ON A TEST WHETHER TWO SAMPLES ARE FROM THE SAME POPULATION[1]

By A. Wald[2] and J. Wolfowitz

1. The Problem.[3] Let X and Y be two independent stochastic variables about whose cumulative distribution functions nothing is known except that they are continuous. Let $x_1, x_2, \cdots, x_m$ be a set of m independent observations on X and let $y_1, \cdots, y_n$ be a set of n independent observations on Y. It is desired to test the hypothesis (the null hypothesis) that the distribution functions of X and Y are identical.

An important step in statistical theory was made when "Student" proposed his ratio of mean to standard deviation for a similar purpose. In the problem treated by "Student" the distribution functions were assumed to be of known (normal) form and completely specified by two parameters. It is clear that in the problem to be considered here the distributions cannot be specified by any finite number of parameters.

It might nevertheless be argued that by virtue of the limit theorems of probability theory, "Student's" ratio might be used in our problem for large samples. Such a procedure is open to very serious objections. The population distributions may be of such form (e.g., Cauchy distribution) that the limit theorems do not apply. Furthermore, the distributions of X and Y may be radically different and yet have the same first two moments; clearly "Student's" ratio will not distinguish between two such distributions.

The Pearson contingency coefficient is a useful test specifically designed for the problem we are discussing here, but one which also possesses some disadvantages. The location of the class intervals is to a considerable extent arbitrary. In order to use the χ^2 distribution, the numbers in each class interval must not be small; often this can be done only by having large class intervals, thus entailing a loss of information.

2. Preliminary remarks. Denote by $P\{X < x\}$ the probability of the relation in braces. Let $f(x)$ and $g(x)$ be the distribution functions of X and Y respectively; e.g., $P\{X < x\} = f(x)$. Throughout this paper we shall assume that $f(x)$ and $g(x)$ are continuous.

Let the set of $m + n$ elements $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$ be arranged in

---

[1] Presented to the Institute of Mathematical Statistics at Philadelphia, December 27, 1939.

[2] Research under a grant-in-aid from the Carnegie Corporation of New York.

[3] The authors are indebted to Prof. S. S. Wilks for proposing this problem to them.

ascending order of magnitude, and let the sequence be designated by Z, thus: $Z = z_1, z_2, \cdots, z_{m+n}$, where $z_1 < z_2 < \cdots < z_{m+n}$. ($f(x)$ and $g(x)$ were assumed to be continuous. Hence the probability is 0 that $z_i = z_{i+1}$ and therefore we may exclude this case.) Let $V = v_1, v_2, \cdots, v_{m+n}$ be a sequence defined as follows: $v_i = 0$ if z_i is a member of the set $x_1, \cdots, x_m$ and $v_i = 1$ if z_i is a member of the set $y_1, \cdots, y_n$. It is easy to show that any statistic S used to test the null hypothesis should be invariant under any continuous, reciprocally one-to-one transformation of the real axis. That is to say, if $t' = \varphi(t)$ is any such transformation, then

$$(1) \quad S(x_1, \cdots, x_m, y_1, \cdots, y_n) \equiv S(\varphi(x_1), \cdots, \varphi(x_m), \varphi(y_1), \cdots, \varphi(y_n)).$$

The reason for this requirement on S is the fact that the transformed stochastic variables $X' = \varphi(X)$ and $Y' = \varphi(Y)$ are continuous and have identical distributions if and only if X and Y have identical distributions. Hence S must be a function of V only, with the added restriction that $S(V) = S(V')$, where $V' = v_{m+n}, v_{m+n-1}, \cdots, v_1$. For if S were a function of $x_1, \cdots, x_m, y_1, \cdots, y_n$ which cannot be expressed as a function of V alone, then there exists a continuous reciprocally one-to-one transformation $t' = \varphi(t)$ such that (1) is not true. On the other hand, any continuous reciprocally one-to-one transformation of the entire line into itself is monotonic and hence either leaves V invariant or else transforms it into V'.

3. Previous results. In an interesting paper on this problem W. R. Thompson [1] proceeds as follows: Let the sets $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$ be ordered in ascending order of magnitude, thus: $x_{p_1}, x_{p_2}, \cdots, x_{p_m}$ and $y_{p'_1}, y_{p'_2}, \cdots, y_{p'_n}$ where $x_{p_1} < x_{p_2} < \cdots < x_{p_m}$ and $y_{p'_1} < y_{p'_2} < \cdots < y_{p'_n}$. Let $P\{x_{p_k} < y_{p'_{k'}}\}$ denote the probability of the relation in braces under the null hypothesis ($f(x) \equiv g(x)$). This probability is shown to be independent of $f(x)$ and the relation

$$(2) \qquad P\{x_{p_k} < y_{p'_{k'}}\} = \psi(m, n, k, k')$$

holds, where the right member, which is given explicitly by Thompson, is a function only of the arguments exhibited. To make a test of the null hypothesis with, say, a 5% level of significance, this writer proposes to choose k and k' so that $\psi(m, n, k, k') = .05$. The test would then consist of noticing whether $x_{p_k} < y_{p'_{k'}}$ or not. In the former case the null hypothesis is to be considered as disproved.

