TESTING FOR AGREEMENT BETWEEN
TWO GROUPS OF JUDGES

by

Myles Hollander\textsuperscript{1} and Jayaram Sethuraman\textsuperscript{2}

AFOSR-N Technical Report No. 2
FSU Statistics Report M398
USARO-D Technical Report No. 21
AFOSR Technical Report No. 68

Revised: March, 1977
The Florida State University
Department of Statistics
Tallahassee, Florida

\textsuperscript{1}Research sponsored by the Air Force Office of Scientific Research, AFSC, USAF, under Grants AFOSR-74-2581B and AFOSR-76-3109.

\textsuperscript{2}Research sponsored by the United States Army Research Office, Durham, under Grant No. DAA29-76-G-0238.
TESTING FOR AGREEMENT BETWEEN TWO GROUPS OF JUDGES

BY MYLES HOLLANDER AND JAYARAM SETHURAMAN

Department of Statistics, Florida State University, Tallahassee

SUMMARY

The "problem of m rankings", so named by Kendall and studied extensively by Kendall and Babington Smith (1939), Kendall (1970), and others, considers the relationship between the rankings that a group of m judges assigns to a set of k objects. Suppose there are two groups of judges ranking the objects. Given that there is agreement within each group of judges, how can we test for evidence of agreement between the two groups? This question, recently posed to us by Kendall, has been studied by Schucany and Frawley (1973) and Li and Schucany (1975). In this paper we show that the test of agreement proposed by Schucany and Frawley, and further advanced by Li and Schucany, is misleading and does not provide a satisfactory answer to Kendall's question.

After pinpointing various defects of the Schucany-Frawley test, we adapt a procedure, proposed by Wald and Wolfowitz (1944) in a slightly different context, to furnish a new test for agreement between two groups of judges.

Some key words: Conditionally distribution-free; Permutation test; Rank correlation.
1. INTRODUCTION

Suppose that a judge is presented with \( k \) objects, say \( k \) science fair projects, and is asked to rank them. Then his ranking is a vector \( r_\cdot = (r_{1\cdot}, \ldots, r_{k\cdot}) \) chosen according to his probability distribution of rankings, \( Q \), on the space \( \Omega \) of \( k! \) possible rankings. When this probability distribution is the uniform probability distribution \( U \) (\( U \) assigns probability \( 1/k! \) to each ranking), we say that the judge has no opinion. Otherwise, we say that the judge has an opinion which is quantified by \( Q \).

Suppose that there are \( m \) like-minded male judges who rank the \( k \) objects independently, producing the rankings \( r_i = (r_{i1}, \ldots, r_{ik}) \), \( i = 1, \ldots, m \). That is, we assume \( r_1, \ldots, r_m \) are independent and identically distributed random vectors in \( \Omega \) with a common distribution \( Q_1 \), the opinion of the male judges. Next suppose that there is a second group of \( n \) like-minded female judges who rank the same \( k \) objects independently and produce the rankings \( r_i = (r_{i1}, \ldots, r_{ik}) \), \( i = m + 1, \ldots, N \), where \( N = m + n \). That is, we assume \( r_{m+1}, \ldots, r_N \) are independent and identically distributed random vectors in \( \Omega \) with a common distribution \( Q_2 \), the opinion of the female judges. How do we test that the male and female judges have a common opinion?

Sir Maurice Kendall posed this question to one of us during his visit to Tallahassee in the Spring of 1976. In our search of the literature, we discovered that Shucany and Frawley (1973) have proposed a test intended to solve this problem. The Shucany-Frawley (SF) test, further advanced by Li and Schucany (1975) and generalized by
Beckett (1975) and Beckett and Shucany (1975), is based on the statistic \( L \) defined by (1.2) below.

