A DISTRIBUTION-FREE TEST FOR BIVARIATE SYMMETRY

by

Myles Hollander

FSU Statistics Report M174-R
ONR Technical Report No. 39

November, 1969
Department of Statistics
Florida State University
Tallahassee, Florida

Research supported by the Office of Naval Research, Contract NONR 988(08), Task Order NR 042-004
A DISTRIBUTION-FREE TEST FOR BIVARIATE SYMMETRY

by

Myles Hollander

The Florida State University

1. Introduction. We observe \( Z = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \) where \( (X_1, Y_1), \ldots, (X_n, Y_n) \) are assumed mutually independent, each distributed according to the distribution \( F \). In this note we introduce a test of

\[
H_0 : F(x,y) = F(y,x), \quad \text{for all } (x,y),
\]

based on the sample distribution function. The test is shown to be consistent against all \( F \) for which \( H_0 \) does not hold.

2. The test procedure. The test is based on the statistic

\[
D_n = \int \int [F_n(x,y) - F_n(y,x)]^2 \, dF_n(x,y),
\]

where \( F_n(x,y) = (n)^{-1} \sum_{j=1}^{n} \phi(X_j, x) \phi(Y_j, y) \) and \( \phi(a,b) = 1 \) if \( a \leq b, 0 \) otherwise. The introduction of \( D_n \) is motivated by the \( B_n \) statistic used for testing independence by Blum, Kiefer, and Rosenblatt (1961); also see Hoeffding (1948).

The statistic \( D_n \) is not distribution-free under \( H_0 \). This is easily seen by taking \( n = 2 \) and \( F \) absolutely continuous. In that case, after some algebra, we find

\[
P_F \left( 2D_2 \leq 4 \right) = P_F(X_1 < X_2 < Y_1 < Y_2) + P_F(X_2 < X_1 < Y_2 < Y_1) + P_F(Y_1 < Y_2 < X_1 < X_2) + P_F(Y_2 < Y_1 < X_2 < X_1).
\]

The four probabilities on the right hand side of (3) are equal whenever \( H_0 \) holds. In Bell and Haller (1969), where parametric and nonparametric

Research supported by the Office of Naval Research, Contract NONR 988(08), Task Order NR 042-004
tests of $H_0$ are discussed, it is pointed out that such pattern
probabilities are not distribution-free in $H_0$. For instance, take
$F(x, y) = R(x) \cdot R(y)$ (the case where $X, Y$ are independent and each has
distribution $R$) to find $P(2D_2 = 5/4) = 1/6$. When $F(x, y)$ is the standardized
bivariate normal with correlation coefficient $\rho$, use of F. N. David's
formula (see equation $(42)$ of Gupta (1963)) yields $P(2D_2 = 5/4) = (2)^{-1} +
(\pi)^{-1} \{\arcsin \rho + 2 \arcsin (-(1+\rho)/2)\}$.

It is also true that $D_n$ is not asymptotically distribution-free
under $H_0$. Consider the distributions $F_1(x, y) = xy$ and
$F_2(x, y) = (2)^{-1} (xy + x^2 y^2)$ on the unit square. Arguments analogous to
those in Blum, Kiefer, and Rosenblatt (1961) and Kiefer (1959) show
the asymptotic null mean of $nD_n$ is equal to

(4) \[ \mu_F = \int \int \mathbb{E}[T^*(x, y)]^2 \, dP(x, y), \]

where $T^*(x, y)$ is a Gaussian process with

(5) \[ \mathbb{E}(T^*(x, y)) = 0, \]

(6) \[ \mathbb{E}(T^*(x, y) \cdot T^*(u, v)) = 2\{F(\min(x, u), \min(y, v)) - F(\min(x, v), \min(u, y))\}. \]

A direct calculation shows $\mu_{F_1} = 1/6$ and $\mu_{F_2} = 29/180$. This enables us
to conclude that $nD_n$ is not asymptotically distribution-free under $H_0$.

Since $nD_n$ lacks the (unconditionally) distribution-free property,
we turn to a conditional test. Consider the group $G$ of $2^n$ transformations
where

$$g(j_1, \ldots, j_n)(z) = \{(X_1, Y_1)^{(j_1)}, \ldots, (X_n, Y_n)^{(j_n)}\},$$

each $j_i$ is either "0" or "1", and $(X_1, Y_1)^{(0)} = (X_1, Y_1), (X_1, Y_1)^{(1)} = (Y_1, X_1).$
The hypothesis \( H_0 \) implies that for each \( g \in G \), \( g(Z) \) has the same distribution as \( Z \). We thus consider the conditional measures

\[
P_c\{ (X_1, Y_1) = (x_1, y_1), \ldots, (X_n, Y_n) \} = 2^{-n}, \quad \text{for each } (j_1, \ldots, j_n).
\]