It is clear that this test cannot be very efficient, ignoring as it does so many of the relations among the observations. Except under certain rather narrow restrictions on the admissible alternatives, for example, that $g(x) \equiv f(x + c)$, where c is an arbitrary constant, the test suffers the further defect of not being "consistent" in a way which will be discussed below. Hence the test suggested by Thompson can scarcely be regarded as a satisfactory solution of the problem. This criticism, of course, does not apply to those sections of Thompson's paper which deal with the question of estimating the so-called normal range.

4. The statistic U. A subsequence $v_{s+1}, v_{s+2}, \cdots, v_{s+r}$ of V (where r may also be 1) will be called a "run" if $v_{s+1} = v_{s+2} = \cdots = v_{s+r}$ and if $v_s \neq v_{s+1}$ when $s > 0$ and if $v_{s+r} \neq v_{s+r+1}$ when $s + r < m + n$. For example, $V = 1, 0, 0, 1, 1, 0$ contains the following runs: 1; 0, 0; 1, 1; 0. The statistic[4] U defined as the number of runs in V seems a suitable statistic for testing the hypothesis that $f(x) \equiv g(x)$. In the event that the latter identity holds, the distribution of U is independent of $f(x)$. A difference between $f(x)$ and $g(x)$ tends to decrease U. U is consistent in a sense which will be discussed below.

In order to derive the distribution of U under the null hypothesis, we first note that all the $\frac{(m+n)!}{m!n!}$ $(= {}^{m+n}C_m)$ possible sequences V have the same probability $\left(= \frac{m!n!}{(m+n)!}\right)$. To see this, consider the sequence V where $v_i = 0$ $(i = 1, 2, \cdots, m)$ and $v_i = 1$ $(i = m+1, m+2, \cdots, m+n)$. Clearly the probability of the sequence is

$$q = \frac{m(m-1)\cdots 1 \cdot n(n-1)\cdots 1}{(m+n)(m+n-1)\cdots(n+1)n(n-1)\cdots 1}.$$

Furthermore, the probability of any other sequence is equal to the product of the factors in the numerator of q taken in a different order, divided by the product of the factors in the denominator taken in the same order. The quotient is, of course, $= q$.

Let e_0 be the number of runs in V whose elements are 0 and let e_1 be the number of runs whose elements are 1. Obviously $U = e_0 + e_1$. Let the runs of each kind be arranged in the ascending order of the indices of the v_i. Let r_{0j} be the number of elements 0 in the j^{th} run of that kind $(j = 1, 2, \cdots, e_0)$ and let $r_{1j'}$ be the number of elements 1 in the j'^{th} run of that kind $(j' = 1, 2, \cdots, e_1)$. The following relations obviously hold:

$$\sum_{j=1}^{e_0} r_{0j} = m, \tag{3}$$

$$\sum_{j'=1}^{e_1} r_{1j'} = n, \tag{4}$$

$$1 \leq e_0 \leq m, \qquad 1 \leq e_1 \leq n, \tag{5}$$

$$|e_0 - e_1| \leq 1. \tag{6}$$

[4] When this paper was already in proof, our attention was called to a paper by W. L. Stevens, entitled "Distribution of groups in a sequence of alternatives," *Annals of Eugenics*, Vol. 9 (1939). There a statistic, which is essentially the U statistic, is proposed for a problem different from that considered by us and the distribution of U is obtained in a different manner. However, the application of the U statistic for the purpose herein described, the proof of consistency and the other results of our paper are not contained in it.

Hence if $U = 2k$, then $e_0 = e_1 = k$, and if $U = 2k - 1$, then either $e_0 = k$, $e_1 = k - 1$ or $e_0 = k - 1$, $e_1 = k$. The element v_1 of V together with the numbers $r_{01}, r_{02}, \cdots, r_{0e_0}, r_{11}, r_{12}, \cdots, r_{1e_1}$, completely determines the sequence V whose probability is q.

Without loss of generality we may assume that $m \leq n$. If $U = 2k$, $1 \leq k \leq m$, $v_1 = 0$, any two sequences of k positive numbers each may constitute a sequence of $r_{01}, \cdots, r_{0e_0}, r_{11}, \cdots, r_{1e_1}$ provided only that (3) and (4) are satisfied. The number of sequences $r_{01}, r_{02}, \cdots, r_{0k}$ which satisfy (3) is the coefficient of a^m in the purely formal expansion of

$$(a + a^2 + a^3 + \cdots)^k = \left(\frac{a}{1 - a}\right)^k$$

and hence is ${}^{m-1}C_{k-1}$. Similarly the number of sequences $r_{11}, r_{12}, \cdots, r_{1k}$ which satisfy (4) is found to be ${}^{n-1}C_{k-1}$. Bearing in mind the case $U = 2k$, $v_1 = 1$, we obtain

$$P\{U = 2k\} = \frac{2({}^{m-1}C_{k-1} \cdot {}^{n-1}C_{k-1})}{{}^{m+n}C_m}, \quad (k = 1, 2, \cdots, m), \tag{7}$$

where the left member denotes the probability of the relation in braces under the null hypothesis. In a similar manner we obtain

$$P = \{U = 2k - 1\} = \frac{({}^{m-1}C_{k-1} \cdot {}^{n-1}C_{k-2} + {}^{m-1}C_{k-2} \cdot {}^{n-1}C_{k-1})}{{}^{m+n}C_m}, \tag{8}$$

$$(k = 2, \cdots, m + 1),$$

with the proviso that ${}^aC_b = 0$ if $a < b$.

