A NON-PARAMETRIC TEST OF INDEPENDENCE[1]

By Wassily Hoeffding

Institute of Statistics, University of North Carolina

1. Summary. A test is proposed for the independence of two random variables with continuous distribution function (d.f.). The test is consistent with respect to the class Ω'' of d.f.'s with continuous joint and marginal probability densities (p.d.). The test statistic D depends only on the rank order of the observations. The mean and variance of D are given and $\sqrt{n}(D - ED)$ is shown to have a normal limiting distribution for any parent distribution. In the case of independence this limiting distribution is degenerate, and nD has a non-normal limiting distribution whose characteristic function and cumulants are given. The exact distribution of D in the case of independence for samples of size $n = 5, 6, 7$ is tabulated. In the Appendix it is shown that there do not exist tests of independence based on ranks which are unbiased on any significance level with respect to the class Ω''. It is also shown that if the parent distribution belongs to Ω'' and for some $n \geq 5$ the probabilities of the $n!$ rank permutations are equal, the random variables are independent.

2. Introduction. In a non-parametric test of a statistical hypothesis we do not make any assumptions about the functional form of the population distribution. A general theory of non-parametric tests is not yet developed, and a satisfactory definition of "best" non-parametric tests does not seem to be available. Desirable properties of a "good" non-parametric test are unbiasedness and consistency. A test of a hypothesis H_0 is said to be consistent with respect to a specified class of admissible hypotheses if the probability of accepting H_0 tends to zero with increasing sample size whenever a hypothesis $\neq H_0$ of this class is true.

In this paper we consider the problem of testing the independence of two random variables X, Y on the basis of a random sample of size n. In all that follows the d.f. $F(x, y)$ of (X, Y) is assumed to be continuous. We will denote by Ω' the class of continuous d.f.'s $F(x, y)$ and by Ω'' the class of d.f.'s having continuous joint and marginal p.d.'s,

$$f(x, y) = \partial^2 F(x, y)/\partial x\, \partial y, \quad f_1(x) = \int f(x, y)\, dy, \quad f_2(y) = \int f(x, y)\, dx.$$

The hypothesis H_0 to be tested is that $F(x, y)$ is of the form

$$F(x, y) = F(x, \infty)F(\infty, y).$$

Several tests of this hypothesis have been proposed. Among them those deserve particular attention which depend only on the rank order of the obser-

[1] Research under a contract with the Office of Naval Research for development of multivariate statistical theory.

vations. They will be referred to as rank tests. The critical region of a rank test of independence with respect to the class Ω' is similar to the sample space; the rank tests share this property with other tests obtained by the method of randomization (cf. Scheffé [1]). A characteristic feature of a rank test is that it remains invariant under order preserving transformations of X or Y.

Rank tests of independence have been studied by Hotelling and Pabst [2], Kendall [3] and Wolfowitz [4]. While nothing is yet known about the power of the last test, the author [5] has shown that the two former tests are asymptotically biased for certain alternatives belonging to Ω'. By a slight modification of the examples given in [5] it can be shown that these tests are asymptotically biased even with respect to the class Ω''.

In the Appendix it is shown that there do not exist rank tests of independence which are unbiased on any level of significance with respect to the classes Ω' or Ω''. It will appear from this paper that there do exist rank tests of independence which are consistent, and hence asymptotically unbiased, at least with respect to Ω''.

3. The Functional $\Delta(F)$. Given a random sample from a population with a d.f. belonging to a class Ω, we want to test the hypothesis H_0 that F is in a subclass ω of Ω. It is easy to construct a consistent test of H_0 if there exist (a) a functional $\theta(F)$ defined for every F in Ω and such that $\theta(F) = 0$ if and only if $F \in \omega$; and (b) a consistent estimate of $\theta(F)$. There are many ways of devising by this method consistent tests of independence. The particular test described in the sequel has been chosen mainly for its relative simplicity.

If $F(x, y)$ is a bivariate d.f., let

$$D(x, y) = F(x, y) - F(x, \infty)F(\infty, y)$$

and

$$\Delta = \Delta(F) = \int D^2(x, y)\, dF(x, y). \tag{3.1}$$

Here and in the following, when no domain of integration is indicated, the (Lebesgue-Stieltjes) integral is extended over the entire space (here R_2).

The random variables X, Y with the d.f. $F(x, y)$ are independent if and only if $D(x, y) \equiv 0$.