Our \( \alpha \)-level test is as follows. Let \( K^{(1)}(z) \leq K^{(2)}(z) \leq \ldots \leq K^{(2^n)}(z) \)
denote the \( 2^n \) ordered values of \( nD_n(gz) \) for \( g \in G \). If \( \phi_n(z) \) denotes the probability of rejecting \( H_0 \) when \( Z = z \), we set

\[
\phi_n(z) = \begin{cases} 
1 & \text{if } nD_n(z) > K^{(m)}(z) \\
\ell_n(z) & \text{if } nD_n(z) = K^{(m)}(z) \\
0 & \text{if } nD_n(z) < K^{(m)}(z),
\end{cases}
\]

where \( m = 2^n - [2^n \alpha] \), \([2^n \alpha] \) is the greatest integer less than or equal to \( 2^n \alpha \), and \( \ell_n(z) \) is determined to give the test size \( \alpha \). The test defined by (8) is conditionally distribution-free.

3. Consistency. Let \( \gamma_F = \iint (F(x, y) - F(y, x))^2 dF(x, y) \).

**Lemma 1:** \( \gamma_F = 0 \) if and only if \( F(x, y) = F(y, x) \)

**Proof:** The "if" is obvious. To prove "only if" let \( A = \{ (x, y) : F(x, y) = F(y, x) \} \).

Then \( \gamma_F = 0 \) implies \( P_F(A) = 1 \). Since \( P_F(A) = 1 \), we have for all \( (x_o, y_o) \),

\[
F(x_o, y_o) = \iint_{A^*(x_o, y_o)} dF(x, y),
\]

where \( A^*(x_o, y_o) = \{ (x, y) : x \leq x_o; y \leq y_o; (x, y) \in A \} \). That \( F(x_o, y_o) = F(y_o, x_o) \)
now follows by computing \( F(x_o, y_o) \) and \( F(y_o, x_o) \) via formula (9), using the fact that, on \( A \), \( F(x, y) = F(y, x) \).

**Theorem:** The sequence of tests defined by (8) is consistent when \( H_0 \) does not hold.

**Proof:** Let \( a_j = \min (X_j, Y_j), \ b_j = \max (X_j, Y_j) \) and rewrite \( D_n \) as
(10) \[ D_n = \sum_{j=1}^{n} \frac{(F_n(a_j, b_j) - F_n(b_j, a_j))^2}{n}, \]

where, without loss of generality, we take \( a_1 \leq a_2 \leq \ldots \leq a_n \). Now,

\[ n(F_n(a_j, b_j) - F_n(b_j, a_j)) \]

(11) \[ = \sum_{i=1}^{n} \{ I(X_i < a_j; a_j < Y_i < b_j) - I(a_j < X_i < b_j; Y_i < a_j) \} \]

where \( I(A) \) denotes the indicator of the event \( A \). Let \( L = \{ i : a_i \neq b_i \} \) and for \( i \in L \), set \( X_i = r_i a_i + (1-r_i)b_i, Y_i = (1-r_i)a_i + r_i b_i \). Thus for \( i \in L \),

\[ r_i = 0 \text{ if } X_i = b_i \text{ and } 1 \text{ if } X_i = a_i. \]

Under \( P_c \), the \( r_i \), \( i \in L \), are independent and identically distributed (iid), \( P_c \{ r_i = 1 \} = P_c \{ r_i = 0 \} = 1/2 \).

For \( i \notin L \), we may define random variables \( r_i \) with this distribution such that \( r_1, \ldots, r_n \) are iid under \( P_c \). Then, from (11) we have.

(12) \[ n(F_n(a_j, b_j) - F_n(b_j, a_j)) = \sum_{i=1}^{n} \{ I(a_j < b_i \leq b_j; a_i \neq b_i) r_i \}
\]

\[ - I(a_j < b_i \leq b_j; a_i \neq b_i)(1-r_i) \].

