A TWO-SAMPLE NONPARAMETRIC TEST BASED ON SPACING-FREQUENCIES

J. S. RAO

University of California, Santa Barbara

V. K. MURTHY

Aerospace Corporation, El Segundo, California

ABSTRACT

For the usual two-sample problem, let $\{\nu_i\}$ denote the numbers of observations of say, the second sample falling in between the spacings formed by the first sample. Asymptotic distribution theory and efficiencies of tests based on these so called "spacing-frequencies" have been studied in Holst and Rao (1980). In particular they show that among test statistics based symmetrically on $\{\nu_i\}$, the statistic corresponding to the sum of squares of $\{\nu_i\}$, suggested by Dixon (1940), asymptotically is locally most powerful. It is also shown there that tests based on symmetric functions of $\{\nu_i\}$ are inefficient compared to those that are not symmetric in these frequencies. These considerations suggest a natural extension of the Dixon statistic, namely $(\sum \nu_i^2 + \sum \omega_j^2)$ where $\{\omega_j\}$ are the frequencies of the first sample in between the gaps made by the second sample. This statistic is shown to have better power performance compared to Dixon's test by Monte Carlo methods. More complete results of a theoretical nature will be presented elsewhere.

I. INTRODUCTION

Let $X_1, \ldots, X_{m-1}$ and $Y_1, \ldots, Y_{n-1}$ be independent random samples from two populations with continuous distribution functions (d. f. 's) $F(x)$ and $G(y)$ respectively. The usual two sample problem is to test if these two populations are identical ie., $F=G$. Since the spacing-frequencies, as well as statistics based on them remain invariant under probability integral transformations, we may without loss of generality make such a transformation on both samples. This permits us to assume that the support of both the samples i the interval $[0, 1]$ and the first of these

d. f. 's namely F(x) is the d. f of uniform distribution on [0, 1]. The d. f of the second sample is $G^* = G \circ F^{-1}$ and the null hypothesis is to test if

(1.1) $$H_0: G^*(y) = y, \quad 0 \le y \le 1.$$

Let $0 = X'_0 \le X'_1 \le \cdots \le X'_{m-1} \le X'_m = 1$ be the order statistics from the first sample. The sample spacings of X are defined by

(1.2) $$D_k = X'_k - X'_{k-1}, \quad k = 1, \cdots, m,$$

and the spacing-frequencies of Y by

(1.3) ν_k = number of Y_j's in the interval $[X'_{k-1}, X'_k]$ $k = 1, \cdots m$.

Statistics of the symmetric type $\Sigma h(\nu_k)$ and more general ones of the type $\Sigma h_k(\nu_k)$ have been studied in Holst and Rao (1980). It is shown that the test based on

(1.4) $$T_1 = \sum_{k=1}^{m} \nu_k^2$$

suggested by Dixon (1940) has the maximum asymptotic relative efficiency (ARE) amongst all the symmetric statistics of the form $\Sigma h(\nu_k)$, for testing the null hypothesis (1.1) against a close sequence of alternatives of the form

(1.5) $$G_n^*(y) = y + L(y)/n^{\delta}, \quad 0 \le y \le 1,$$

with $\delta = \frac{1}{4}$. It is however an unfortunate fact that any test symmetric in $\{\nu_1, \cdots, \nu_m\}$ can not distinguish alternatives of the form (1.5) when $\delta > \frac{1}{4}$, while several standard two-sample tests, like for instance the Kolmogorov-Smirnov statistic, do have non-zero power against alternatives (1.5) with $\delta = \frac{1}{2}$.

The aim of this note is to present a more powerful but simple alternative to (1.4). In order to do this, we introduce the conjugate or dual spacing frequencies $\{\omega_j\}$ defined as follows. Let

$0 = Y'_0 \le Y'_1 \le \cdots \le Y'_{n-1} \le Y'_n = 1$ be the order statistics from the second sample. Define the sample spacings of Y as

(1.6) $$E_k = Y'_k - Y'_{k-1}, \quad k = 1, \cdots n,$$

and the dual frequencies

(1.7) ω_k = number of X_i's in the interval $[Y'_{k-1}, Y'_k]$ $k = 1, \cdots n$

The proposed statistic for testing H_0 in (1.1) is

(1.8) $$T_2 = \sum_{k=1}^{m} \nu_k^2 + \sum_{k=1}^{n} \omega_k^2,$$

which is a natural extension of the statistic in (1.4).

