A COMBINATORIAL TEST FOR THE PROBLEM OF TWO SAMPLES FROM CONTINUOUS DISTRIBUTIONS

S. S. WILKS
PRINCETON UNIVERSITY

1. Introduction

The general problem of testing the null hypothesis that two independent samples come from identical continuous distributions against the alternative hypothesis that they come from *any* pair of different continuous distributions, has been considered by Smirnov [11], Wald and Wolfowitz [13], and others. The two-sample problem for testing the null hypothesis against various *restricted classes* of pairs of alternatives has been considered by Dixon [3], Wilcoxon [14], Mann and Whitney [7], Lehmann [5], Mood [8], Savage [10], Sukhatme [12], and other authors.

The purpose of this paper is to consider a simple combinatorial test for the general two-sample problem based on what are called "cell frequency counts" which one sample generates with respect to the other. The test proposed is consistent for testing the null hypothesis against alternatives in the class of all pairs of different continuous distributions subject to mild assumptions. The test criterion suggested, defined by (4.8), has as its limiting distribution in large samples a chi-square distribution under the null hypothesis. The power of the test is considered in some detail for alternatives in which the two distributions are "nearly" equal.

To be more precise let $\mathcal{C}$ be the class of all pairs of continuous c.d.f.'s $(F(x), G(x))$ and let $\mathcal{C}_0$ be the subset of $\mathcal{C}$ for which $F(x) \equiv G(x)$.

Let $(x_{(1)}, \cdots, x_{(n)})$, with $x_{(1)} < \cdots < x_{(n)}$, be the order statistics of a sample O_n from $F(x)$ and let $I_1, \cdots, I_{n+1}$ be the intervals $(-\infty, x_{(1)}]$, $(x_{(1)}, x_{(2)}]$, $\cdots$, $(x_{(n-1)}, x_{(n)}]$, $(x_{(n)}, +\infty)$, respectively. In an independent sample O'_m from $G(x)$, let $r_1, \cdots, r_{n+1}$ be the numbers of elements in O'_m which fall into $I_1, \cdots, I_{n+1}$, respectively. Then $(r_1, \cdots, r_{n+1})$ is a discrete vector random variable where $r_1, \cdots, r_{n+1}$ are nonnegative integers satisfying

$$r_1 + \cdots + r_{n+1} = m. \tag{1.1}$$

Next we define a new vector random variable $(s_0, s_1, \cdots, s_m)$ where

$$s_i = \text{number of } (r_1, \cdots, r_{n+1}) \text{ which are equal to } i, \tag{1.2}$$

Research partially supported by the Office of Naval Research.

$i = 0, 1, \cdots, m$. The $s_0, s_1, \cdots, s_m$ will be called *cell frequency counts* which O'_m makes with respect to O_n. They can take only nonnegative integral values subject to the two conditions

$$\begin{aligned} s_0 + s_1 + \cdots + s_m &= n + 1, \\ s_1 + 2s_2 + \cdots + ms_m &= m. \end{aligned} \tag{1.3}$$

The components of the vector $(s_0, s_1, \cdots, s_m)$ are invariant under permutations of $r_1, \cdots, r_{n+1}$. There does not appear to be a more basic set of elements from which to construct a general nonparametric two-sample test than these components. For any fixed k, if we allow $m, n \to \infty$ so that $m/n \to \lambda > 0$, then $s_0, s_1, \cdots, s_k$ are asymptotically normal. The basic quadratic form, given in (4.6) and in an alternative and more convenient form in (4.8), involved in this asymptotic normal distribution takes a particularly simple form which provides a large-sample test for $(F, G) \in \mathcal{C}_0$, which is consistent against alternatives in a large subset $\mathcal{C}^*$ of $\mathcal{C} - \mathcal{C}_0$ to be defined more precisely later. It is this test which will be considered in this paper.

2. Distribution of cell frequency counts for (F, G) $\in \mathcal{C}_0$

If $(x_{(1)}, \cdots, x_{(n)})$ are the order statistics O_n let $(u_1, \cdots, u_n)$ denote the continuous vector random variable $(F[x_{(1)}], F[x_{(2)}] - F[x_{(1)}], \cdots, F[x_{(n)}] - F[x_{(n-1)}])$. The components of this vector have constant probability density equal to $n!$ over the simplex defined by $u_1 + \cdots + u_n \leqq 1$ and $u_i \geqq 0$, for $i = 1, \cdots, n$.