Let

\[
S_j = \sum_{i=1}^{m} r_{ij}, \quad T_j = \sum_{i=m+1}^{N} r_{ij}, \quad j = 1, \ldots, k. \tag{1.1}
\]

The SF statistic is

\[
L = \sum_{j=1}^{k} S_jT_j. \tag{1.2}
\]

It is easily seen that \( L \) is equivalent to the statistic \( \bar{\rho} \), the average value of all \( mn \) Spearman rank order correlations of a ranking from a male judge with a ranking from a female judge. More precisely,

\[
\bar{\rho} = \frac{12L - 3mn(k+1)^2}{mn(k^3-k)}, \tag{1.3}
\]

where

\[
\bar{\rho} = (mn)^{-1} \sum_{i=1}^{m} \sum_{i'=m+1}^{N} \rho_{ii'}, \tag{1.4}
\]

and

\[
\rho_{ii'} = 1 - \frac{\{6 \sum_{j=1}^{k} (r_{ij} - \bar{r}_{ij})^2\}/(k^3-k)}{\{6 \sum_{j=1}^{k} (r_{ij} - \bar{r}_{ij})^2\}/(k^3-k)}. \tag{1.5}
\]

Shucany and Frawley reasoned that large values of \( \bar{\rho} \), or equivalently large values of \( L \), should constitute evidence for the hypothesis \( H_{11} \) of two-group agreement. (\( H_{11} \) is defined precisely by (2.4) of Section 2.)

In Section 2 of this paper we show that the SF test is misleading, and does not constitute a satisfactory answer to Kendall's question.
The defects of the SF test include:

I. When \( m < n \), the statistic \( l \) gives too much weight to the rankings of male judges, and not enough weight to the rankings of female judges. When \( m > n \), the situation is reversed.

II. Critical values for the SF test are obtained by referring \( l \) to its distribution under an irrelevant (for the problem under discussion) hypothesis \( H_{00} \) of complete accordance within each group. The hypothesis \( H_{00} \) [see (2.1) of Section 2] specifies that \( Q_1 = Q_2 = U \).

III. In Section 2, equation (2.2) defines the alternative \( H_{01} \) which specifies that the male judges have no opinion (\( Q_1 = U \)) but the female judges have an opinion (\( Q_2 \neq U \)). The alternative \( H_{10} \) is defined by (2.3) analogously. Then, in Proposition 2 of Section 2, we prove that \( l \) has the same mean under \( H_{00} \) as it does under \( H_{01} \cup H_{10} \). This is used (Proposition 3) to show that the SF test is not consistent against \( H_{01} \cup H_{10} \), a hypothesis where the two groups of judges are governed by different distributions, one of which is uniform.

IV. Proposition 3 also shows that the SF test is not consistent against a large class of alternatives where the two groups of judges have different opinions, neither of which is uniform. That is, there are \((Q_1, Q_2)\) pairs in \( A_{11} \) (defined by (2.5) of Section 2) where \( Q_1 \neq Q_2 \), neither \( Q_1 \) nor \( Q_2 \) is uniform, but for which, even as \( m \) and \( n \) get arbitrarily large, the SF test leads to the decision that the two groups agree.
In Section 3 we show that we can apply a permutation test based on the Mahalanobis $D^2$ statistic, proposed in a different setting by Wald and Wolfowitz (1944), to obtain a conditionally distribution-free test for the hypothesis of agreement between the two groups of judges. A convenient large sample approximation is available, and the test is consistent for a large class of alternatives.

Section 4 contains an application of our conditional test, and the SF test, to a set of leisure activity preferences data provided by Sutton (1976).

