y<\infty} |F_n(x, y)-F(x, y)|$$

and

$$G_n(r) = \text{Prob}\,[n^{\frac{1}{2}}K_n < r].$$

For every $F(x, y)$, there exists a distribution function G such that the sequence of distribution functions G_n converges to G at every continuity point of G as $n \to \infty$.

In [1] it is stated that the asymptotic distribution of $n^{\frac{1}{2}}A_n$ exists. Using this result one derives

THEOREM 2.2. *Under the null hypothesis, H_0: $F(x, y) = F_1(x)F_2(y)$ the following random variables converge to each other in probability,*

$$C_n = n \int\int Z_n^{2}(x, y)\, dF_n(x, y)$$

$$C_n^{*} = n \int\int Z_n^{2}(x, y)\, dF_1(x)\, dF_2(x)$$

$$C_n^{**} = n \int\int [Z_n^{**}(x, y)]^2\, dF_1(x)\, dF_2(y)$$

where

$$Z_n^{**} = T_n(x, y) + \int\int T_n(x, y)\, dF_1(x)\, dF_2(y) - \int T_n(x, y)\, dF_{n1}(x) - \int T_n(x, y)\, dF_{n2}(y)$$

$$D_n = n \int\int Z_n^{*2}(x, y)\, dF_1(x)\, dF_2(y).$$

PROOF. It is first shown that $C_n - C_n^{*} \to 0$ in probability. Clearly

$$|C_n - C_n^{*}| \leqq K_1(n^{\frac{1}{2}} \sup |T_n(x, y)|)^2 \int\int d|F_n(x, y) - F(x, y)|,$$

where K_1 is a constant. Since $F_n(x, y)$ converges to $F(x, y)$ uniformly with probability 1 and Prob $(n^{\frac{1}{2}} \sup |T_n(x, y)| < r)$ converges to a distribution function, then as $n \to \infty$, $C_n - C_n^{*} \to 0$ in probability. Again

$$|C_n^{*} - C_n^{**}| \leqq K_2(n^{\frac{1}{2}} \sup T_n(x, y))^2 \int\int [\int\int d|F_n(x, y) - F(x, y)|]\, dF_1(x)\, dF_2(y),$$

and the above reasoning shows that $C_n^{*} \to C_n^{**}$ as $n \to \infty$. Similarly $C_n^{**} - D_n \to 0$ in probability.

3. Invariance of C_n. The result of this section is contained in the following:

LEMMA 3.1. *C_n is invariant with respect to choice of origin on the torus.*

PROOF. The term { } in equation (9) may be shown to be

$$(10)\quad T_n(X_i, Y_i) - T_n(X_i, Y_j) - T_n(X_j, Y_i) + T_n(X_j, Y_j) = \pm((n_{ij,ij}/n) - (n_{\cdot ij}/n)(n_{ij\cdot}/n))$$

where $n_{ij,ij}$ is the number of observations in the rectangle with corners (X_i, Y_i), (X_i, Y_j), (X_j, Y_i), (X_j, Y_j), including the "northeast" corner while

$$n_{ij\cdot} = n|F_{1n}(X_i) - F_{1n}(X_j)| \quad \text{and} \quad n_{\cdot ij} = n|F_{2n}(Y_i) - F_{2n}(Y_j)|.$$

Let X_k, $k = 1, 2, \cdots, n$ be replaced by $X_k^{1} = X_k + c$ (modulo 1) where c is some constant. It is sufficient to consider the effect of this transformation on the term in { } when the "bottom" X "rolls" off the bottom to the top. This will be accomplished by examining each of 3 nontrivial cases in turn. First assume that $X_i < X_j$ and $Y_i < Y_j$ then if $X_i^1 > X_j^1$ we have

$$(11) \quad F_n(X_i^1, Y_i) - F_n(X_i^1, Y_j) - F_n(X_j^1, Y_i) + F_n(X_j^1, Y_j) = (n_{\cdot ij} - n_{ij,ij})/n$$

and

$$(12) \quad -F_{n1}(X_i^1)F_{n2}(Y_j) + F_{n1}(X_i^1)F_{n2}(Y_i) + F_{n1}(X_j^1)F_{n2}(Y_i) - F_{n1}(X_j^1)F_{n2}(Y_j)$$
$$= [1-(n_{ij\cdot}/n)][n_{\cdot ij}/n].$$

Subtracting (12) from (11) shows that only the sign of the term in { } is altered. Thus the result is verified in this case. Suppose now that $X_i < X_j$ and $Y_i > Y_j$ and $X_i^1 > X_j^1$. Then the sign of the right member of (10) can be shown to change from a minus to a plus, and hence the result is proved in this case too. Similarly, we can treat the cases $X_i > X_j$, $Y_i > Y_i$ or $Y_j < Y_j$ with $X_j^1 < X_i^1$. By symmetry different choices of the Y coordinate of the origin will leave C_n unchanged.