The conditional probability that the vector random variable $(r_1, \cdots, r_{n+1})$ has a specific value $(r'_1, \cdots, r'_{n+1})$, given $(u_1, \cdots, u_n)$, is

$$\frac{m!}{r'_1! \cdots r'_{n+1}!} u_1^{r_1'} \cdots u_n^{r_n'} (1 - u_1 - \cdots - u_n)^{r_{n+1}'}. \tag{2.1}$$

Multiplying this expression by the probability element of $(u_1, \cdots, u_n)$, namely $n!\, du_1 \cdots du_n$, and integrating over the simplex $u_1 + \cdots + u_n \leqq 1$, where $u_i \geqq 0$, for $i = 1, \cdots, n$, it is seen that the probability that $(r_1, \cdots, r_{n+1})$ takes on any specified value $(r'_1, \cdots, r'_{n+1})$ is given by

$$\frac{1}{\binom{m+n}{n}}, \tag{2.2}$$

and hence constant for every possible sample point in the space of $(r_1, \cdots, r_{n+1})$.

The problem of finding the probability that $(s_0, s_1, \cdots, s_m)$ has a particular value $(s'_0, s'_1, \cdots, s'_m)$ is one of counting all points in the sample space of $(r_1, \cdots, r_{n+1})$ for which $(s_0, s_1, \cdots, s_m) = (s'_0, s'_1, \cdots, s'_m)$, subject to conditions (1.3), and multiplying by $1/\binom{m+n}{n}$. It will be seen that the number of such points is the coefficient of $t_0^{s_0'} t_1^{s_1'} \cdots t_m^{s_m'} u^m$ in the formal expansion of

$$(t_0 + t_1 u + \cdots + t_m u^m)^{n+1}. \tag{2.3}$$

Extracting this coefficient and multiplying it by $1/\binom{m+n}{n}$, and dropping the dashes, we find the probability function $p(s_0, s_1, \cdots, s_m)$ of $(s_0, s_1, \cdots, s_m)$ to be given by

$$p(s_0, s_1, \cdots, s_m) = \frac{(n+1)!}{s_0!s_1! \cdots s_m! \binom{m+n}{n}}. \tag{2.4}$$

To obtain the probability function of only a fixed number of the s_i, say $(s_0, s_1, \cdots, s_k)$, with $k \leqq m$, we set $t_{k+1} = \cdots = t_m = t$ in (2.3) and select the coefficient of $t_0^{s_0}t_1^{s_1} \cdots t_k^{s_k}t^s u^m$, where

$$\begin{aligned} s_0 + s_1 + \cdots + s_k + s &= n + 1, \\ s &= s_{k+1} + \cdots + s_m. \end{aligned} \tag{2.5}$$

In particular, if $n + 1 \geqq m$, the probability function of s_0, the number of *empty cells*, is given by

$$p(s_0) = \frac{\binom{n+1}{s_0}\binom{m-1}{n-s_0}}{\binom{m+n}{n}}, \tag{2.6}$$

the sample space of s_0 being $n - m + 1, n - m + 2, \cdots, n + 1$.

3. Means, variances, and covariances of cell frequency counts

We shall need the means and covariance matrix of $(s_0, s_1, \cdots, s_k)$. For this purpose it will be convenient once and for all to evaluate the general factorial moment

$$E(s_0^{[g_0]}s_1^{[g_1]} \cdots s_k^{[g_k]}), \tag{3.1}$$

where

$$s_i^{[g_i]} = s_i(s_i - 1) \cdots (s_i - g_i + 1), \qquad s_i - g_i + 1 \geqq 0. \tag{3.2}$$

In view of the fact that the sum $\sum_s$ of $p(s_0, s_1, \cdots, s_m)$ over the sample space of $(s_0, s_1, \cdots, s_m)$ is unity, we see that

$$\sum_s \frac{1}{s_0!s_1! \cdots s_m!} = \frac{\binom{m+n}{n}}{(n+1)!}. \tag{3.3}$$