2. THE SCHUCANY-FRAWLEY TEST

To understand the contents of the Schucany-Frawley (1972) paper, and the SF test advocated there and in the subsequent paper by Li and Schucany (1975), it is helpful to consider the following five subclasses of possible opinions $(Q_1, Q_2)$ for the two groups of judges. Thus, let

\[ H_{00} = \{ (Q_1, Q_2) : Q_1 = Q_2 = U \}, \quad (2.1) \]

\[ H_{01} = \{ (Q_1, Q_2) : Q_1 = U, Q_2 \neq U \}, \quad (2.2) \]

\[ H_{10} = \{ (Q_1, Q_2) : Q_1 \neq U, Q_2 = U \}, \quad (2.3) \]

\[ H_{11} = \{ (Q_1, Q_2) : Q_1 = Q_2, Q_1 \neq U, Q_2 \neq U \}, \quad (2.4) \]

and

\[ A_{11} = \{ (Q_1, Q_2) : Q_1 \neq Q_2, Q_1 \neq U, Q_2 \neq U \}. \quad (2.5) \]
The hypothesis of agreement between the two groups of judges corresponds to $H = H_{11} \cup H_{00}$. However, the hypothesis of agreement, given that each group of judges has an opinion, corresponds to $H_{11}$, and the hypothesis that the judges have no opinion (in Kendall's terminology, the hypothesis of complete accordance) corresponds to $H_{00}$.

Schucany and Frawley (1972) state that "... it is meaningless to make any comparison between groups unless each group 'has an opinion' i.e., there is concordance within each group." They then incongruously designate $H_{00}$ as the "null hypothesis." At the $\alpha$ level they propose to reject $H_{00}$ in favor of $H_{11}$ when $L \geq l_{00}$ where $l_{00}$ is determined by

$$P_{H_{00}} (L \geq l_{00}) = \alpha.$$  \hfill (2.6)

If $m$ and $n$ are large, the normal approximation to the distribution of $L$ under $H_{00}$ yields

$$l_{00} = E_{00}(L) + z_{\alpha}\{\text{var}_{00}(L)\}^{1/2},$$  \hfill (2.7)

where

$$E_{00}(L) = mnk(k + 1)^2/4,$$  \hfill (2.8)

and

$$\text{var}_{00}(L) = mn(k - 1)k^2(k + 1)^2/144,$$  \hfill (2.9)

are the mean and variance, respectively, of $L$ under $H_{00}$ and $z_{\alpha}$ is the upper $\alpha$ percentile point of the standard normal distribution.
There are many defects with the SF test. First, it is clear that 
$L$, or equivalently $\bar{\rho}$, is not a suitable test statistic when $m = n$. Consider, 
for example, an extreme case where $m = 1$ and $n = 10$. Then, as summarized 
by the $L$ statistic, or equivalently $\bar{\rho}$, a direct averaging of the 10 rank 
correlation coefficients gives too much weight to the rank vector of the 
male judge.

Secondly, the test is defined by Schucany and Frawley to discriminate 
between $H_{00}$ and $H_{11}$ when in fact they state it is meaningless to compare 
the groups unless each group has an opinion. The hypothesis $H_{00}$ asserts 
that each group does not have an opinion.

Thirdly, Proposition 3 shows that the SF test is not consistent against 
$H_{01} \cup H_{10}$, which consists of situations where the two groups of judges do not 
agree. (Propositions 1-2, which are used to prove Proposition 3, may be of 
some interest in their own right. For example, Proposition 1, pertaining to 
the case $m = n = 1$, shows that when one judge has no opinion, the distribution 
of a general class of statistics, including Spearman's rank order correlation 
coefficient, does not depend on the opinion of the other judge.)

Proposition 3 also shows that the SF test is further seriously 
weakened by the fact that it is not consistent against a large class of 
$(Q_1, Q_2)$ pairs in $A_{11}$.

PROPOSITION 1. Let $g(s_1, \ldots, s_k; t_1, \ldots, t_k)$ be a function of $2k$ 
arguments with an invariance property given by 

$$g(s_1, \ldots, s_k; t_1, \ldots, t_k) = g(s_{p_1}, \ldots, s_{p_k}; t_{p_1}, \ldots, t_{p_k}), \quad (2.10)$$

-7-
for each permutation \((p_1, \ldots, p_k)\) of \((1, \ldots, k)\). Then the statistic
\[
g(r_{11}, \ldots, r_{1k}; r_{m+1,1}, \ldots, r_{m+1,k})
\]
is distribution-free under \(H_{00} \cup H_{01} \cup H_{10}\).