4. The asymptotic distribution of C_n and D_n. It is most convenient to find the asymptotic distribution of $E_n = nD_n/(n-1)$. Clearly E_n and D_n will have the same asymptotic distribution. Since $F(x, y)$ is continuous, so are the marginal distribution functions. Hence, we may use the probability integral transformation as in [1] to obtain the following

$$(13) \quad E_n = \frac{n^2}{n-1}\int_0^1\int_0^1\Big[\tilde{T}_n(x, y) + \int_0^1\int_0^1 \tilde{T}_n(x, y)\,dx\,dy - \int_0^1 \tilde{T}_n(x, y)\,dy$$
$$- \int_0^1 \tilde{T}_n(x, y)\,dy\Big]^2 dx\,dy$$

where

$$(14) \qquad \tilde{T}_n(x, y) = \tilde{F}_n(x, y) - \tilde{F}_{n1}(x)\tilde{F}_{n2}(y)$$

$\tilde{F}_n(x, y)$ is the empirical cdf of $(F_1(X_1), F_2(Y_1)), \cdots, (F_1(X_n), F_2(Y_n))$ and $\tilde{F}_{n1}$, F_{n2} are the corresponding marginals. Therefore under H_0 we may assume that $X_1, X_2, \cdots, X_n$ and $Y_1, Y_2, \cdots, Y_n$ are independent $U(0, 1)$ rv's. Let Z_{km} denote the double Fourier coefficient of the term in braces in (13).

$$(15) \qquad Z_{km} = \textstyle\int_0^1\int_0^1 Z_n^*(x, y)\exp(-2\pi ikx)\exp(-2\pi imy)\,dx\,dy.$$

It is easily seen that

$$(16) \qquad \begin{aligned} Z_{00} &= 0, \\ Z_{k0} &= 0, \qquad \text{all } k, \\ Z_{0m} &= 0, \qquad \text{all } m. \end{aligned}$$

For $k \neq 0$, $m \neq 0$

$$\int_0^1\int_0^1 F_n(x, y)\exp(-2\pi ikx)\exp(-2\pi imy)\,dx\,dy$$

$$= \frac{1}{n}\sum_{j=1}^{n}\int_0^1 H(x-X_j)\exp(-2\pi ikx)\,dx\int_0^1 H(y-Y_j)\exp(-2\pi imy)\,dy \tag{17}$$

$$= \frac{1}{4\pi^2 km}\Big[-\frac{1}{n}\sum_{j=1}^{n}\exp(-2\pi ikX_j)\exp(-2\pi imY_j) \tag{18}$$
$$+\frac{1}{n}\sum_{j=1}^{n}[\exp(-2\pi ikX_j)+\exp(-2\pi imY_j)]-1\Big].$$

Also

$$\int_0^1 \tilde{F}_{n1}(x)\exp(-2\pi ikx)\,dx = \frac{1}{n}\sum_{j=1}^{n}\frac{\exp(-2\pi ikX_j)-1}{2\pi ik} \tag{19}$$

$$\int_0^1 \tilde{F}_{n2}(y)\exp(-2\pi imy)\,dy = \frac{1}{n}\sum_{j=1}^{n}\frac{\exp(-2\pi imY_j)-1}{2\pi im}, \tag{20}$$

which imply that

$$\int_0^1\int_0^1 \tilde{F}_{n1}(x)\tilde{F}_{n2}(y)\exp(-2\pi ikx)\exp(-2\pi imy)\,dx\,dy$$

$$= \frac{1}{4\pi^2 km}\Big[-\Big(\frac{1}{n}\sum_{j=1}^{n}\exp(-2\pi ikX_j)\Big)\Big(\frac{1}{n}\sum_{j=1}^{n}\exp(-2\pi imY_j)\Big) \tag{21}$$
$$+\frac{1}{n}\sum_{j=1}^{n}(\exp(-2\pi ikX_j)+\exp(-2\pi imY_j))-1\Big].$$