To determine the value of the expression in (3.1) we must evaluate the sum

$$\sum_s \frac{s_0^{[g_0]}s_1^{[g_1]} \cdots s_k^{[g_k]}}{s_0!s_1! \cdots s_m!}, \tag{3.4}$$

which is equal to the sum

$$\sum_{s'} \frac{1}{(s_0 - g_0)!(s_1 - g_1)! \cdots (s_k - g_k)!s_{k+1}! \cdots s_m!}, \tag{3.5}$$

where s' is the space of all nonnegative integral values of $s_0 - g_0, s_1 - g_1, \cdots, s_k - g_k, s_{k+1}, \cdots, s_m$ subject to the conditions

$$\begin{aligned}(s_0 - g_0) + (s_1 - g_1) + \cdots + (s_k - g_k) + s_{k+1} + \cdots + s_m \\ = n + 1 - g_0 - g_1 - \cdots - g_k \\ (s_1 - g_1) + 2(s_2 - g_2) + \cdots + k(s_k - g_k) + (k+1)s_{k+1} + \cdots + ms_m \\ = m - g_1 - 2g_2 - \cdots - kg_k,\end{aligned} \tag{3.6}$$

where $g_0 + g_1 + \cdots + g_k \leqq n + 1$ and $g_1 + 2g_2 + \cdots + kg_k \leqq m$.

It follows from (3.3) that the sum (3.5), subject to conditions (3.6), is given by the right side of (3.3) upon replacing n by $n - g_0 - g_1 - \cdots - g_k$ and m by $m - g_1 - 2g_2 - \cdots - kg_k$. This gives for the sum (3.5) the value

$$\frac{\binom{m + n - g_0 - 2g_1 - \cdots - (k+1)g_k}{n - g_0 - g_1 - \cdots - g_k}}{(n + 1 - g_0 - g_1 - \cdots - g_k)!}. \tag{3.7}$$

Multiplying (3.7) by $(n+1)!/\binom{m+n}{n}$, we obtain for the general factorial moment

$$E(s_0^{[g_0]} s_1^{[g_1]} \cdots s_k^{[g_k]}) = \frac{(n+1)!\binom{m + n - g_0 - 2g_1 - \cdots - (k+1)g_k}{n - g_0 - g_1 - \cdots - g_k}}{(n + 1 - g_0 - g_1 - \cdots - g_k)!\binom{m+n}{n}}. \tag{3.8}$$

We shall be particularly interested in the means and variances of $(s_0, s_1, \cdots, s_k)$ for large values of m and n such that $m = \lambda n + O(1)$, where $\lambda > 0$ and $(1/n)O(1)$ converges to zero as $n \to \infty$.

Under these conditions and putting

$$p_i = \frac{\lambda^i}{(1+\lambda)^{i+1}}, \qquad i = 0, 1, \cdots, k, \tag{3.9}$$

it is straightforward but tedious to verify that for $i = 0, 1, \cdots, k$,

$$\begin{aligned}E(s_i) &= np_i + O(1), \\ \text{Var}\,(s_i) &= np_i^2\left[\frac{1}{p_i} - 2 - \frac{i^2}{\lambda} + \frac{(i+1)^2}{1+\lambda}\right] + O(1),\end{aligned} \tag{3.10}$$

and for $i \neq j = 0, 1, \cdots, k$,

$$\text{Cov}\,(s_i, s_j) = np_ip_j\left[-2 - \frac{ij}{\lambda} + \frac{(i+1)(j+1)}{1+\lambda}\right] + O(1). \tag{3.11}$$

4. Asymptotic distribution of the cell frequency counts in large samples

If we put

$$Z_{in} = \frac{(s_i - np_i)}{\sqrt{n}}, \tag{4.1}$$

and, for a fixed k, let $\varphi_n(t_0, t_1, \cdots, t_k)$ be the characteristic function of $(Z_{0n}, Z_{1n}, \cdots, Z_{kn})$, that is,

$$\varphi_n(t_0, t_1, \cdots, t_k) = E[\exp(it_0Z_{0n} + it_1Z_{in} + \cdots + it_kZ_{kn})], \tag{4.2}$$

it can be verified by methods similar, for instance, to those used by Okamoto [9] and Kitabatake [4] that for $(F, G) \in \mathcal{C}_0$

$$\lim_{n\to\infty} \log \varphi_n = -\frac{1}{2}\sum_{i,j=0}^{k} \sigma_{ij}t_it_j, \tag{4.3}$$

where

$$\sigma_{ij} = \begin{cases} p_i^2\left[\frac{1}{p_i} - 2 - \frac{i^2}{\lambda} - \frac{(i+1)^2}{1+\lambda}\right], & i = j, \\ p_ip_j\left[-2 - \frac{ij}{\lambda} + \frac{(i+1)(j+1)}{1+\lambda}\right], & i \neq j. \end{cases} \tag{4.4}$$