Setting \( d_{ij} = I(a_j < b_i \leq b_j; a_i \neq b_i) \) and \( s_i = 2r_i - 1 \), we have from (10), (12),

(13) \[ nD_n = n^{-2} \sum_{j=1}^{n} T_j^2, \]

where

(14) \[ T_j = \sum_{i=1}^{n} d_{ij}s_i. \]

Then,

(15) \[ E_c(T_j^2) = E_c \left( \sum_{i=1}^{n} d_{ij} + \sum_{i \neq j}^{n} d_{ij}d_{i\neq j}s_is_{i\neq j} \right) = \sum_{i=1}^{n} d_{ij}, \]

since, with respect to \( P_c \), the \( d \)'s are constant and the \( s \)'s are independent with \( E_c(s_i) = 0 \). Thus

(16) \[ E_c(nD_n) = n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} d_{ij} \text{ def } c_n. \]
Note \( q_n \leq 1 \), and since \( d_{i1} = 1 \) if \( i \in L \), we see that \( q_n > 0 \) provided \( z \) is such that \( x_i \neq y_i \) for at least one \( i \). Letting \( W = \{ z : x_i = y_i, \text{ all } i \} \) we have (cf. Theorem (2.1) of Hoeffding (1952))

\[
K^m(z) < q_n/a \leq 1/a, \ z \in W
\]

(17)

\[
K^m(z) = 0, \quad z \in W.
\]

Also,

(18) \( \lim \text{p-}D_n = Y_F \).

Equation (18), where the probability limit is unconditional (i.e., for \( P_F \)) follows from \( \lim E(D_n) = Y_F, \lim \text{Var}(D_n) = 0 \), and Chebychev's inequality. Now, when \( H_o \) is false, (18) and \( Y_F > 0 \) (Lemma 1) imply

\[
\lim \text{P}_F(nD_n) = 0 \quad \text{for all } t.
\]

This, along with (17), yields consistency.

4. An unsolved problem. Equations (13), (14) provide a convenient representation of \( D_n \) for calculating the values of \( nD_n(gZ) \). These values depend on \( Z \) only through the ordering pattern of the 2n \( X \)'s and \( Y \)'s. For example, for any observed \( Z \) (\( n=3 \)) yielding the ordering pattern \( (X_1 < Y_3 < Y_1 < X_2 < X_3 < Y_2) \) we obtain the \( P_c \) distribution of \( nD_n \) as follows. We have \( a_1 = X_1, a_2 = Y_3, a_3 = X_2, b_1 = Y_1, b_2 = X_3, b_3 = Y_2 \), \( d_{11} = 1 \), \( d_{12} = 1, d_{13} = 0, d_{22} = 1, d_{23} = 1, d_{33} = 1. \) (Whenever \( a_1 < a_2 < \ldots < a_n \), we have \( d_{ij} = 0 \) for \( i > j \).) From (13), (14), we find \( nD_n = 1/9 \) for our observed ordering pattern, and also \( K^{(1)} = K^{(2)} = 1/9, K^{(3)} = K^{(4)} = K^{(5)} = K^{(6)} = 5/9, K^{(7)} = K^{(8)} = 1, \) so there is no reason to reject \( H_o \).

For large \( n \), it is possible, though extremely tedious, to table the distribution. It is thus useful, and of statistical interest, to obtain the asymptotic \( P_c \) distribution of \( nD_n \). We have not been able to do this.
Using (13), and proceeding formally, we write

\[
D_n = n^{-2} \sum_{i=1}^{n} \sum_{i'=1}^{n} \left( \sum_{j=1}^{n} d_{ij} d_{i'j} \right) s_is_{i'} = n^{-2} \sum_{j=1}^{n} \lambda_j^{(n)} \left( \sum_{i=1}^{n} e_je_{i} \right)^2,
\]

where \( c_{ii'} = \sum_{j=1}^{n} d_{ij} d_{i'j}, \ C_n = (c_{ii'}), \) and \( \lambda_j^{(n)}, e_j^{(n)} = (e_{j1}, \ldots, e_{jn}) \)

are the characteristic roots and vectors corresponding to \( C_n \). However, in order to remove the difficulty of computing \( \lambda \)'s and \( e \)'s for many \( C_n \) matrices, corresponding to different ordering patterns, another normalization seems necessary. Representation (19), and related work by Blum, Kiefer, and Rosenblatt (1961) and Hoeffding (1948) suggests the asymptotic distribution will be that of a weighted sum of independent chi-square-1 variables.

In order to increase the utility of the conditional test, a Fortran program for obtaining the significance level achieved has been written by Douglas Whitten. For \( n \leq 15 \), complete enumeration is used. For \( n > 15 \), we use a less time consuming modified procedure (as suggested by Dwass (1957)) based on a random sample of the \( 2^n \) permutations. The program deck and instructions can be obtained by writing to the Department of Statistics at Florida State.

5. Acknowledgments. I am extremely grateful to Jayaram Sethuraman for numerous valuable discussions. Thanks are also due to Jaroslav Hájek and Richard Savage for helpful comments, to Douglas Whitten and Duane Meeter for computing support.
REFERENCES


Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be independent and identically distributed according to the bivariate continuous distribution \(F\). For testing \(H_0: F(x, y) = F(y, x)\), all \((x, y)\), we introduce a conditionally distribution-free test based on the sample distribution function.