Combining (14), (19), (20) and (21) one obtains,

$$\int_0^1\int_0^1 \tilde{T}_n(x, y)\exp(-2\pi ikx)\exp(-2\pi imy)\,dx\,dy$$

$$= \frac{1}{4\pi^2 mk}\Big[\frac{1}{n^2}\Big(\sum_{j=1}^{n}\exp(-2\pi ikX_j)\Big)\Big(\sum_{j=1}^{n}\exp(-2\pi imY_j)\Big) \tag{22}$$
$$-\frac{1}{n}\sum_{j=1}^{n}\exp(-2\pi i(kX_j+mY_j))\Big].$$

Hence, if $k \neq 0$, $m \neq 0$ it follows that

$$Z_{km} = \frac{1}{4\pi^2 mk}\Big[\frac{1}{n}\Big(\sum_{j=1}^{n}\exp(-2\pi ikX_j)\Big)\Big(\frac{1}{n}\sum_{j=1}^{n}\exp(-2\pi imY_j)\Big) \tag{23}$$
$$-\frac{1}{n}\sum_{j=1}^{n}\exp(-2\pi i(kX_j+mY_j))\Big].$$

Applying Parseval's theorem to (13) and (15) yields

$$E_n = \frac{n^2}{n-1} \sum_{|k|=1}^{\infty} \sum_{|m|=1}^{\infty} |Z_{km}|^2. \tag{24}$$

This representation has an immediate consequence. Suppose

$$\begin{aligned} X_i &= X_i + c_1 \qquad \text{(modulo 1)} \\ Y_i &= Y_i + c_2 \qquad \text{(modulo 1)} \qquad i = 1, 2, \cdots, n \end{aligned} \tag{25}$$

where c_1 and c_2 are arbitrary constants. Under this transformation $|Z_{km}|$ becomes

$$|\exp(-2\pi i k c_1)\exp(-2\pi i m c_2)|\,|Z_{km}| = |Z_{km}|,$$

whenever k and m are both $\neq 0$, therefore D_n is invariant with respect to the choice of the origin. Under the null hypothesis of independence

$$EZ_{km} = 0 \quad \text{all} \quad k, m, \tag{26}$$

$$\begin{aligned} EZ_{km}Z_{k'm'} &= \frac{\delta_{kk'}\delta_{mm'}}{16\pi^2k^2m^2}\left(\frac{n-1}{n^2}\right) \qquad &&\text{if} \quad k \neq 0, m \neq 0; \\ &= 0 &&\text{otherwise,} \end{aligned} \tag{27}$$

where δ_{jk} is the Kronecker delta.

Let $C(x, y; u, v)$ denote the covariance kernel of the random process

$$n^2/(n-1)\, Z_n^*(x, y), \text{ then}$$

$$C(x, y; u, v) = \frac{n^2}{n-1}\{[EZ_n^*(x, y)Z_n^*(u, v)] - [EZ_n^*(x, y)][EZ_n^*(u, v)]\}. \tag{28}$$

Letting $\{C_{kmrs}\}$ be the Fourier coefficients of $C(x, y; u, v)$ relative to the basis

$$\{\exp(2\pi ikx)\exp(2\pi imy)\exp(-2\pi iru)\exp(-2\pi isv);\ -\infty < k, m, r, s < \infty\},$$

$$C_{kmrs} = \frac{n^2}{n-1}E[Z_{km}Z_{rs}] - \frac{n^2}{n-1}[EZ_{km}][EZ_{rs}] \tag{29}$$

$$\begin{aligned} &= \frac{1}{16\pi^4m^2k^2} \qquad &&\text{if } k = r, m = s, \text{ and } k, m \neq 0 \\ &= 0 &&\text{otherwise.} \end{aligned} \tag{30}$$

The Fourier series for $C(x, y; u, v)$, which converges in mean square, is

$$C(x, y; u, v) = \textstyle\sum_k \sum_m \sum_r \sum_s C_{kmrs} \exp(2\pi ikx)\exp(2\pi imy)\exp(-2\pi iru)\exp(-2\pi isv),$$

which in view of (27), reduces to

$$= \sum_{|k|=1} \sum_{|m|=1} \frac{1}{16\pi^4 m^2 k^2} \exp(2\pi i k(x-u)) \exp(2\pi i m(y-v)). \tag{31}$$