But the right side of (4.3) is the logarithm of the characteristic function of a vector random variable $(w_0, w_1, \cdots, w_k)$ having a normal distribution with zero means and covariance matrix $||\sigma_{ij}||$. Thus, from Lévy's theorem [6] on the uniqueness of a limiting distribution as determined by the limit of a sequence of characteristic functions, we have

$$\lim_{n\to\infty} P(Z_{in} \leqq y_i; i = 0, 1, \cdots, k) = \frac{(|\sigma_{ij}|)^{1/2}}{(2\pi)^{k/2}} \int_{R_{k+1}} \exp\left(-\frac{1}{2}\sum \sigma^{ij}w_iw_j\right) dw_0 \cdots dw_k, \tag{4.5}$$

where R_{k+1} is the portion of the Euclidean $(k+1)$-space for which $w_i \leqq y_i$ with $i = 0, 1, \cdots, k$ and where $||\sigma^{ij}|| = ||\sigma_{ij}||^{-1}$.

Furthermore, as $m, n \to \infty$ with $m/n \to \lambda > 0$, the quadratic form

$$Q_{kn} = \sum_{i,j=0}^{k} \sigma^{ij}Z_{in}Z_{jn} \tag{4.6}$$

has, as its limiting distribution, the chi-square distribution with $k+1$ degrees of freedom if $(F, G) \in \mathcal{C}_0$.

Using the fact that Q_{kn} can be written as

$$|\sigma_{ij}|Q_{kn} = -\begin{vmatrix} 0 & Z_{0n} & Z_{1n} & \cdots & Z_{kn} \\ Z_{0n} & \sigma_{00} & \sigma_{01} & \cdots & \sigma_{0k} \\ Z_{1n} & \sigma_{10} & \sigma_{11} & \cdots & \sigma_{1k} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ Z_{kn} & \sigma_{k0} & \sigma_{k1} & \cdots & \sigma_{kk} \end{vmatrix} \tag{4.7}$$

and performing some simple operations on the determinant on the right it can be shown that Q_{kn} can be expressed in a more convenient and more easily computable form as

$$Q_{kn} = \sum_{i=0}^{k} \frac{(s_i - np_i)^2}{np_i} + \frac{u^2 + v^2}{n\lambda^2(1 + \lambda)p_k}, \tag{4.8}$$

where

$$u = \sum_{i=0}^{k} (s_i - np_i)(i - \lambda - k - 1),$$
$$v = [\lambda(1 + \lambda)]^{1/2} \sum_{i=0}^{k} (s_i - np_i). \tag{4.9}$$

Note that the form of the sum appearing on the right side of (4.8) is essentially that of the classical Pearson chi-square associated with the first $k + 1$ cell frequency counts $s_0, s_1, \cdots, s_k$. This sum must be augmented by the term $(u^2 + v^2)/[n\lambda^2(1 + \lambda)p_k]$ in order to produce a quantity having a limiting chi-square distribution with $k + 1$ degrees of freedom as $n \to \infty$.

5. Q_{kn} as a test for the hypothesis that $(F, G) \in \mathcal{C}_0$

As indicated in the preceding section, Q_{kn} has the chi-square distribution with $k + 1$ degrees of freedom as $n \to \infty$ under the null hypothesis that $(F, G) \in \mathcal{C}_0$, that is, if $F(x) \equiv G(x)$. As a matter of fact Q_{kn} provides a test for the null hypothesis which is consistent against a large class of alternative pairs (F, G) in $\mathcal{C} - \mathcal{C}_0$, which we shall call $\mathcal{C}^*$, where $\mathcal{C}^*$ is the subset of $\mathcal{C} - \mathcal{C}_0$ such that for any $(F, G) \in \mathcal{C}^*$ the function $G(F^{-1}(v)) = H(v)$, say, has a bounded derivative $h(v)$ on $[0, 1]$, $F^{-1}(v)$ being the inverse of $F(x)$. Then, since $\mathcal{C}^*$ is a subset of $\mathcal{C} - \mathcal{C}_0$, we have for $(F, G) \in \mathcal{C}^*$ that $H(v) \not\equiv v$, and hence $h(v) \not\equiv 1$, over a set of positive probability on $[0, 1]$.

To examine the consistency of Q_{kn} we shall use a method similar to that used on a related problem by Blum and Weiss [1]. Let T_α, with $\alpha = 1, \cdots, n + 1$ be a random variable where $T_\alpha = 1$ if i elements of the second sample O'_m lie on the interval I_α generated by the first sample O_n where $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$; $T_\alpha = 0$ otherwise.