THEOREM 3.1. *When $F(x, y)$ belongs to Ω'', $\Delta(F) = 0$ if and only if $D(x, y) \equiv 0$.*

PROOF. Evidently $D(x, y) \equiv 0$ implies $\Delta(F) = 0$.

Now suppose that $D(x, y) \not\equiv 0$. Since $F(x, y)$ is in Ω'', the function $d(x, y) = f(x, y) - f_1(x)f_2(y)$ is continuous. We have

$$D(x, y) = \int_{-\infty}^{x}\int_{-\infty}^{y} d(u, v)\, du\, dv.$$

$D(x, y) \not\equiv 0$ implies $d(x, y) \not\equiv 0$, and since

$$\iint d(x, y)\, dx\, dy = 0,$$

there exists a rectangle Q in R_2 such that $d(x, y) > 0$ if (x, y) is in Q. Hence $D(x, y) \neq 0$ almost everywhere in Q, and $f(x, y) > 0$ in Q. Thus

$$\Delta(F) \geq \iint_Q D^2(x, y)\, f(x, y)\, dx\, dy > 0.$$

This completes the proof.

If $F(x, y)$ is discontinuous, we can have $\Delta(F) = 0$ and $D(x, y) \not\equiv 0$. This is, for instance, the case for the distribution

$$P\{X = 0, Y = 1\} = P\{X = 1, Y = 0\} = \tfrac{1}{2}.$$

The question remains open whether $\Delta = 0$ implies $D(x, y) \equiv 0$ if $F(x, y)$ is continuous or absolutely continuous.

In Section 7 it will be shown that

$$0 \leq \Delta \leq \tfrac{1}{30}$$

The upper bound $\frac{1}{30}$ is attained when $F(x, y)$ is the (continuous) d.f. of a random variable (X, Y) such that X has any continuous d.f. and $Y = X$ (or, more generally, Y is a monotone function of X).

Let

$$C(u) = \begin{cases} 1 & \text{if } u \geqq 0, \\ 0 & \text{if } u < 0, \end{cases}$$

$$\psi(x_1, x_2, x_3) = C(x_1 - x_2) - C(x_1 - x_3), \tag{3.2}$$

$$\phi(x_1, y_1; \cdots; x_5, y_5) = \tfrac{1}{4}\psi(x_1, x_2, x_3)\psi(x_1, x_4, x_5)\psi(y_1, y_2, y_3)\psi(y_1, y_4, y_5).$$

Then we can write

$$\Delta = \int \cdots \int \phi(x_1, y_1; \cdots; x_5, y_5)\, dF(x_1, y_1) \cdots dF(x_5, y_5). \tag{3.3}$$

4. The Statistic D. Let $(X_1, Y_1), \cdots, (X_n, Y_n)$ be a random sample from a population with the d.f. $F(x, y)$, $n \geq 5$, and let

$$D = D_n = \frac{1}{n(n-1)\cdots(n-4)} \Sigma'' \phi(X_{\alpha_1}, Y_{\alpha_1}; \cdots; X_{\alpha_5}, Y_{\alpha_5}), \tag{4.1}$$

where Σ'' denotes summation over all α such that

$$\alpha_i = 1, \cdots, n; \qquad \alpha_i \neq \alpha_j \text{ if } i \neq j, \quad (i, j = 1, \cdots, 5).$$

Since the number of terms in Σ'' is $n(n-1)\cdots(n-4)$, we have by (3.3),

$$ED = \Delta. \tag{4.2}$$

Since in the case of independence $ED = 0$, D can assume both positive and negative values. It will be seen in Section 7 that $-\frac{1}{60} \leq D_n \leq \frac{1}{30}$, the upper bound $\frac{1}{30}$ being attained for every n, while the minimum of D_n apparently increases with n.