We shall now briefly indicate a method of obtaining the mean $E(U)$ and variance $\sigma^2(U)$ of U. For example, $E(U)$ may be obtained by performing several summations of the type

$$\sum_{i=0}^{m-1} i \cdot {}^{m-1}C_i \cdot {}^{n-1}C_i. \tag{9}$$

It is easy to verify that the expression (9) is the term free of a in the purely formal expansion in a of:

$$(m - 1) \cdot (1 + a)^{m-2} \cdot a \cdot \left(1 + \frac{1}{a}\right)^{n-1}, \tag{10}$$

and hence is

$$(m - 1) \cdot {}^{m+n-3}C_{n-2}. \tag{11}$$

The other summations required for the mean and variance can be carried out in a similar manner. We shall omit these tedious calculations. The results are:

$$E(U) = \frac{2mn}{m+n} + 1, \tag{12}$$

$$\sigma^2(U) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}. \tag{13}$$

The critical region for testing the null hypothesis on a level of significance β is given by the inequality $U < u_0$, where u_0 is a function of m and n such that $P\{U < u_0\} = \beta$.

5. The asymptotic distribution of U. Let $m/n = \alpha$, a positive constant. Then, as $m \to \infty$,

$$E(U) \sim \frac{2m}{1+\alpha},$$

$$\sigma^2(U) \sim \frac{4\alpha m}{(1+\alpha)^3}.$$

THEOREM I. *If t is any real number, the probability of the relation* $U < \frac{2m}{1+\alpha} + 2\left[\frac{\alpha m}{(1+\alpha)^3}\right]^{\frac{1}{2}} t$ *converges uniformly in t to*

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}w^2}\, dw$$

as $m \to \infty$.

The proof of this theorem is essentially the same as the classical proof that the binomial law converges to the normal distribution (see, for example, Fréchet [2], p. 89) and it will be unnecessary to give the details. Since the asymptotic distribution of the subpopulation of even U is the same as that of odd U, it will be sufficient to consider only the right member of (7). Let $m' = m - 1$, $n' = n - 1$, and $k' = k - 1$. We make the substitution

$$w = \frac{k' - \frac{m'}{1+\alpha'}}{\sqrt{m'}}, \quad \text{where} \quad \alpha' = \frac{m'}{n'}, \tag{14}$$

$$dw = \frac{1}{\sqrt{m'}}, \tag{15}$$

and evaluate the factorials by Stirling's formula. We shall give here only the results of successive simplifications. At each step we shall omit the factors free of k or w, since their product may be reconstructed from the final exponential form Thus instead of the right member of (7) we can consider the expression:

$$^{m-1}C_{k-1} \cdot {}^{n-1}C_{k-1}. \tag{16}$$

Omitting factors free of k, we get

$$\frac{1}{(k-1)!\,(m-k)!\,(k-1)!\,(n-k)!} \tag{17}$$

and by Stirling's formula, since k and m are both large:

$$\frac{1}{k'^{(2k'+1)}(m'-k')^{(m'-k'+\frac{1}{2})}(n'-k')^{(n'-k'+\frac{1}{2})}}. \tag{18}$$

Now apply (14). We obtain

$$\left(\sqrt{m'}\,w+\frac{m'}{1+\alpha'}\right)^{-2\sqrt{m'}w-\frac{2m'}{1+\alpha'}-1}\cdot\left(-\sqrt{m'}\,w+\frac{m'\alpha'}{1+\alpha'}\right)^{\sqrt{m'}w-\frac{m'\alpha'}{1+\alpha'}-\frac{1}{2}} \cdot\left(-\sqrt{m'}\,w+\frac{m'}{\alpha'(1+\alpha')}\right)^{\sqrt{m'}w-\frac{m'}{\alpha'(1+\alpha')}-\frac{1}{2}}. \tag{19}$$

Dividing inside the parentheses by $\frac{m'}{1+\alpha'}$, $\frac{m'\alpha'}{1+\alpha'}$, $\frac{m'}{\alpha'(1+\alpha')}$, respectively, and again omitting factors free of w, we get

$$\left(1+\frac{(1+\alpha')w}{\sqrt{m'}}\right)^{-2\sqrt{m'}w-\frac{2m'}{1+\alpha'}-1}\cdot\left(1-\frac{(1+\alpha')w}{\alpha'\sqrt{m'}}\right)^{\sqrt{m'}w-\frac{m'\alpha'}{1+\alpha'}-\frac{1}{2}} \cdot\left(1-\frac{\alpha'(1+\alpha')w}{\sqrt{m'}}\right)^{\sqrt{m'}w-\frac{m'}{\alpha'(1+\alpha')}-\frac{1}{2}}. \tag{20}$$