Then if $(F, G) \in \mathcal{C}^*$, we have

$$E(s_i|\mathcal{C}^*) = E\left(\sum_{\alpha=1}^{n+1} T_\alpha|\mathcal{C}^*\right) = \sum_{\alpha=1}^{n+1} E(T_\alpha|\mathcal{C}^*). \tag{5.1}$$

Since $E(T_\alpha|\mathcal{C}^*) = P\{T_\alpha = 1|\mathcal{C}^*\}$, we have

$$E(s_i|\mathcal{C}^*) = \sum_{\alpha=1}^{n+1} P\{T_\alpha = 1|\mathcal{C}^*\}. \tag{5.2}$$

But for $\alpha = 2, \cdots, n$ we have

$$P\{T_\alpha = 1|\mathcal{C}^*\} = \frac{n!\binom{m}{i}}{(\alpha - 2)!(n - \alpha)!} \iint_T [1 - \{H(u_2) - H(u_1)\}]^{m-1} \tag{5.3}$$
$$[H(u_2) - H(u_1)]^i u_1^{\alpha-2}(1 - u_2)^{n-\alpha}\, du_1\, du_2,$$

where $u_1 = F(x_{(\alpha-1)})$, $u_2 = F(x_{(\alpha)})$, and T is the triangle in the u_1u_2-plane defined by $0 < u_1 < u_2 < 1$. For $\alpha = 1, n+1$ the formulas are slightly different, but for our purposes they need not be written down explicitly since $(1/n)P\{T_1 = 1|\mathcal{C}^*\}$ and $(1/n)P\{T_{n+1} = 1|\mathcal{C}^*\}$ both $\to 0$ as $n \to \infty$.

Substituting from (5.3) into the right side of (4.2) and summing from $\alpha = 2$ to $\alpha = n$, we obtain

$$E(s_i|\mathcal{C}^*) = \delta + n(n-1)\binom{m}{i}\iint_T [1 - \{H(u_2) - H(u_1)\}]^{m-i} [H(u_2) - H(u_1)]^i [1 - (u_2 - u_1)]^{n-2}\, du_1\, du_2, \tag{5.4}$$

where $\delta = P\{T_1 = 1|\mathcal{C}^*\} + P\{T_{n+1} = 1|\mathcal{C}^*\}$.

Performing the transformation $u_1 = v$ and $u_2 = v + y/n$, and considering the mean value of s_i/n we have

$$E\left(\frac{s_i}{n}\middle|\mathcal{C}^*\right) = \frac{\delta}{n} + \int_0^1 \int_0^{(1-v)n} f_{mn}(v, y)\, dv\, dy, \tag{5.5}$$

where

$$f_{mn}(v, y) = \frac{n(n-1)}{n^2}\binom{m}{i}\left(1 - \frac{y}{n}\right)^{n-2} \left[1 - \left\{\frac{H(v + y/n) - H(v)}{y/n}\right\}\frac{y}{n}\right]^{m-i} \left[\frac{H(v + y/n) - H(v)}{y/n}\right]^i \left(\frac{y}{n}\right)^i. \tag{5.6}$$

If we let $m, n \to \infty$ so that $m/n \to \lambda > 0$, and making use of the assumption that $H(v)$ has a bounded derivative $h(v)$ on $[0, 1]$ it follows that $\delta/n \to 0$, and

$$\int_0^1 \int_0^{(1-v)n} f_{mn}(v, y)\, dv\, dy \to \frac{1}{i!}\int_0^1 \int_0^\infty e^{-y[1+\lambda h(v)]} y^i [\lambda h(v)]^i\, dy\, dv = \int_0^1 \frac{[\lambda h(v)]^i\, dv}{[1 + \lambda h(v)]^{i+1}} = p_i(\lambda, h), \tag{5.7}$$

say.

Summarizing, we have the following result:

Let O_n and O'_m be samples of sizes n and m from $F(x)$ and $G(x)$, respectively. Then if $m, n \to \infty$ so that $m/n \to \lambda > 0$, we have for $(F, G) \in \mathcal{C}^*$

$$E\left(\frac{s_i}{n}\middle|\mathcal{C}^*\right) \to p_i(\lambda, h). \tag{5.8}$$

Note that if $(F, G) \in \mathcal{C}_0$, then $h(v) \equiv 1$ on $[0, 1]$ and we obtain the result

$$E\left(\frac{s_i}{n}\middle|\mathcal{C}_0\right) \to p_i \tag{5.9}$$

as $m, n \to \infty$ so that $m/n \to \lambda > 0$, as implied in (3.10).