The random variable D as defined by (4.1) belongs to the class of U-statistics considered by the author [5]. The following properties of D follow immediately from the results of that paper:

I. *Let*

$$\Phi(x_1, y_1; \cdots; x_5, y_5) = D_5 = \frac{1}{5!}\Sigma'' \phi(x_{\alpha_1}, y_{\alpha_1}; \cdots; x_{\alpha_5}, y_{\alpha_5}),$$

$$\Phi_k(x_1, y_1; \cdots; x_k, y_k) = \int \cdots \int \Phi(x_1, y_1; \cdots; x_k, y_k; x_{k+1}, y_{k+1}; \cdots x_5, y_5)$$
$$dF(x_{k+1}, y_{k+1}) \cdots dF(x_5, y_5), \ (k = 1, \cdots, 5),$$

$$\zeta_k = \int \cdots \int \{\Phi_k(x_1, y_1; \cdots; x_k, y_k) - \Delta\}^2 dF(x_1, y_1) \cdots dF(x_k, y_k).$$

Then the variance of D_n *is*

$$\text{var } D_n = \binom{n}{5}^{-1} \sum_{k=1}^{5} \binom{5}{k}\binom{n-5}{5-k}\zeta_k. \tag{4.3}$$

We have

$$25\,\zeta_1 \leq n \text{ var } D_n \leq 5\,\zeta_5.$$

n var D_n is a decreasing function of n, and

$$\lim_{n\to\infty} n \text{ var } D_n = 25\,\zeta_1. \tag{4.4}$$

II. By Theorem 7.1, [5], *the random variable* $\sqrt{n}(D_n - \Delta)$ *has a normal limiting distribution with mean zero and variance* $25\,\zeta_1$.

It will be seen in section 6 that in the case of independence $\zeta_1 = 0$, so that the normal limiting distribution of $\sqrt{n}D_n$ is a degenerate one. In this case nD_n has a non-normal limiting distribution. (See section 8).

5. Computation of D. From (4.1) and (3.2) we get after reduction

$$D = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)}, \tag{5.1}$$

where

$$A = \sum_{\alpha=1}^{n} a_\alpha(a_\alpha - 1)\, b_\alpha(b_\alpha - 1),$$

$$B = \sum_{\alpha=1}^{n} (a_\alpha - 1)(b_\alpha - 1)\, c_\alpha, \tag{5.2}$$

$$C = \sum_{\alpha=1}^{n} c_\alpha(c_\alpha - 1),$$

and

$$a_\alpha = \sum_{\beta=1}^{n} C(X_\alpha - X_\beta) - 1, \qquad b_\alpha = \sum_{\beta=1}^{n} C(Y_\alpha - Y_\beta) - 1,$$

$$c_\alpha = \sum_{\beta=1}^{n} C(X_\alpha - X_\beta)C(Y_\alpha - Y_\beta) - 1.$$

$a_\alpha + 1$ and $b_\alpha + 1$ are the ranks of X_α and Y_α, respectively. c_α is the number of sample members (X_β, Y_β) for which both $X_\beta < X_\alpha$ and $Y_\beta < Y_\alpha$. (Since $F(x, y)$ is continuous we may assume that $X_\alpha \neq X_\beta$ and $Y_\alpha \neq Y_\beta$ if $\alpha \neq \beta$.)

Thus, to compute D for a given sample we have to determine the numbers a_α, b_α, c_α for each sample member, calculate A, B, C from (5.2) and insert them in (5.1).

6. The variance of D in the case of independence. Since $F(x, y)$ is assumed to be continuous, so are $F(x, \infty)$ and $F(\infty, y)$. The inequalities $x_1 < x_2$ and $F(x_1, \infty) < F(x_2, \infty)$ are then equivalent unless $F(x_1, \infty) = F(x_2, \infty)$. The same is true of $y_1 < y_2$ and $F(\infty, y_1) < F(\infty, y_2)$. This shows that the function ϕ, (3.2), does not change its value if x_i, y_i is replaced by $F(x_i, \infty)$, $F(\infty, y_i)$, except perhaps on a set of zero probability. Hence Δ and D are invariant under the transformation

$$u = F(x, \infty), \qquad v = F(\infty, y); \qquad U = F(X, \infty), \qquad V = F(\infty, Y).$$

In the case of independence we have $F(x, y) = uv$, and

$$\zeta_k = \int_0^1 \cdots \int_0^1 \{\Phi_k'(u_1, v_1; \cdots; u_k, v_k)\}^2 \, du_1 \, dv_1 \cdots du_k \, dv_k,$$

where Φ_k' is defined as Φ_k, with x_i, y_i and $F(x_i, y_i)$ replaced by u_i, v_i and $u_i v_i$ respectively. On evaluation of these definite integrals we get

$$\zeta_1 = 0, \qquad 200 \cdot 30^2 \zeta_2 = \tfrac{2}{9}, \qquad 600 \cdot 30^2 \zeta_3 = \tfrac{14}{3},$$
$$600 \cdot 30^2 \zeta_4 = \tfrac{164}{9}, \qquad 120 \cdot 30^2 \zeta_5 = 12.$$

On inserting these values in (4.3) we obtain

$$\text{(6.1)} \qquad \operatorname{var}(30D) = \frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)}.$$

Another way to determine the coefficients ζ_k in the case of independence is to compute var D_n for $n = 5, 6, 7$ from the exact distributions given in section 7, and $\lim_{n \to \infty} n^2 \operatorname{var} D_n$ from the asymptotic distribution of nD_n (section 8).