Taking logarithms, expanding in powers of $\frac{w}{\sqrt{m'}}$ and neglecting terms in $\frac{w^3}{m'^{\frac{1}{2}}}$ and higher orders, the results are

$$\begin{aligned} &-\left(2\sqrt{m'}\,w+\frac{2m'}{1+\alpha'}+1\right)\left(\frac{(1+\alpha')w}{\sqrt{m'}}-\frac{(1+\alpha')^2w^2}{2m'}\right)\\ &-\left(\sqrt{m'}\,w-\frac{m'\alpha'}{1+\alpha'}-\frac{1}{2}\right)\left(\frac{(1+\alpha')w}{\alpha'\sqrt{m'}}+\frac{(1+\alpha')^2w^2}{2\alpha'^2m'}\right)\\ &-\left(\sqrt{m'}\,w-\frac{m'}{\alpha'(1+\alpha')}-\frac{1}{2}\right)\left(\frac{\alpha'(1+\alpha')w}{\sqrt{m'}}+\frac{\alpha'^2(1+\alpha')^2w^2}{2m'}\right) \end{aligned} \tag{21}$$

which equals

$$-\frac{w^2(1+\alpha')^3}{2\alpha'}+O(m'^{-\frac{1}{2}}). \tag{22}$$

The proof of the fact that the distribution of w converges uniformly to the normal distribution with zero mean and variance $\frac{\alpha'}{(1+\alpha')^3}$ can be carried out in the same way as the classical proof that the binomial law converges to the normal distribution.

It is obvious that

$$w^* = \frac{k - \frac{m}{1+\alpha}}{\sqrt{m}}$$

has the same distribution as w. From this and from the fact that $U = 2k$ or $2k - 1$ THEOREM I follows.

In using conventional tables of the Gaussian function to make tests of significance on U when m and n are large, the reader is urged not to forget that the critical region of U lies in only one tail of the curve.

6. An example. We give here a simple example illustrating the use of the statistic U and THEOREM I.

Suppose 50 observations were made on X and 50 observations on Y. Suppose further that these observations are arranged in ascending order and that the i^{th} element of this sequence is said to have the rank i. The observations on X occupy the following ranks: 1, 5, 6, 7, 12, 13, 14, 15, 16, 17, 19, 20, 21, 25, 26, 27, 28, 31, 32, 38, 42, 43, 44, 45, 50, 51, 52, 53, 54, 56, 57, 58, 62, 63, 64, 65, 68, 69, 75, 79, 80, 81, 86, 87, 89, 90, 91, 93, 94, 95.

The observations on Y occupy the remaining ranks.

In this case, $U = 34$.

For $m = n = 50$,

$$E(U) = 51,$$
$$\sigma^2(U) = 24.747.$$

The probability of getting 34 runs or less when the distribution functions of X and Y are continuous and identical is therefore less than $5 \cdot 10^{-4}$.

7. Consistency. We shall say that a test is "consistent" if the probability of rejecting the null hypothesis when it is false (i.e., the complement of the probability of a type II error, cf. Neyman and Pearson, [3]) approaches one as the sample number approaches infinity. In the literature of statistics a function of the observations which converges stochastically to a population parameter as the sample number approaches infinity, is called a "consistent" statistic. If a test of a hypothesis about a population parameter is made by a proper use of a consistent (statistic) estimate of the parameter, the test will be consistent also according to our definition, which thus furnishes an extension of the idea of consistency to the case where the alternatives to the null hypothesis cannot be specified by a finite number of parameters.

It is obvious that consistency ought to be a minimal requirement of any good test. It is the purpose of this section to prove that, subject to some slight and from the practical statistical point of view, unimportant, restrictions on the distribution functions, the test furnished by the statistic U is consistent.

We shall say that the distribution functions $f(x)$ and $g(x)$ satisfy the condition A, if, for any arbitrarily small positive δ, there exist a finite number of

closed intervals, such that the probability of the sum I of these intervals is $> 1 - \delta$ according to at least one of the distribution functions $f(x)$ and $g(x)$, and such that $f(x)$ and $g(x)$ have positive continuous derivatives $f'(x)$ and $g'(x)$ in I.

In all that follows, although m and n are considered as variables, their ratio m/n is to be a constant, denoted by α. Let $\beta > 0$ denote the level of significance on which the test is to be made, so that, if $f(x) \equiv g(x)$,

$$P\{U < u_0(m)\} = \beta \tag{23}$$

where the critical region for two samples of size m and n, respectively, is given by

$$U < u_0(m).$$

THEOREM II. *If $f(x)$ and $g(x)$ satisfy condition* A, *and if*

$$f(x) \not\equiv g(x), \tag{24}$$

then

$$\lim_{m\to\infty} P\{U < u_0(m)\} = 1. \tag{25}$$

The proof of this theorem will be given in several stages.

Let $E\left(\frac{U}{m}; f; g\right)$ and $\sigma^2\left(\frac{U}{m}; f; g\right)$ denote the mean and variance, respectively, of $\frac{U}{m}$, when X and Y have the distribution functions $f(x)$ and $g(x)$, respectively, and the sample numbers are m and n. Let the set $x_1 \cdots x_m$; $y_1 \cdots y_n$ be arranged in ascending order of magnitude, thus:

$$Z = z_1, z_2, \cdots, z_{m+n}, \tag{26}$$

where $z_1 < z_2 < \cdots < z_{m+n}$. The sequence

$$V = v_1, v_2, \cdots, v_{m+n} \tag{27}$$

is defined as follows: $v_i = 0$ if z_i is a member of the set $x_1 \cdots x_m$ and $v_i = 1$ if z_i is a member of the set $y_1 \cdots y_n$.