Proof. Let \((Q_1, Q_2) \in H_{10}\). Then \(Q_2 = U\). Define the random permutation
\((p_1, \ldots, p_k)\), depending on \((r_{11}, \ldots, r_{1k})\) only, by
\[
r_{1p_j} = j, j = 1, \ldots, k.
\]

Using the invariance property (2.10) we have
\[
P_{Q_1,U}(g(r_{11}, \ldots, r_{1k}; r_{m+1,1}, \ldots, r_{m+1,k}) = g_0)
\]
\[
= P_{Q_1,U}(g(1, \ldots, k; r_{m+1,p_1}, \ldots, r_{m+1,p_k}) = g_0)
\]
\[
= P_U(g(1, \ldots, k; r_{m+1,1}, \ldots, r_{m+1,k}) = g_0).
\]
The last equality above follows since \((p_1, \ldots, p_k)\) is independent of
\((r_{m+1,1}, \ldots, r_{m+1,k})\) and the distribution of \((r_{m+1,1}, \ldots, r_{m+1,k})\) is
permutation invariant. This proves that the distribution of
\[
g(r_{11}, \ldots, r_{1k}; r_{m+1,1}, \ldots, r_{m+1,k})
\]
under \(H_{10}\) is the same as under \(H_{00}\). The same argument shows that the distribution of \(g(r_{11}, \ldots, r_{1k}; r_{m+1,1}, \ldots, r_{m+1,k})\) under \(H_{01}\) is the same as under \(H_{00}\). This completes
the proof.

PROPOSITION 2. The mean of \(L\) is distribution-free under \(H_{00} \cup H_{01} \cup H_{10}\).

Proof. The function
\[ g_1(s_1, \ldots, s_k; t_1, \ldots, t_k) = \sum_{j=1}^{k} s_j t_j \]

satisfies invariance property (2.10). Thus \( g_1 \) is distribution-free under \( H_{00} \cup H_{01} \cup H_{10} \). The proof is completed by noting that the statistic \( L \) is of the form

\[ L(r_1, \ldots, r_N) = \sum_{i=1}^{m} \sum_{j=m+1}^{N} g_1(r_{1i}, \ldots, r_{ki}; r_{j1}, \ldots, r_{jk}). \]

Thus \( E(L) = mnE(g_1) \), the value under \( H_{00} \cup H_{01} \cup H_{10} \) being given by the right-hand-side of (2.8).

Before proving Proposition 3, it is helpful to introduce some notation. Define the vector of mean rankings of the two groups of judges as follows:

\[ \mu = (\mu_1, \ldots, \mu_k), \quad \nu = (\nu_1, \ldots, \nu_k), \quad (2.13) \]

where

\[ \mu_j = E_{Q_1}(r_{j.}), \quad \nu_j = E_{Q_2}(r_{j.}), \quad j = 1, \ldots, k, \quad (2.14) \]

and \( E_{Q_1}, E_{Q_2} \) denote that the expectation is taken with respect to \( Q_1, Q_2 \) respectively. Thus

\[ E_{Q_1Q_2}(L) = mn \sum_{j=1}^{k} \mu_j \nu_j. \]

**PROPOSITION 3.** The SF test is not consistent against any \((Q_1, Q_2)\) pair for which

\[ \sum_{j=1}^{k} \mu_j \nu_j - \{k(k+1)/4\} \leq 0. \quad (2.15) \]