7. The exact distribution of D in the case of independence for $n = 5, 6, 7$. Let $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ be a sample from a population with a continuous d.f. We may confine ourselves to samples with $x_i \neq x_j$ and $y_i \neq y_j$ if $i \neq j$. Let $(x_1', y_{\beta_1}'), \cdots, (x_n', y_{\beta_n}')$ be a rearrangement of $(x_1, y_1), \cdots, (x_n, y_n)$ such that $x_1' < x_2' < \cdots < x_n'$ and $y_1' < y_2' < \cdots < y_n'$. The permutation $\Pi = (\beta_1, \cdots, \beta_n)$ of $(1, \cdots, n)$ will be referred to as the ranking of the sample S.

D_n depends only on the ranking of the sample. We shall express this by writing $D_n = D_n(\Pi) = D_n(\beta_1, \cdots, \beta_n)$. If $(\beta'_{\alpha_1}, \cdots, \beta'_{\alpha_m})$ is a permutation of $m(< n)$ of the integers $1, \cdots, n$ such that $\beta'_1 < \beta'_2 < \cdots < \beta'_m$, $D_m(\beta'_{\alpha_1}, \cdots, \beta'_{\alpha_m})$ is defined to be equal to $D_m(\alpha_1, \cdots, \alpha_m)$. Replacing in (4.1) (X_α, Y_α) by (α, β_α) we find

$$D_n(\beta_1, \cdots, \beta_n) = \binom{n}{5}^{-1} \Sigma' D_5(\beta_{\alpha_1}, \cdots, \beta_{\alpha_5}), \tag{7.1}$$

where Σ' stands for summation over all α such that $1 \leq \alpha_1 < \alpha_2 < \cdots < \alpha_5 \leq n$.

Denoting by $\Pi^{(i)}$ the permutation obtained from $\Pi = (\beta_1, \cdots, \beta_n)$ by omitting β_i, we have the recursion formula

$$nD_n(\Pi) = (n - 5) \sum_{i=1}^{n} D_{n-1}(\Pi^{(i)}). \tag{7.2}$$

From (4.1) and (3.2) we obtain

$$60D_5(\beta_1, \cdots, \beta_5) = \psi(\beta_3, \beta_1, \beta_4)\psi(\beta_3, \beta_2, \beta_5) + \psi(\beta_3, \beta_1, \beta_5)\psi(\beta_3, \beta_2, \beta_4)$$

or

$$60D_5(\beta_1, \cdots, \beta_5) = \begin{cases} 0 & \text{if } \beta_3 \neq 3; \\ 2 & \text{if } \beta_3 = 3 \text{ and } \beta_1, \beta_2 < 3 \quad \text{or } \beta_1, \beta_2 > 3; \\ -1 & \text{if } \beta_3 = 3 \text{ and } \beta_1 < 3, \beta_2 > 3 \text{ or } \beta_1 > 3, \beta_2 < 3. \end{cases} \tag{7.3}$$

We have

$$\begin{aligned} D_n(\beta_1, \cdots, \beta_n) &= D_n(\beta_2, \beta_1, \beta_3, \cdots, \beta_n) \\ &= D_n(\beta_1, \cdots, \beta_{n-2}, \beta_n, \beta_{n-1}) = D_n(\beta_n, \beta_{n-1}, \cdots, \beta_1) \end{aligned} \tag{7.4}$$

For $n = 5$ this follows from (7.3) and for general n from (7.1).

Also, by the symmetry of D_n with respect to x and y, D_n does not change its value if in the permutation $(\beta_1, \cdots, \beta_n)$ the numbers 1, 2 or $n - 1, n$ are interchanged or the permutation is replaced by its inverse.

In the case of independence all $n!$ rankings have the same probability $1/n!$. To find the distribution of D_n we have to determine the number of rankings giving rise to particular values of D_n.