For the case $i = 0$ it follows from the Schwarz inequality

$$\int_0^1 \frac{dv}{1+\lambda h(v)} \int_0^1 [1+\lambda h(v)]\, dv \geqq \left\{\int_0^1 \left[\frac{1}{1+\lambda h(v)}\right]^{1/2} [1+\lambda h(v)]^{1/2}\, dv\right\}^2 \tag{5.10}$$

that for any $\lambda > 0$,

$$\int_0^1 \frac{dv}{1+\lambda h(v)} \geqq \frac{1}{1+\lambda}, \tag{5.11}$$

that is,

$$p_0(\lambda, h) \geqq p_0 \tag{5.12}$$

with equality holding if and only if $h(v) \equiv 1$ on $[0, 1]$ except possibly for a set of probability 0. This means that as $m, n \to \infty$ with $m/n \to \lambda > 0$,

$$\lim_{n\to\infty} E\left(\frac{s_0}{n}\middle|\mathcal{C}^*\right) > \lim_{n\to\infty} E\left(\frac{s_0}{n}\middle|\mathcal{C}_0\right). \tag{5.13}$$

It can be shown by some tedious computations, which would require too much space here, that for the variance of s_0 under $\mathcal{C}^*$, we have

$$\operatorname{Var}\left(\frac{s_0}{n}\middle|\mathcal{C}^*\right) \leqq \frac{n+1}{4n^2}. \tag{5.14}$$

It follows from (5.13) and (5.14) that Q_{0n} is a test for $(F, G) \in \mathcal{C}_0$ which is consistent against all alternatives $(F, G) \in \mathcal{C}^*$.

It should be pointed out, however, that for no other value of i than $i = 0$ is it true for all $(F, G) \in \mathcal{C}^*$ and all $\lambda > 0$ that we have an inequality of form $p_i(\lambda, h) > p_i$ or $p_i(\lambda, h) < p_i$. This means that we cannot construct a test for $(F, G) \in \mathcal{C}_0$ which is consistent against all alternatives $(F, G) \in \mathcal{C}^*$ for all $\lambda > 0$ from any single s_i except s_0. In the case of s_0 it can be seen from the structure of Q_{0n} that for $(F, G) \in \mathcal{C}_0$, we have s_0 asymptotically normal $N[n/(1+\lambda), n\lambda^2/(1+\lambda)^3]$ for large n.

On the other hand it follows from (5.13) that for fixed k each of the tests Q_{in} with $i = 0, 1, \cdots, k$ is consistent for testing $(F, G) \in \mathcal{C}_0$ against alternatives $(F, G) \in \mathcal{C}^*$. It can be shown that for large n and for $i = 1, \cdots, k$ the power of Q_{in} for testing $(F, G) \in \mathcal{C}_0$ against alternatives $(F, G) \in \mathcal{C}^*$ is greater than that of $Q_{i-1,n}$.

6. Optimum choice of λ for members of $\mathcal{C}^*$ "near" those of $\mathcal{C}_0$

A thorough study of the power of the test Q_{kn} for the hypothesis described above would require the determination and careful examination of the asymptotic distribution of Q_{kn} for $(F, G) \in \mathcal{C}^*$ for large n. For any choice of (F, G) the power of the test depends on λ, the ratio of the size of the second sample to that of the first. It can be shown by methods similar to those used by Kitabatake [4] that for $(F, G) \in \mathcal{C}^*$ and for the fixed k, the $s_0, s_1, \cdots, s_k$ are asymptotically

jointly normal. We have shown that the value of $E(s_i|\mathcal{C}^*)$ to terms of order n is $np_i(\lambda, h)$, but the actual computation of the covariance matrix of $(s_0, s_1, \cdots, s_k)$ for $(F, G) \in \mathcal{C}^*$ to terms of order n is a tedious job which remains to be done.

Studies made by Mood [8] and others indicate that the Wald-Wolfowitz run test has low efficiency if used for testing hypotheses concerning differences of population means or variances. Such a study is likely to show the same to be true of tests based on cell frequency counts. On the other hand there may be some interest in considering such tests for testing the hypothesis that $(F, G) \in \mathcal{C}_0$ against alternatives $(F, G) \in \mathcal{C}^*$ "near" those in $\mathcal{C}_0$. We shall examine the power of Q_{kn} in some detail for this case.