LEMMA 1. *If the following are fulfilled*:

a) $f(x) \equiv 0 \quad x < 0,$

$f(x) \equiv x \quad 0 \leq x \leq 1,$

$f(x) \equiv 1 \quad x > 1.$

b) $g(x) \equiv 0 \quad x \leq 0,$

$g(x) \equiv 1 \quad x \geq 1.$

c) *The derivative $g'(x)$ of $g(x)$ exists, is continuous and positive everywhere in the interval $0 \leq x \leq 1$.*

d) *k is an arbitrary but fixed positive integer. For every m, $i_{1m} < i_{2m} < \cdots < i_{km}$ are a set of k positive integers subject only to the restriction that the least upper bound γ of the sequence $\frac{i_{km}}{m+n}$ is less than* 1.

Then the expected value

$$E\left(\prod_{j=1}^{k} v_{i_{jm}}\right) \text{ of } \prod_{j=1}^{k} v_{i_{jm}}$$

satisfies the inequality

$$\left| E\left(\prod_{j=1}^{k} v_{i_{jm}}\right) - \prod_{j=1}^{k} \frac{g'(a_{\lambda_{jm}})}{\alpha + g'(a_{\lambda_{jm}})} \right| < \varphi(m) \tag{28}$$

where $\lambda_{jm} = \frac{i_{jm}}{m+n}$ and $a_{\lambda_{jm}}$ $(j = 1 \cdots k)$ *is the root of*

$$ma_{\lambda_{jm}} + ng(a_{\lambda_{jm}}) = \lambda_{jm}(m+n) \tag{29}$$

and $\varphi(m)$ depends only on m and is such that

$$\operatorname*{Lim}_{m\to\infty} \varphi(m) = 0. \tag{30}$$

It is easy to verify that the root $a_{\lambda_{jm}}$ of (29) exists and is unique.

PROOF: It will be sufficient to show that, for any specified set of values of

$$v_{i_{1m}} \cdots v_{i_{(r-1)m}}, \qquad v_{i_{(r+1)m}} \cdots v_{i_{km}} \qquad (r = 1 \cdots k)$$

the conditional probability $P\{v_{i_{rm}} = 1\}$ of the relation in braces satisfies the inequality

$$\left| \frac{g'(a_{\lambda_{rm}})}{\alpha + g'(a_{\lambda_{rm}})} - P\{v_{i_{rm}} = 1\} \right| < \psi(m), \tag{31}$$

where $\psi(m)$ depends only on m and is such that

$$\operatorname*{Lim}_{m\to 0} \psi(m) = 0. \tag{32}$$

For each m let

$$V'_m = v'_{i_{1m}}, v'_{i_{2m}} \cdots v'_{i_{(r-1)m}}, v'_{i_{(r+1)m}} \cdots v'_{i_{km}} \tag{33}$$

be a fixed sequence whose elements are either 0 or 1. We shall consider the conditional probability $P\{v_{i_{rm}} = s\}$, $(s = 0, 1)$ of the relation in braces subject to the condition that

$$v_{i_{jm}} = v'_{i_{jm}}, \qquad (j = 1, 2, \cdots (r-1), (r+1), (r+2), \cdots k). \tag{34}$$

Let a and b be two numbers such that $0 < a < b < 1$, and let m^* be a non-negative integer such that $m^* \leq m$, and $m^* \leq [\gamma(m+n)]$ where $[\gamma(m+n)]$ denotes the largest integer $\leq \gamma(m+n)$. Let $Q_m(a, b, m^*)$ denote the proba-

bility that, if m^* observations are made on X and $[\gamma(m+n)] - m^*$ observations are made on Y, the following conditions will be fulfilled:

(a) the total number of observations $< a'$ is exactly $i_{rm} - 1$

(b) all observations are $< b$

(c) if the $[\gamma(m+n)]$ observations are arranged in ascending order and if $v_j^* = 0$ or 1 according as the j^{th} element is an observation on X or on Y, then

$$v^*_{i_{jm}} = v'_{i_{jm}} \qquad (j = 1, 2, \cdots, r-1), \tag{35}$$

and

$$v^*_{i_{jm}-1} = v'_{i_{jm}} \qquad (j = r+1, r+2 \cdots k). \tag{36}$$

It is easy to see that the probability P_0 of the simultaneous fulfillment of the relations (34) and of $v_{i_{rm}} = 0$ is given by

$$P_0 = \int_0^1 \int_0^b \sum_{m^*} R_m(a, b, m^*) m'(1-b)^{m'-1}(1-g(b))^{n'}\, da\, db, \tag{37}$$

where

$$R_m(a, b, m^*) = {}^mC_{m^*}\, {}^nC_{[\gamma(m+n)]-m^*} \frac{dQ_m}{db}(a, b, m^*), \tag{38}$$

$$m' = m - m^*, \tag{39}$$

and

$$n' = n - [\gamma(m+n)] + m^*. \tag{40}$$

Similarly, the probability P_1 of the simultaneous fulfillment of the relations (34) and of $v_{i_{rm}} = 1$ is given by

$$P_1 = \int_0^1 \int_0^b \sum_{m^*} R_m(a, b, m^*)\, n' g'(a)(1-b)^{m'}(1-g(b))^{n'-1}\, da\, db. \tag{41}$$

Then

$$\frac{P\{v_{i_{rm}} = 0\}}{P\{v_{i_{rm}} = 1\}} = \frac{P_0}{P_1}. \tag{42}$$

Let $n_0 = \sum_{j > [\gamma(m+n)]} v_j$ and $m_0 = m + n - [\gamma(m+n)] - n_0$. The variables $(z_{i_{rm}} - a_{\lambda_{rm}})$, $(z_{[\gamma(m+n)]} - a_\gamma)$, $\left(\frac{m_0}{n_0} - \frac{\alpha(1-a_\gamma)}{(1-g(a_\gamma))}\right)$ all converge stochastically to zero.