If $n = 5$ there are $5! = 120$ rankings. Owing to (7.4) we need consider only those with $\beta_1 < \beta_2$, $\beta_4 < \beta_5$, $\beta_1 < \beta_4$. Their number is $\frac{120}{8} = 15$. Among them those with $\beta_3 \neq 3$ yield $D_5 = 0$; this leaves only the three permutations

$$(1, 2, 3, 4, 5), \qquad (1, 4, 3, 2, 5), \qquad (1, 5, 3, 2, 4).$$

By (7.3) the respective values of $60D_5$ are $2, -1, -1$. Thus we have

$$P\{60D_5 = 2\} = \tfrac{1}{15}, \qquad P\{60D_5 = -1\} = \tfrac{2}{15},$$
$$P\{60D_5 = 0\} = \tfrac{12}{15}.$$

The distribution of D_6, D_7, $\cdots$ can be obtained in a similar way using the relations (7.1) to (7.4). The distribution of D_n for $n = 5, 6, 7$ is given in Table I.

From (7.3) and (7.1) it follows that $-\frac{1}{60} \leq D_n \leq \frac{1}{30}$ for $n = 5, 6, \cdots$. The upper bound $\frac{1}{30}$ is attained for $\Pi = (1, 2, \cdots, n)$ and every n. To judge by the cases $n = 5, 6, 7$, the minimum of D_n apparently increases with n. From $ED_n = \Delta$ it also follows that $\Delta \leq \frac{1}{30}$.

8. The Asymptotic Distribution of nD_n in the Case of Independence.

THEOREM 8.1. *If $F(x, y) = F(x, \infty)F(\infty, y)$ and $F(x, \infty)$ and $F(\infty, y)$ are continuous, the random variable $nD_n + \frac{1}{36}$ has a limiting distribution whose characteristic function* (c.f.) *is*

$$g(t) = \prod_{k=1}^{\infty} \left(1 - \frac{2it}{k^2 \pi^4}\right)^{-\frac{1}{2}\tau(k)} \tag{8.1}$$

where $\tau(k)$ is the number of divisors of k.

Note that $\tau(k)$ is the number of divisors of k including 1 and k. Thus $\tau(1) = 1$, $\tau(2) = 2$, $\tau(3) = 2$, $\tau(4) = 3$, $\cdots$.

The author has not been able to bring the d.f. corresponding to the c.f. $g(t)$ into a form suitable for numerical computation. Thus Theorem 8.1 may be considered as a preliminary result. For this reason only a brief indication of the proof is given here.

If $(X_1, Y_1), \cdots, (X_n, Y_n)$ is a random sample from a population with d.f. $F(x, \infty)F(\infty, y)$, let $nS_n(x, y)$ be the number of sample members (X_i, Y_i) such that $X_i \leq x$, $Y_i \leq y$. $S_n(x, y)$ is a d.f. depending on the random sample. If we put $F(x, y) = S_n(x, y)$ in $\Delta(F)$ as defined by (3.3), we get

$$\Delta(S_n) = \frac{1}{n^5} \sum_{\alpha_1=1}^{n} \cdots \sum_{\alpha_5=1}^{n} \phi(X_{\alpha_1}, Y_{\alpha_1}; \cdots; X_{\alpha_5}, X_{\alpha_5}).$$

It is easy to prove that if $n\{\Delta(S_n) - E\Delta(S_n)\}$ has a limiting distribution, it is the same as that of nD_n.

Now it can be shown that $n\Delta(S_n)$ has a limiting distribution with the c.f. (8.1). This can be done either analogously to Smirnoff's [6] derivation of the limiting distribution of the goodness of fit statistic ω_n^2, or applying von Mises' [7] general results on the asymptotic distribution of a differentiable statistical function. Though the latter paper deals only with univariate distributions, its results can be extended to the multivariate case.

By expanding $\log g(t)$ in powers of it we obtain for the j-th cumulant κ_j

$$\kappa_j = \frac{2^{5j-3}(j-1)!}{[(2j)!]^2} B_{2j-1}^2,$$

where B_{2j-1} are Bernoulli's numbers,

$$B_1 = \tfrac{1}{6}, \qquad B_3 = \tfrac{1}{30}, \qquad B_5 = \tfrac{1}{42}, \qquad B_7 = \tfrac{1}{30}, \cdots.$$

In particular, $\kappa_1 = \frac{1}{36}$, and since $ED_n = 0$, the limiting distribution of $n\Delta(S_n)$ is that of $nD_n + \frac{1}{36}$.