More precisely, we shall compare approximate values of $p_i(\lambda, h)$ with p_i, where $i = 0, 1, \cdots, k$, as a function of λ assuming $h(v)$ to be of form $1 + \epsilon(u)$ where values of $\int_0^1 \epsilon(u)^r\, du$, for $r = 3, 4, \cdots$, are negligible compared with that for $r = 2$. Note that the value of the integral is zero for $r = 1$. If we denote by Δ^2 the value of the integral for $r = 2$ then for the degree of approximation indicated we have $p_i(\lambda, h) \doteq \bar{p}_i(\lambda, h)$, where

$$\bar{p}_i(\lambda, h) = p_i\left\{1 + \frac{\Delta^2}{2}\left[i(i-1) - 2i(i+1)t + (i+1)(i+2)t^2\right]\right\}, \qquad i = 0, 1, \cdots, k, \tag{6.1}$$

where $t = \lambda/(1 + \lambda)$, and p_i is defined in (3.9).

6.1. *Case of $k = 0$.* First consider the test Q_{0n}. This test, of course, is equivalent to using s_0 as a test, s_0 having the normal distribution $N[n/(1 + \lambda), n\lambda^2/(1 + \lambda)^3]$ as its asymptotic distribution for $(F, G) \in \mathcal{C}_0$. The value of λ which suggests itself as the optimum choice (the one to maximize the power of the Q_{0n} or s_0 test) for discriminating between members of $\mathcal{C}_0$ and of "nearby" members of $\mathcal{C}^*$ is that which makes the difference $\bar{p}_0(\lambda, h) - p_0$ as large as possible. Putting $i = 0$ in (6.1) we find

$$\bar{p}_0(\lambda, h) - p_0 = p_0 t^2 \Delta^2. \tag{6.2}$$

Noting that $p_0 t^2 = \lambda^2/(1 + \lambda)^3$, it is seen that the value of λ which maximizes the difference in (6.2) is $\lambda = 2$. This value of λ gives p_i the value $2^i/3^{i+1}$ and hence $n/3, 2n/9, 4n/27, 8n/81, \cdots$ as approximate mean values of cell frequency counts $s_0, s_1, s_2, s_4, \cdots$.

It should be pointed out that Q_{0k}, or equivalently s_0, is closely related to the Wald-Wolfowitz [13] run test. The number of runs u in the Wald-Wolfowitz test has $N[2n\lambda/(1 + \lambda), 4n\lambda^2(1 + \lambda)^3]$ as its limiting distribution as $m, n \to \infty$ with $m/n \to \lambda > 0$. Since the variance of u is four times that of s_0, to terms of order n, the value $\lambda = 2$ also maximizes the variance of u.

We further remark that s_0 may be regarded as a two-sample version of a one-sample test proposed by David [2] for testing the hypothesis that a sample of size m comes from a specified continuous distribution $F_0(x)$. In her problem $I_1, \cdots, I_{n+1}$ are disjoint intervals of the x-axis such that the probability on

each as computed from $F_0(x)$ is $1/(n+1)$ and her test is the number s_0^* of these intervals containing no elements of the sample. It was shown by Okamoto [9] that s_0^* is consistent for testing F_0 against any continuous distribution differing from F_0 and satisfying certain mild restrictions. Kitabatake [4] showed that s_0^* has a limiting normal distribution in samples from F_0 as well as from any alternative to F_0 satisfying some mild conditions.

6.2. *Case of $k \geqq 1$.* If one considers the problem of an optimum choice of λ for Q_{kn} with $k \geqq 1$, the situation is much more complicated. The choice $\lambda = 2$, which is optimum in the sense discussed earlier for $k = 0$, produces nonzero values of $\bar{p}_i(\lambda, h) - p_i$ except for $i = 1$ and 8 which indicates that for this choice of λ the cell frequency counts s_1 and s_8 would contribute virtually nothing to the power of Q_{kn} in large samples for discriminating between members of $\mathcal{C}_0$ and "nearby" members of $\mathcal{C}^*$. For the case $k = 1$ the prospect that s_1 would contribute almost nothing to the power of Q_{1n} is not very attractive!

One procedure which might suggest itself is to choose λ so as to maximize

$$\sum_{i=0}^{k} [\bar{p}_i(\lambda, h) - p_i]^2. \tag{6.3}$$

This solution, however, neglects the direction of the vector

$$\bar{p}_0(\lambda, h) - p_0, \quad \bar{p}_1(\lambda, h) - p_1, \quad \cdots, \quad \bar{p}_k(\lambda, h) - p_k, \tag{6.4}$$

which is an important matter on account of the dispersion among the eigenvalues of the covariance matrix $||\sigma_{ij}||$ defined in (4.4).