Let $P_0(\epsilon)$ and $P_1(\epsilon)$ denote the values of the right members of (37) and (41), respectively, if the integration is restricted to the region where $a \leq b$, $|a - a_{\lambda_{rm}}| < \epsilon$, $|b - a_\gamma| < \epsilon$ and the summation is restricted to those values

of m^* for which $\left|\frac{m'}{n'} - \frac{\alpha(1-a_\gamma)}{(1-g(a_\gamma))}\right| < \epsilon$. Hence, because of the aforementioned stochastic convergence, for all sufficiently large m

$$|P_s(\epsilon) - P_s| < \epsilon \qquad s = 1, 2. \tag{43}$$

Since $P_s > 0$, for sufficiently large m, also

$$\left|\frac{P_0(\epsilon)}{P_1(\epsilon)} - \frac{P_0}{P_1}\right| < \epsilon. \tag{44}$$

Since $g(x)$ and $g'(x)$ are continuous in the interval $[0, 1]$ and hence uniformly continuous, it is clear that

$$\left|\frac{P_0(\epsilon)}{P_1(\epsilon)} - \frac{\alpha}{g'(a_{\lambda_{rm}})}\right| < c\epsilon, \tag{45}$$

where c is a fixed constant independent of m. From (44) and (45) it follows easily that, for any arbitrarily small ϵ',

$$\left|\frac{P_0}{P_1} - \frac{\alpha}{g'(a_{\lambda_{rm}})}\right| < \epsilon' \tag{46}$$

for sufficiently large m.

Since $P\{v_{i_{rm}} = 1\} = \frac{P_1}{P_0 + P_1}$, the required relation (31) follows. This completes the proof of LEMMA 1.

LEMMA 2. *If conditions a, b, and c of Lemma* 1 *are satisfied, then*

$$\lim_{m\to\infty} E\left(\frac{U}{m}; f; g\right) = 2\int_0^1 \frac{g'(x)}{\alpha + g'(x)}\,dx \tag{47}$$

and

$$\lim_{m\to\infty} \sigma^2\left(\frac{U}{m}; f; g\right) = 0. \tag{48}$$

PROOF: Since

$$\begin{aligned}\frac{U}{m} &= \frac{1}{m} + \frac{1}{m}\sum_{j=2}^{m+n}(v_j - v_{j-1})^2 \\ &= \frac{1 + v_1 + v_{m+n}}{m} + \frac{2}{m}\sum_{j=2}^{m+n-1} v_j - \frac{2}{m}\sum_{j=2}^{m+n} v_{j-1}v_j,\end{aligned} \tag{49}$$

we have from LEMMA 1,

$$\begin{aligned}E\left(\frac{U}{m}\right) &= \frac{2}{m}\left[\sum_j \frac{g'(a_{jm})}{\alpha + g'(a_{jm})} - \sum_j \left(\frac{g'(a_{jm})}{\alpha + g'(a_{jm})}\right)^2\right] + \eta(m) + \eta^*(\gamma) \\ &= \frac{2}{m}\sum\left[\frac{\alpha g'(a_{jm})}{(\alpha + g'(a_{jm}))^2}\right] + \eta(m) + \eta^*(\gamma),\end{aligned} \tag{50}$$

where

$$\operatorname{Lim}_{m\to\infty} \eta(m) = \operatorname{Lim}_{\gamma\to 1} \eta^*(\gamma) = 0 \tag{51}$$

and a_{jm} is the root of the equation

$$ma_{jm} + ng(a_{jm}) = j \qquad (j = 2 \cdots m + n). \tag{52}$$

From equation (52) it follows that

$$\operatorname{Lim}_{m\to\infty} (a_{jm} - a_{(j-1)m})(m + ng'(a_{jm})) = 1 \tag{53}$$

uniformly in j. Since γ may be chosen arbitrarily near to 1, the required result (47) follows easily from (50).

It remains to consider the variance of $\frac{U}{m}$. The expression

$$\frac{1 + v_1 + v_{m+n}}{m} + \frac{2}{m} \sum_{j=2}^{m+n-1} v_j$$

differs from $\frac{2}{\alpha}$ by at most $\frac{1}{m}$, so that its variance converges to zero with $m \to \infty$. In order to prove (48), it will be sufficient to show that the variance of

$$W = \frac{1}{m} \sum_{j=2}^{m+n} v_{j-1}v_j \tag{54}$$

goes to zero with increasing m. From LEMMA 1 it follows that

$$-z(m) < [E(v_iv_jv_kv_e) - E(v_iv_j)E(v_kv_e)] < z(m), \tag{55}$$

where $\operatorname{Lim}_{m\to\infty} |z(m)| = 0$, provided only that the integers i, j, k, l are distinct and $< \gamma(m + n)$. The variance of mW is the sum of terms of the type occurring in (55). The number of terms for which i, j, k, l are distinct is of the order m^2. All other terms are of size at most 2 and their number is of the order m. Since the number γ may be chosen arbitrarily near to 1, the variance of W converges to zero with $m \to \infty$.