It may be shown that

$$C(x, y; u, v) = [\tfrac{1}{2}(x-u)^2 - \tfrac{1}{2}|x-u| + \tfrac{1}{12}][\tfrac{1}{2}(y-v)^2 - \tfrac{1}{2}|y-v| + \tfrac{1}{12}]$$

which may be rewritten in the form of a product of two covariance kernels obtained by Watson (1961) for U_n^2 namely,

$$C(x, y; u, v) = (\min(x, u) - \tfrac{1}{2}(x+u) + \tfrac{1}{2}(x-u)^2 + \tfrac{1}{12}) \cdot (\min(y, v) - \tfrac{1}{2}(y+v) + \tfrac{1}{2}(y-v)^2 + \tfrac{1}{12}).$$

Consider the integral equation

$$\int_0^1 \int_0^1 C(x, y; u, v)\phi(u, v)\, du\, dv = \lambda\phi(x, y). \tag{32}$$

From the series representation of $C(x, y; u, v)$ and (32), the four complex eigenfunctions $\exp(-2\pi i k u)$, $\exp(-2\pi i m v)$, $\exp(2\pi i m v)$, $\exp(2\pi i k v)\exp(-2\pi i m v)$ and $\exp(2\pi i k u)\exp(2\pi i m v)$ correspond to the eigenvalue $\frac{1}{16}\pi^4 k^2 m^2 (k, m > 0)$. Since any linear combination of these is also an eigenfunction with the same eigenvalue, one can take $2\sin(2\pi k u)\sin(2\pi m v)$, $2\sin(2\pi k u)\cos(2\pi m v)$, $2\cos(2\pi k u)\sin(2\pi m v)$, $2\cos(2\pi k u)\cos(2\pi m v)$ as a basis for the eigenmanifold. Having obtained the eigenvalues and eigenfunctions, the usual argument, cf. [1], gives the asymptotic characteristic function,

$$\phi(t) = \prod_{k=1}^{\infty} \prod_{m=1}^{\infty} \left[\left(1 - \frac{2it}{16\pi^4 k^2 m^2}\right)^{-\frac{1}{2}}\right]^4 = \prod_{k,m=1}^{\infty} \left(1 - \frac{it}{8\pi^4 k^2 m^2}\right)^{-2}. \tag{33}$$

Applying a result of Zolatorev's [11], one may approximate the upper tail of the distribution function, $T(x)$, of E_n as follows:

$$\lim_{x\to\infty} \frac{1 - T(x)}{\operatorname{Prob}[\chi^2(4) > 16\pi^4 x]} = \prod_{m,k=1;(m,k)\neq(1,1)}^{\infty} \left(1 - \frac{1}{k^2 m^2}\right)^{-2}. \tag{34}$$

Acknowledgment. The author is grateful to Professor S. Schach for his assistance in the preparation of this manuscript and to the referee for many helpful comments.

REFERENCES

[1] Blum, J. R., Kiefer, J. and Rosenblatt, M. (1961). Distribution free test of independence based on the sample distribution function. *Ann. Math. Statist.* **32** 485–492.

[2] Gould, S. H (1966). *Variational Methods for Eigenvalue Problems*. University of Toronto Press.

[3] Hoeffding, W. (1948). A nonparametric test of independence. *Ann. Math. Statist.* **19** 546–557.

[4] KAC, M. (1951). On some connections between probability theory and differential and integral equations. *Proc. Second Berkeley Symp. Math. Statist. Prob.* University of California Press.

[5] KAC, M. and SIEGERT, A. J. F. (1947). An explicit representation of a stationary Gaussian process. *Ann. Math. Statist.* **18** 438–442.

[6] KIEFER, J. and WOLFOWITZ, J. (1958). On the deviations of the empirical distribution function of vector chance variables. *Trans. Amer. Math. Soc.* **87** 173–186.

[7] KUIPER, N. H. (1960). Tests concerning random points on a circle. *Nederl. Akad. Wetensch. Proc. Ser. A* 383–397.

[8] ROSENBLATT, M. (1952). Limit theorems associated with variants of the von Mises Statistics. *Ann. Math. Statist.* **23** 617–623.

[9] ROTHMAN, E. D. (1969). Tests for uniformity of a circular distribution. Technical Report No. 128, Department of Statistics, Johns Hopkins Univ.

[10] WATSON, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika* **48** 109–114.

[11] ZOLOTAREV, V. M. (1961). Concerning a certain probability problem. *Theor. Probability Appl.* 201–204.