Proof. Consider first a \((Q_1, Q_2)\) pair such that the inequality in (2.15) is in fact an equality. (Proposition 2 shows all \((Q_1, Q_2)\) pairs in \( H_{01} \cup H_{10} \) have this property.) From the asymptotic theory of \( U \)-statistics, we have under \((Q_1, Q_2),\)
\[ (L - E_{Q_1,Q_2}(L))/\sqrt{\text{var}_{Q_1,Q_2}(L)} \xrightarrow{d} N(0,1). \]

Since \( E_{Q_1,Q_2}(L) = E_{00}(L) \), the asymptotic power at \((Q_1, Q_2)\), of the SF test of size \( \alpha \), is \( \leq 1 - \Phi(\theta \alpha) \), where \( \Phi \) is the standard normal cumulative distribution function and

\[ \theta^2 = \lim \inf \{ \text{var}_{00}(L)/\text{var}_{Q_1,Q_2}(L) \}. \]

Since \( \theta \geq 0 \), \( 1 - \Phi(\theta \alpha) \) is bounded above by \( \frac{1}{2} \).

Next, consider a \((Q_1, Q_2)\) pair such that the inequality in (2.15) is strict. (Of course, there are many such pairs in \( A_{11} \).) Then, under such a \((Q_1, Q_2)\), the statistic \( (L - E_{00}(L))/\sqrt{\text{var}_{00}(L)} \) will tend to \( -\infty \) and the SF test will not lead to the rejection of the SF "null hypothesis" \( H_{00} \) and thus the hypothesis of complete accordance will be (erroneously) accepted.
3. A CONDITIONALLY DISTRIBUTION-FREE TEST

The basic hypothesis testing problem of "agreement" versus "disagreement" between the two groups is, in terms of the hypotheses defined by (2.1) - (2.5), to discriminate between \( H = H_{00} \cup H_{11} \) versus \( A = H_{01} \cup H_{10} \cup A_{11} \). Since each judge's rank vector can assume only \( k! \) values, it appears at first glance that a test based on a multinomial distribution with \( k! \) cells could provide a solution to the testing problem. However, since \( k! \) is usually large, and many of the \( k! \) rankings will not occur in the data, such a test based on the multinomial would not be satisfactory.

We therefore modify the testing problem slightly by restricting the class of alternatives to those \((Q_1, Q_2)\) pairs whose vectors of mean ranks for the \( k \) objects are unequal. That is, in the notation of (2.13), we will test the hypothesis

\[
H = \{(Q_1, Q_2): Q_1 = Q_2\}, \quad \text{(3.1)}
\]

versus the alternative

\[
A^* = \{(Q_1, Q_2): \mu \neq \nu\}. \quad \text{(3.2)}
\]

We have thus reduced to problem to that of testing for the equality of the mean vectors in two multivariate populations. After this reduction, we can use a test suggested by Wald and Wolfowitz (1944) (in the context of testing for equality of two mean vectors) for our specific problem of two group agreement.
If the distributions $Q_1$, $Q_2$ were multivariate normal with the same covariance matrix, the appropriate test for equality of mean vectors would be the normal theory test based on the Mahalanobis $D^2$-distance between the two sample means. Clearly, here $Q_1$ and $Q_2$ are not multivariate normal, we thus use the Wald-Wolfowitz (1944) conditionally distribution-free test.

Notice that the covariance matrix of $(r_{11}, \ldots, r_{kk})$ under any distribution $Q$ on $\Omega$ will be singular, since $\sum_{i=1}^{k} r_{ij} = k(k+1)/2$. We will therefore omit the ranking of the $k^{th}$ object and use only the rankings of the first $(k-1)$ objects in computing the Mahalanobis distance. In this we tacitly assume that the covariance matrix of $(r_{11}, \ldots, r_{(k-1)(k-1)})$ under $Q$ is non-singular. Certain obvious modifications will have to be made if this covariance is singular.