This proves LEMMA 2.

LEMMA 3. *If conditions a, b, and c of Lemma* 1 *are fulfilled, and if* (24) *holds, then*

$$T = \int_0^1 \frac{g'(x)}{\alpha + g'(x)}\,dx < \frac{1}{1 + \alpha}. \tag{56}$$

Let $a_1 < a_3$ be any two real numbers and designate $\frac{a_1 + a_3}{2}$ by a_2. Let $F(x)$ be defined as follows:

$$\begin{aligned} &F(a_1) = 0, \\ &F(x) = (x - a_i)b_i + F(a_i), \qquad (a_i \leq x \leq a_{i+1}\,;\, i = 1, 2). \end{aligned} \tag{57}$$

Let c be defined by

$$F(a_3) = c(a_3 - a_1). \tag{58}$$

Then it is easy to verify that the maximum of

$$T^* = \int_{a_1}^{a_3} \frac{F'(x)}{\alpha + F'(x)} dx \tag{59}$$

with respect to b_1 and b_2, subject to the restrictions that b_1 and b_2 be non-negative, and that a_1, a_3 and c be fixed ($c > 0$), occurs when and only when

$$b_1 = b_2 = c. \tag{60}$$

Now define

$$\begin{aligned} P_{ij} &= \frac{i}{2^j}, \qquad P_{0j} = 0, \\ l_{ij} &= \frac{g(P_{ij}) - g(P_{(i-1)j})}{2^j} \end{aligned} \tag{61}$$

and

$$S_j = \frac{1}{2^j} \sum_{i=1}^{2^j} \frac{l_{ij}}{\alpha + l_{ij}}, \qquad (i = 1, 2, \cdots 2^j; j = 0, 1, 2 \cdots).$$

Repeated application of the result of the previous paragraph easily gives

$$S_j \geq S_{j+1}. \tag{62}$$

From (24) it follows that there exists a positive integer j' such that $S_{j'} > S_{j'+1}$. Obviously

$$S_0 = \frac{1}{1 + \alpha} \tag{63}$$

and

$$\lim_{j \to \infty} S_j = T. \tag{64}$$

Hence LEMMA 3 is proved.

Proof of Theorem II: Let $\delta_1 > \delta_2 > \cdots > \delta_j > \cdots$ be an arbitrary but fixed sequence such that $\lim \delta_j = 0$. For $\delta = \delta_j$, let $I_1, \cdots, I_{k(j)}$ be a set of closed intervals such that no two intervals have an interior point in common and within which, by condition (A), $f'(x)$ and $g'(x)$ exist, are positive, and continuous. Let I_{0j} be the complementary set (with respect to the whole line). (It is easy to see that, if condition (A) is fulfilled, such a system can be constructed.) Let $U_i (i = 1, 2 \cdots k(j)$ and U_{0j} denote, respectively, the runs caused by the observations which fall in the intervals I_i, I_{0j}. Then

$$\left| U - \sum_{i=1}^{k(j)} U_i - U_{0j} \right| \leq 2(k(j)). \tag{65}$$

From condition (A) it follows that, with a probability arbitrarily close to 1, for sufficiently large m,

$$U_{0j} < 3pm\delta_j, \tag{66}$$

where

$$p = \max\left[1, \frac{1}{\alpha}\right], \qquad (j = 1, 2 \cdots).$$

Let $[a_i \leq x < b_i]$, $i = 1, 2 \cdots$ denote the interval I_i, and let m_i and n_i denote the number of observations on X and Y, respectively, which fall in the interval I_i. Then $\frac{m_i}{m}$ and $\frac{n_i}{n}$ converge stochastically with increasing m to $[f(b_i) - f(a_i)]$ and $[g(b_i) - g(a_i)]$, respectively.

Within the interval $I_i (i = 1, 2 \cdots k)$ we perform the transformation

$$X^* = f(X), \qquad Y^* = f(Y), \tag{67}$$

which leaves U_i invariant. For fixed m_i, n_i the relative distribution of X^* is uniform and the relative distribution of Y^* fulfills condition (c) of LEMMA 1. Hence from LEMMA 2 we obtain that $\frac{U_i}{m}$ converges stochastically to

$$\operatorname*{Lim}_{m\to\infty} E\left(\frac{U_i}{m}; f; g\right) \leq \frac{2[f(b_i) - f(a_i)][g(b_i) - g(a_i)]}{[g(b_i) - g(a_i)] + \alpha[f(b_i) - f(a_i)]}. \tag{68}$$

It can be verified that the sum of the second members in (68) over all values i is less than or equal to $\frac{2}{1 + \alpha}$.