The problem of properly controlling the direction of the vector can be handled by selecting the value of λ which maximizes the quantity

$$Q_k^* = \sum_{i,j=0}^{k} \sigma^{ij}[\bar{p}_i(\lambda, h) - p_i][\bar{p}_j(\lambda, h) - p_j]. \tag{6.5}$$

The structure of Q_k^* is identical with that of Q_{kn} as given in (4.8) except that $(s_i - np_i)/\sqrt{n}$ is replaced by $[\bar{p}_i(\lambda, h) - p_i]$, which, as will be seen in (4.1), has the value $(\Delta^2/2)p_i f_i$, where

$$f_i = [i(i-1) - 2i(i+1)t + (i+1)(i+2)t^2] \tag{6.6}$$

and, as before, $t = \lambda/(1+\lambda)$.

Thus, we have

$$Q_k^* = \frac{\Delta^4}{4}\left[\sum_0^k p_i f_i^2 + \frac{\alpha^2 + \beta^2}{\lambda^2(1+\lambda)p_k}\right], \tag{6.7}$$

where

$$\begin{aligned} \alpha &= \sum_0^k p_i f_i(i - \lambda - k - 1), \\ \beta &= [\lambda(1+\lambda)]^{1/2} \sum_0^k p_i f_i. \end{aligned} \tag{6.8}$$

For the case $k = 0$, we find that $\lambda = 2$ maximizes Q_1^*, which is the same value

of λ which, as we pointed out earlier, maximizes $\bar{p}_i(\lambda, h) - p_i$ and also maximizes the variance of s_0.

For the case $k = 1$ the value of λ which maximizes $\bar{p}_i(\lambda, h) - p_i$ is 1.88 (to two decimals). The problem of determining the value of λ for $k \geqq 2$ requires a considerable amount of computation and this remains to be done. It is conjectured that these values of λ would lie in the interval $(2 \pm .1)$. In actual applications of Q_{kn} values of k not exceeding 2 or 3 would probably be sufficient for practical purposes.

REFERENCES

[1] J. R. BLUM and L. WEISS, "Consistency of certain two-sample tests," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 242–246.

[2] F. N. DAVID, "Two combinatorial tests of whether a sample has come from a given population," *Biometrika*, Vol. 37 (1950), pp. 97–110.

[3] W. J. DIXON, "A criterion for testing the hypothesis that two samples are from the same population," *Ann. Math. Statist.*, Vol. 11 (1940), pp. 199–204.

[4] S. KITABATAKE, "A remark on a non-parametric test," *Math. Japon.*, Vol. 5 (1958), pp. 45–49.

[5] E. L. LEHMANN, "Consistency and unbiasedness of certain non-parametric tests," *Ann. Math. Statist.*, Vol. 22 (1951), pp. 165–179.

[6] P. LÉVY, *Théorie de l'Addition des Variables Aléatoires*, Paris, Gauthier-Villars, 1937.

[7] H. B. MANN and D. R. WHITNEY, "On a test of whether one or two random variables is stochastically larger than the other," *Ann. Math. Statist.*, Vol. 18 (1947), pp. 50–60.

[8] A. M. MOOD, "On the asymptotic efficiency of certain non-parametric tests," *Ann. Math. Statist.*, Vol. 25 (1954), pp. 514–522.

[9] M. OKAMOTO, "On a non-parametric test," *Osaka Math. J.*, Vol. 4 (1952), pp. 77–85.

[10] I. R. SAVAGE, "Contributions to the theory of rank order statistics—the two-sample case," *Ann. Math. Statist.*, Vol. 27 (1956), pp. 590–615.

[11] N. V. SMIRNOV, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples," *Bull. Math. Univ. Moscou*, Vol. 2 (1939), pp. 3–14.

[12] B. V. SUKHATME, "On certain two-sample non-parametric tests for variances," *Ann. Math. Statist.*, Vol. 28 (1957), pp. 188–194.

[13] A. WALD and J. WOLFOWITZ, "On a test of whether two samples are from the same population," *Ann. Math. Statist.*, Vol. 11 (1940), pp. 147–162.

[14] F. WILCOXON, "Individual comparison by ranking methods," *Biometrics*, Vol. 1 (1945), pp. 80–83.