9. The D-test of Independence. Given a random sample from a bivariate population with continuous d.f., a test for independence can now be carried out as follows:

If $\alpha(0 < \alpha < 1)$ is the desired level of significance, let ρ_n be the smallest number satisfying the inequality

$$P\{D_n > \rho_n \mid F \in \omega\} \leq \alpha,$$

where ω is the class of d.f.'s of the form $F(x, \infty)F(\infty, y)$.

Compute D_n as shown in section 5. Reject the hypothesis H_0 of independence if and only if $D_n > \rho_n$.

For $n = 5, 6, 7$ the numbers ρ_n can be obtained from Table I.

From Tchebychef's inequality and (6.1) we have

$$P\left\{ 30D_n > \sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}} \right\} \leq \alpha.$$

Hence

$$30\rho_n \leq \sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}}.$$

It follows that $\rho_n = O(n^{-1})$.

If $\Delta > 0$, we have $\Delta - \rho_n > 0$ for sufficiently large n. Then

$$P\{D_n > \rho_n\} \geq P\{ \mid D_n - \Delta \mid \leq \Delta - \rho_n\} \geq 1 - (\text{var } D_n)/(\Delta - \rho_n)^2.$$

By (4.4) the right hand side tends to 1.

This, together with Theorem 3.1, shows that the D-test is consistent with respect to the class Ω''.

Since $P\{D_n \leq 0\}$ tends to 0 if $\Delta > 0$, it is safe not to reject H_0 whenever $D_n \leq 0$. An inspection of Table I shows that at least for small n this will happen in more than one-half of the cases if H_0 is true.

10. Concluding Remarks. It would be interesting to compare the power of the D-test with that of other tests with respect to particular alternatives, for instance with the product moment correlation test when the population is normal with correlation ρ. A preliminary investigation seems to indicate that for small values of $\mid \rho \mid$ and $n \to \infty$ the power efficiency of the D-test as compared with the product moment correlation test is rather low. This result may not be conclusive for values of n which are of practical interest. On the other hand, it may be expected that a test which is consistent with respect to a large class of alternatives will have a lower power with regard to a sub-class of alternatives than a test which has optimum properties with respect to this particular sub-class. These considerations suggest the problem of selecting from a given class of non-parametric tests (such as those consistent with respect to Ω'') a test which is most powerful with respect to certain parametric alternatives (such as normal distributions).

TABLE I

The distribution of D_n in the case of independence for $n = 5, 6, 7$.

$n = 5$

x	$15P\{60D_5 = x\}$	$P\{60D_5 \geq x\}$
−1	2	1.0000
0	12	0.8667
2	1	0.0667

$n = 6$

x	$90P\{180D_6 = x\}$	$P\{180D_6 \geq x\}$
−2	4	1.0000
−1	28	0.9556
0	36	0.6444
1	16	0.2444
2	1	0.0667
3	4	0.0556
6	1	0.0111

$n = 7$

x	$630P\{1260D_7 = x\}$	$P\{1260D_7 \geq x\}$
−11	8	1.0000
−8	32	0.9873
−7	32	0.9365
−6	8	0.8857
−5	28	0.8730
−4	88	0.8286
−3	64	0.6889
−2	56	0.5873
−1	8	0.4984
0	88	0.4857
2	77	0.3460
3	24	0.2238
4	4	0.1857
6	56	0.1794
8	8	0.0905
9	4	0.0778
12	24	0.0714
14	2	0.0333
18	12	0.0302
24	2	0.0111
30	4	0.0079
42	1	0.0016

APPENDIX

A. Equiprobable rankings and independence. Let $\Pi_{n\nu}$, $(\nu = 1, 2, \cdots, n!)$ be the $n!$ possible rankings of samples of size n from a bivariate population with continuous d.f. $F(x, y)$ (cf. section 7).

If $F(x, y) = F(x, \infty)F(\infty, y)$ we have

$$P\{\Pi_{n\nu}\} = 1/n! \qquad (\nu = 1, \cdots, n!) \tag{A1}$$

for every n.