Let

$$s_j = s_j/m, \quad t_j = t_j/n, \quad j = 1, \ldots, k - 1, \quad (3.3)$$

and let

$$c_{jj'} = \sum_{i=1}^{N} (r_{ij} - \bar{r}_j) (r_{ij'} - \bar{r}_{j'})/(N - 1), \quad 1 \leq j, j' \leq k - 1, \quad (3.4)$$

where $\bar{r}_j = \sum_{i=1}^{N} r_{ij}/N, \quad j = 1, \ldots, k - 1$. Setting $s = (s_1, \ldots, s_{k-1})$, $t = (t_1, \ldots, t_{k-1})$, and $C$ to be the $(k - 1) \times (k - 1)$ matrix of the $c_{jj'}$'s, our proposed test rejects $H$ in favor of $A^*$ if

$$B(r_1, \ldots, r_N) = mnN^{-1}(s - t)C^{-1}(s - t)' \quad (3.5)$$
is large. The statistic $B$ is monotonically related to the Mahalanobis $D^2$-statistic. Since $B$ is not distribution-free under $H$, we turn to a permutation test which is formally described as follows.

Consider the group $\Pi$ of permutation transformations $\pi$ that apply to a vector of $N$ rank vectors $(\omega_1, \ldots, \omega_N)$, $\omega_i = (\omega_{i1}, \ldots, \omega_{ik}) \in \Omega$, $i = 1, \ldots, N$, as follows:

$$\pi(\omega_1, \ldots, \omega_N) = (\omega_{\pi_1}, \ldots, \omega_{\pi_N}),$$

where $\pi = (\pi_1, \ldots, \pi_N)$ is a permutation of $(1, 2, \ldots, N)$. There are $N!$ transformations in $\Pi$. Let $\Pi(\omega_1, \ldots, \omega_N)$ denote the orbit of $(\omega_1, \ldots, \omega_N)$, that is

$$\Pi(\omega_1, \ldots, \omega_N) = \{\pi(\omega_1, \ldots, \omega_N) : \pi \in \Pi\}.$$

Under $H$,

$$P\{r_1, \ldots, r_N = (r_1^*, \ldots, r_N^*)| (r_1, \ldots, r_N) \in \Pi(\omega_1, \ldots, \omega_N)\} = \begin{cases} (N!)^{-1} & \text{if } (r_1^*, \ldots, r_N^*) \in \Pi(\omega_1, \ldots, \omega_N) \\ 0 & \text{otherwise}. \end{cases}$$

Thus the conditional distribution of $(r_1, \ldots, r_N)$ given that it belongs to the orbit of $(\omega_1, \ldots, \omega_N)$ is distribution-free under $H$. This conditional distribution is called the permutation distribution.

Define the critical value $B_0(\omega_1, \ldots, \omega_N)$ by the equation

$$P_H(\text{B}(r_1, \ldots, r_N) \geq B_0(\omega_1, \ldots, \omega_N)| (r_1, \ldots, r_N) \in \Pi(\omega_1, \ldots, \omega_N)) = \alpha.$$
Our \( \alpha \) - level permutation test rejects \( H \) if

\[
B(r_1, \ldots, r_N) \geq B^*_0(r_1, \ldots, r_N).
\]  

(3.6)

Since the statistic \( B \) is invariant under the \( m! \) permutations of the male rank vectors among themselves and invariant under the \( n! \) permutations of the female rank vectors among themselves, our proposed test requires the calculation of \( B \), not for each \( \pi \in \Pi \), but only for the \( M = \binom{N}{m} \) \( \pi \)'s corresponding to the possible choices of \( m \) rank vectors to serve as the male rank vectors. Thus when \( R = (r_1, \ldots, r_N) \) is observed, let 
\[ b_1(R) \leq \ldots \leq b_M(R) \] 
denote the ordered values of \( B(\pi(R)) \) for these \( M \) transformations. When \( \alpha = d/M \), our test rejects \( H \) in favor of \( A^* \) if \( B(R) \) is one of the \( d \) largest \( b \) values.