From (24) and condition (A) we get that, for sufficiently small δ_j, there exists at least one interval for which the first member of (68) is less than the second member. Hence

$$\Sigma < \frac{2}{1 + \alpha}, \tag{69}$$

where

$$\Sigma = \sum_{i=1}^{\infty} \operatorname*{Lim}_{m\to\infty} E\left(\frac{U_i}{m}; f; g\right). \tag{70}$$

Now take j so large that

$$3p\delta_j < \epsilon, \tag{71}$$

where

$$0 < 3\epsilon < \frac{2}{1 + \alpha} - \Sigma. \tag{72}$$

Since $\frac{U_i}{m}$ converges stochastically to its expected value, from (65), (66), (70), (71), and (72), it follows that, with a probability arbitrarily close to 1, for sufficiently large m,

$$\frac{U}{m} < \frac{2}{1+\alpha} - \epsilon. \tag{73}$$

From (23) and THEOREM I we get

$$\lim_{m\to\infty} \frac{u_0(m)}{m} = \frac{2}{1+\alpha}. \tag{74}$$

THEOREM II follows easily from (73) and (74).

8. Remarks on a proposed test. We have already remarked in Section 3 that the test proposed by W. R. Thompson is not consistent. To show this, we shall give two distribution functions $f(x)$ and $g(x)$ such that, although these functions will be very different, the probability of rejecting the hypothesis that they are the same will not approach one as the sample number approaches infinity.

Suppose, to simplify the notation, that the observations have been ordered according to size, i.e., that $x_1 < x_2 < \cdots < x_m$ and $y_1 < y_2 < \cdots < y_n$. Suppose further than $m = n$, and that the test is to be made on a level of significance $\beta > 0$. In the right member of (2) we need not exhibit n and shall replace k and k' by $k(m)$ and $k'(m)$ to show the dependence on m. We have, under the null hypothesis,

$$P\{x_{k(m)} < y_{k'(m)}\} = \psi(m, k(m), k'(m)) = \beta. \tag{75}$$

The sequence $\frac{k(m)}{m}$ is bounded, so that there exists a monotonically increasing subsequence $m_1, m_2 \cdots$ of the sequence of integers $1, 2 \cdots$ and a number h, $0 \leq h \leq 1$, such that

$$\lim_{i\to\infty} \frac{k(m_i)}{m_i} = h. \tag{76}$$

It is easy to see that then also

$$\lim_{i\to\infty} \frac{k'(m_i)}{m_i} = h. \tag{77}$$

We shall now assume that $0 < h < 1$. If $h = 0$ or 1 only a trivial alteration will be needed in the argument to follow. Let ϵ and δ be arbitrarily small positive numbers. We now consider two populations, A and B described as follows:

A) $$f(x) \equiv g(x) \equiv x \qquad (0 \leq x \leq 1),$$

B) $$f(x) \equiv x \qquad (0 \leq x \leq 1),$$

$$g(x) \equiv g(a_i) + \frac{(x - a_i)(g(a_{i+1}) - g(a_i))}{(a_{i+1} - a_i)} \qquad (a_i \leq x \leq a_{i+1}; i = 0, 1, \cdots, 4),$$

where

$$\begin{aligned} a_0 &= 0 & g(a_0) &= 0 \\ a_1 &= h - 2\delta > 0 & g(a_1) &= 0 \\ a_2 &= h - \delta & g(a_2) &= a_2 \\ a_3 &= h + \delta < 1 - \delta & g(a_3) &= a_3 \\ a_4 &= 1 - \delta & g(a_4) &= a_3 \\ a_5 &= 1 & g(a_5) &= 1 \end{aligned}$$

The definition of $f(x)$ and $g(x)$ outside the interval $0 \leq x \leq 1$ is obvious. It will be shown that even for such different populations as A and B and for samples of size greater than that of any arbitrarily assigned number, the probability of rejecting the null hypothesis if B is true will be at most $\beta + \epsilon$.

Let h_1, h_2, h_3 denote the number of observations on X which fall in the intervals $0 < x \leq a_2$, $a_2 < x \leq a_3$, $a_3 < x \leq 1$, respectively (m fixed, of course). Let h_1', h_2', h_3' be the corresponding numbers for Y. For a fixed m, the probability of a set h_1, h_2, h_3, h_1', h_2', h_3' is the same whether the sample be drawn from the population A or B. From (76), (77), and multinomial law it follows that for all sufficiently large m_i the probability is at least $1 - \epsilon$ of the occurrence of a set h_1, h_2, h_3, h_1', h_2', h_3' for which $x_{k(m_i)}$ and $y_{k'(m_i)}$ will both fall in the interval $a_2 < x \leq a_3$. Furthermore it is obvious that for all samples with fixed h_2, h_2' the distribution within the interval $a_2 < x \leq a_3$ is the same whether the sample came from the population A or B. Hence even when the sample is drawn from the population B, the first member of (75) is $< \beta + \epsilon$. This completes the proof of the inconsistency of the test based on (75).

This test is consistent if the alternatives to the null hypothesis are limited, for example, to those where $g(x) \equiv f(x + c)$, c a constant.

REFERENCES

[1] WILLIAM R. THOMPSON, *Annals of Math. Stat.*, Vol. 9, (1938), p. 281.
[2] MAURICE FRÉCHET, *Généralités sur les Probabilités. Variables aléatoires*, Paris, (1937).
[3] J. NEYMAN AND E. S. PEARSON. *Statistical Research Memoirs.* University College, London. Vol. 1, (1936).

COLUMBIA UNIVERSITY,
NEW YORK, N. Y.