Does (A1) for some particular n imply independence? This is not true for $n = 2$. In this case (A1) is equivalent to $P\{(1, 2)\} = \frac{1}{2}$. If the distribution has a p.d. $f(x, y)$, we have

$$P\{(1, 2)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{x}\int_{-\infty}^{y} f(u, v)\, du\, dv + \int_{x}^{\infty}\int_{y}^{\infty} f(u, v)\, du\, dv\right] f(x, y)\, dx\, dy,$$

which equals $\frac{1}{2}$ whenever $f(x, y) = f(-x, y)$. However, we have the following theorem:

THEOREM. *If $F(x, y)$ is in Ω'' and* (A1) *holds for some $n \geq 5$, then*

$$F(x, y) = F(x, \infty)F(\infty, y). \tag{A2}$$

PROOF. (4.2) can be written in the form

$$\sum_{\nu=1}^{n!} D_n(\Pi_{n\nu}P)\{\Pi_{n\nu}\} = \Delta. \tag{A3}$$

If (A1) holds, the left hand side of (A3) has the same value as when (A2) is true. But in the latter case we have $\Delta = 0$. Hence (A1) implies $\Delta = 0$. By Theorem 3.1 this is sufficient for (A2). The proof is complete.

B. Non-existence of unbiased rank tests of independence.

THEOREM. *There do not exist rank tests of independence which are unbiased on any significance level with respect to the classes Ω' or Ω''.*

PROOF: Let $\Pi_{n\nu}$ have the meaning of Appendix A. Any critical region of a rank test of independence is a set $S_m = \{\Pi_{n\nu_1}, \cdots, \Pi_{n\nu_m}\}$ of m rankings. In the case of independence $P(S_m) = P\{\Pi_{n\nu} \in S_m\} = m/n!$ We may confine ourselves to significance levels $m/n!$, $m = 1, 2, \cdots, n! - 1$. To prove the theorem it is sufficient to show that for every $n = 2, 3, \cdots$, for some $m(1 \leq m \leq n! - 1)$ and every S_m there exists a d.f. F in Ω'' such that

$$P(S_m \mid F) < m/n!.$$

We shall prove the slightly more general proposition that this holds for

$$m = 1, 2, 3.$$

Let the bivariate distribution A_n be such that the probability mass is distributed uniformly on the $n - 1$ segments

$$\frac{k-1}{n-1} < x \leq \frac{k}{n-1}, \qquad y - x = \frac{n-2k}{n-1}, \qquad (k = 1, 2, \cdots, n-1), \tag{B1}$$

and is zero in any region not containing a part of these segments.

Let B_n be the distribution which is uniform on the $n - 1$ segments

$$\frac{k-1}{n-1} < x \leq \frac{k}{n-1}, \qquad x + y = \frac{2k-1}{n-1}, \qquad (k = 1, 2, \cdots, n-1), \tag{B2}$$

and zero elsewhere.

The d.f.'s of both A_n and B_n are continuous, with

$$F(x, \infty) = F(\infty, x) = x \qquad (0 \le x \le 1).$$

Since the probability of (X, Y) lying on any one of the segments (B1) or (B2) is $1/(n - 1)$, the probabilities $P(\Pi/A_n)$ and $P(\Pi/B_n)$ are easily obtained in terms of the multinominal distribution with $n - 1$ equal probabilities. In particular, we have

$$\text{(B3)} \qquad P(1, 2, \cdots, n \mid A_2) = 1; \qquad P(n, n - 1, \cdots, 1 \mid B_2) = 1,$$

$$\text{(B4)} \qquad \begin{aligned} P(1, 2, \cdots, n \mid A_n) = P(n, n - 1, \cdots, 1 \mid B_n) &= (n - 1)\left(\frac{1}{n - 1}\right)^n \\ &= \left(\frac{1}{n - 1}\right)^{n-1}, \\ P(n, n - 1, \cdots, 1 \mid A_n) = P(1, 2, \cdots, n \mid B_n) &= 0. \end{aligned}$$

In general, if Π_n is any permutation of $1, \cdots, n$, we have either $P(\Pi_n \mid A_n) = 0$ or $P(\Pi_n \mid B_n) = 0$. For any Π_n with $P(\Pi_n \mid A_n) \ne 0$ contains at least one "run up" of 2 or more numbers (a sequence of consecutive numbers $i, i + 1, \cdots, i + k$) which is not preceded by smaller numbers or followed by larger numbers. On the other hand, if a Π_n' with $P(\Pi_n' \mid B_n) \ne 0$ contains a "run up", it is either preceded by smaller numbers or followed by larger numbers. Hence if $P(\Pi_n \mid A_n) \ne 0$, then $P(\Pi_n \mid B_n) = 0$. Similarly, $P(\Pi_n \mid B_n) \ne 0$ implies $P(\Pi_n \mid A_n) = 0$.