Even though \( C^{-1} \), appearing in (3.5), is unchanged by permutations, the computations of the \( \binom{N}{m} \) values of \( B \), when \( m \) and \( n \) are large, are formidable. In such cases, the following chi-square approximation can be used.

Wald and Wolfowitz (1944) have shown that under \( H \) the permutation distribution of \( B \) has a limiting chi-square distribution with \( k-1 \) degrees of freedom, assuming that the covariance matrix of \( (r_1, \ldots, r_{(k-1)}) \) under \( Q_1(=Q_2) \) is non-singular. Thus the large sample approximation to the \( \alpha \) level test defined by (3.6) is reject \( H \) if

\[
B \geq \chi^2_{\alpha, k-1}
\]  

(3.7)
where $\chi^2_{\alpha, k-1}$ is the upper $\alpha$ percentile point of a chi-square distribution with $k-1$ degrees of freedom.

Consistency of the permutation test is established as follows. When $(Q_1, Q_2) \in A^*$ and the covariance matrices of $(\tilde{r}, \ldots, \tilde{r}_{(k-1)})$ under $Q_1$ and $Q_2$ are non-singular, the results of Wald and Wolfowitz (1944) show that in the permutation distribution, $\sqrt{N} (s - t)$ has a limiting multivariate normal distribution with a mean vector $\mu - \nu$ [where here $\mu, \nu$ are the corresponding $k-1$ dimensional versions of (2.13)] which is non-zero. Thus $B$ tends to $\infty$ and the permutation test based on $B$ is consistent for all such alternatives in $A^*$.

4. AN EXAMPLE

Sutton (1976) has studied leisure preferences, and attitudes on retirement, of the elderly with the aim of providing leisure programs that meet the needs and goals of those participating. She cites evidence that, in the United States, existing senior programs seem to be geared to fitting clients to activities rather than planning activities with the individual's needs and goals in mind. In a sample of elderly retirees residing in Leon County, Florida, Sutton asked a number of questions designed to determine preferences for selected "activity components." Activity components are elements within activities such as where the activity takes, with whom the activity is done, and the type of leadership preferred during the activity. The data in Table 1 are the responses of $m = 14$ white females and $n = 13$ black females, in the age group 70-79 years, to the question: With which sex do you
prefer to spend your leisure? Each female was asked to rank the
three responses: male(s), female(s), both sexes, scoring 1 for the
most desired or first choice and 3 for the least desired or third
choice.

Table 1. Preferred companions for leisure time
activities of elderly females
(data of C. Sutton)

<table>
<thead>
<tr>
<th></th>
<th>male(s)</th>
<th>female(s)</th>
<th>both sexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>White</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

S's: 41 20 23
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Black Females**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**T's:** 30  32  16

For these data there is evidence that the white females have an opinion and that the black females have an opinion. Friedman's (1937) $\chi^2$ statistic (which, except for constants, is equivalent to the Kendall and Babington Smith coefficient W) for white females is 18.4. Referring this value to the chi-square distribution with two degrees of freedom yields a $P$ value less than .001 for the hypothesis of accordance among white females. The corresponding values for the black females are $\chi^2 = 11.7, P = .003$. 
We now use the conditionally distribution-free test to see whether the white females and black females agree. From Table 1, (3.3), and (3.4) we obtain

\[
(s_1, s_2) = (2.929, 1.429), (t_1, t_2) = (2.308, 2.462),
\]

\[
\mathbf{c} = \begin{bmatrix} .3960 & -.2593 \\ -.2593 & .5328 \end{bmatrix}
\]

\[
\mathbf{c}^{-1} = \begin{bmatrix} 3.706 & 1.804 \\ 1.804 & 2.755 \end{bmatrix}
\]

and from (3.5),

\[
B = 13.8.
\]

There are \(\binom{27}{14} = 20,058,300\) possible ways to pick 14 of the 27 rank vectors to serve as the rank vectors corresponding to the white females. Of these, only 4178 choices yield B values that are greater than or equal to the observed value of B = 13.8. Thus the exact P value for the conditional test is 4178/20,058,300 = .0002. This constitutes very strong evidence that the white female retirees have a different opinion than the black female retirees. The same conclusion is reached using the chi-square approximation, to the conditional distribution of B, given in Section 3. Referring B = 13.8 to the chi-square distribution with two degrees of freedom yields an approximate P value of .001.