From (B3) it follows that for any set S_m of m rankings which does not include $(1, 2, \cdots, n)$ or $(n, n - 1, \cdots, 1)$ we have either $P(S_m \mid A_2) = 0$ or $P(S_m \mid B_2) = 0$. Hence we need only consider critical regions containing both $(1, 2, \cdots, n)$ and $(n, n - 1, \cdots, 1)$. For $m = 1$ there are no such regions. For $m = 2$ there is just one. But from (B4) it follows that for $n > 2$,

$$\begin{aligned} P(1, 2, \cdots, n \mid A_n) &+ P(n, n - 1, \cdots, 1 \mid A_n) \\ &= \left(\frac{1}{n - 1}\right)^{n-1} < \frac{2}{n}\left(\frac{1}{n - 1}\right)^{n-2} < \frac{2}{n!}. \end{aligned}$$

Finally, if Π_n is any permutation other than $(1, 2, \cdots, n)$ or $(n, n - 1, \cdots, 1)$, we have, by the preceding arguments, either for A_n or for B_n,

$$P(1, 2, \cdots, n) + P(n, n - 1, \cdots, 1) + P(\Pi_n) = \left(\frac{1}{n - 1}\right)^{n-1} < \frac{3}{n!}.$$

This completes the proof for d.f.'s in Ω'. To prove the theorem for d.f.'s in Ω'' we can replace the distributions A_n and B_n by distributions A_n' and B_n' having continuous joint and marginal densities and such that the probabilities $P(\Pi \mid A_n')$ and $P(\Pi \mid B_n')$ differ as little as we please from $P(\Pi \mid A_n)$ and $P(\Pi \mid B_n)$, respectively. For instance, A_2' can be defined by the continuous density

$$
\begin{aligned}
f(x, y) &= K(\epsilon - y + x) && \text{if } 0 \le y - x \le \epsilon, && x \le 1 - \epsilon, y \ge \epsilon; \\
&= K(\epsilon - x + y) && \text{if } -\epsilon \le y - x \le 0, && x \ge \epsilon, \quad y \le 1 - \epsilon; \\
&= K(x + y - \epsilon) && \text{if } x + y \ge \epsilon, && x \le \epsilon, \quad y \le \epsilon; \\
&= K(2 - \epsilon - x - y) && \text{if } x + y \le 2 - \epsilon, && x \ge 1 - \epsilon, y \ge 1 - \epsilon; \\
&= 0 && \text{elsewhere,}
\end{aligned}
$$

where $K = 3/(3\epsilon^2 - 4\epsilon^3)$ and $0 < \epsilon \le \frac{1}{2}$. If ϵ is taken sufficiently small, the distribution satisfies the requirements. The details are left to the reader.

The proof also shows the non-existence of an unbiased rank test of independence for $n = 2$ and any level of significance (for we need consider only one level, $\frac{1}{2}$). It also can be shown that for $n = 3$, any $m = 1, 2, \cdots, 5$ and any S_m the inequality $P(S_m) < m/3!$ holds for at least one of the distributions A_2, A_3, B_2, B_3. The question remains open whether there exist rank tests of independence which are unbiased for some sample sizes n and some significance levels $m/n!$.

REFERENCES

[1] H. Scheffé, "Statistical inference in the non-parametric case," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 305–332.

[2] H. Hotelling and M. R. Pabst, "Rank correlation and tests of significance involving no assumptions of normality," *Annals of Math. Stat.*, Vol. 7 (1936), pp. 29–43.

[3] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, Vol. 30 (1938), pp. 81–93.

[4] J. Wolfowitz, "Additive partition functions and a class of statistical hypotheses," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 247–279.

[5] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *Annals of Math. Stat.*, Vol. 19 (1948), pp. 293–325.

[6] N. V. Smirnoff, "On the distribution of Mises' ω^2-criterion," (Russian, with French summary), *Matematicheskii Sbornik*, Nov. Ser., Vol. 2 (1937), pp. 973–993.

[7] R. von Mises, "On the asymptotic distribution of differentiable statistical functions," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 309–348.