Quite the opposite erroneous conclusion is reached by referring \(L\) to its \(\text{H}_0\) distribution as recommended by Schucany and Frawley (1973). We find, from (1.2), (2.8), and (2.9),
\[
\frac{L - E_{00}(L)}{\text{var}_{00}(L)} = \frac{2238 - 2184}{364} = 2.83.
\]

The Schucany-Frawley normal deviate of 2.83 gives the incorrect impression that the observed value of \( L \) is extremely large, and according to the SF test, this "large" value leads to the acceptance of \( H_{11} \).

The authors are grateful to Sir Maurice G. Kendall for suggesting the problem and for a useful discussion. We also thank John Kitchin, Edward Cooley, and Shih-Chuan Cheng for the programming that yielded the exact \( P \) value for the conditional test applied to the data of Table 1, Catherine Dabadie for calling our attention to the leisure activities preference study of Cindy Sutton, and Cindy Sutton for permitting us to use her data. This research was sponsored by the U.S. Air Force Office of Scientific Research and the U.S. Army Research Office.

REFERENCES


<table>
<thead>
<tr>
<th>1. REPORT NUMBERS</th>
<th>2. GOVT. ACCESSION NO.</th>
<th>3. RECIPIENT'S CATALOG NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>USARO-DAA29-76-G-0238 No. 21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFOSR-76-3109 No. 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFOSR-74-2581B No. 68</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. TITLE
TESTING FOR AGREEMENT
BETWEEN TWO GROUPS OF JUDGES

5. TYPE OF REPORT
Technical Report

6. PERFORMING ORGANIZATION REPORT NO.
FSU Statistical Report No. M398

7. AUTHOR(s)
Myles Hollander and Jayaram Sethuraman

8. CONTRACT OR GRANT NUMBER(s)
USARO-DAA29-76-G-0238 No. 21
AFOSR-76-3109 No. 2
AFOSR-74-2581B No. 68

9. PERFORMING ORGANIZATION NAME & ADDRESS
The Florida State University
Department of Statistics
Tallahassee, Florida 32306

10. PROGRAM ELEMENT, PROJECT, TASK
AREA AND WORK UNIT NOS.

11. CONTROLLING OFFICE(s) NAME(s) & ADDRESS(es)
Air Force Office of Scientific Research
Bolling Air Force Base, D.C. 20332
U.S. Army Research Office-Durham
P. O. Box 12211
Research Triangle Park, N.C. 27709

12. REPORT DATE
January, 1977

13. NUMBER OF PAGES
19

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release: distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS
Conditionally distribution-free; Permutation test; Rank correlation.
20. ABSTRACT

The "problem of m rankings", so named by Kendall and studied extensively by Kendall and Babington Smith (1939), Kendall (1970), and others, considers the relationship between the rankings that a group of m judges assigns to a set of k objects. Suppose there are two groups of judges ranking the objects. Given that there is agreement within each group of judges, how can we test for evidence of agreement between the two groups? This question, recently posed to us by Kendall, has been studied by Schucany and Frawley (1973) and Li and Schucany (1975). In this paper we show that the test of agreement proposed by Schucany and Frawley, and further advanced by Li and Schucany, is misleading and does not provide a satisfactory answer to Kendall's question. After pinpointing various defects of the Schucany-Frawley test, we adapt a procedure, proposed by Wald and Wolfowitz (1944) in a slightly different context, to furnish a new test for agreement between two groups of judges.