II. SOME DISTRIBUTION THEORY AND MONTE CARLO POWER COMPARISONS

Denoting by $R(.)$ the rank of the observation in the combined sample, it is easily observed that $R(X_i) = \sum_{k=1}^{i} \nu_k + i$, $i = 1, \cdots, m-1$ while similarly $R(Y_j) = \sum_{k=1}^{j} \omega_k + j$, $j = 1, \cdots n-1$. Thus it is clear that the vector $\underline{\nu} = (\nu_1, \cdots \nu_m)$ somewhat indirectly determines the dual vector $\underline{\omega} = (\omega_1, \cdots \omega_n)$ and vice-versa. Indeed the statistic (1.8) can be written as

(2.1) $$\sum_{k=1}^{m} \nu_k^2 + \sum_{r=0}^{m} r^2 \cdot N_r ,$$

where N_r = { number of j: ω_j = r }. . to explore this second ter
further, define the sum of (r-1) of the ν_k's starting from the inde

(2.2) $$S_i^{(r-1)} = \nu_i + \nu_{i+1} + - - - + \nu_{i+(r-2)}, \quad i = 1, - - -,$$

where for convenience, we take $\nu_i = \nu_{i-m}$ for i > m, circularly
$\{ \omega_j \geq r \}$ if and only if $\{ S_j^{(r-1)} = 0 \}$, we have

(2.3) $$\begin{aligned} N_r &= \sum_{j=1}^{m} I\left(S_j^{(r-1)} = 0,\ S_j^{(r)} > 0,\ S_{j-1}^{(r)} > 0\right) \\ &= \sum_{j=1}^{m} I\left(\nu_{j-1} = 0,\ \nu_j + \nu_{j+1} + - - - + \nu_{j+ \ -2} = 0, \nu_{j+}\right. \end{aligned}$$

where I(.) is the indicator function of the event in the parenthesis.
(2.1) and (2.3), the statistic proposed in (1.8) can be expressed in
of $\{\nu_k\}$ alone as follows

(2.4) $$T_2 = \frac{\sum_{k=1}^{m} \nu_k^2 + \sum_{r=1}^{m} r^2 \sum_{j=1}^{m} I\left(\nu_{j-1} = 0,\ S_j^{(r-1)} = 0,\ \nu_{j+(r-1)} > 0\right)}{\sum_{k=1}^{m} \nu_k^2 + \sum_{j=1}^{m} \sum_{r=1}^{m} r^2 \ I\left(\nu_{j-1} = 0,\ S_j^{(r-1)} = 0,\ \nu_{j+(r-1)} > 0\right)} .$$

From this alternate form (2.4) for the proposed statistic T, it is
that it is not symmetric in $(\nu_1, - - -, \nu_m)$ and hence it is possible to
superior ARE compared to T_1 (cf. Holst and Rao (1980)).

A brief outline of the distribution theory follows: Under the nul
hypothesis (1.1), the vector $\underset{\sim}{\nu} = (\nu_1, - - -, \nu_m)$ has the same probabili
distribution as the occupancy numbers in the indistinguishable ball
problem where (m-1) balls are distributed among n cells. Thus

(2.5) $$P\left(\underset{\sim}{V} = \underset{\sim}{\nu}\right) = 1 \Big/ \binom{n + m - 2}{m - 1},$$

see for instance the discussion on Bose-Einstein statistics, Feller
(1968 p. 40), see also Holst (1979), example 2 on p. 552. It may b
easily verified that

(2.6) $$P\left(\underset{\sim}{V} = \underset{\sim}{\nu}\right) = P\left(\underset{\sim}{\eta} = \underset{\sim}{\nu} \ \Big| \ \sum_{1}^{m} \eta_i = (\eta - 1)\right),$$

where $\underset{\sim}{\eta} = (\eta_1, ---, \eta_m)$ are independent, identically distributed
geometric random variables with pdf $p(\eta_i = k) = pq^k$, k = 0,

where $p = \frac{m-1}{m+n-2}$. This representation (2.6) of the spacing freque
$\underset{\sim}{\nu}$ in terms of independent geometric random variables is valuable
one can then derive the distribution of any statistic $T = f(\nu_1, ---,$

as the conditional distribution of $T^* = f(\eta_1, ---, \eta_m)$ conditioned
event $\sum_{1}^{m} \eta_i = (n-1)$. Indeed an application of Theorem 2 Holst (197

an appropriate central limit theorem yield the asymptotic null distr

of T_2 as given in (2.4). The details are quite messy and will be p
ted elsewhere.

The superior power performance of T_2 over T_1 is shown by the
lowing simulated powers as given in Table 2.1, Instead of using th
asymptotic distributions, we find the upper 5% and 10% values of th
distributions of T_1 and T_2, using 200 samples. The values of m
are allowed to be 10, 20, 50, 100 and 200. The powers are then s
ted against the alternative

$$G_1^*(y) = y^2 + \quad 0 \le y \le 1 \tag{2.7}$$

It appears that as both samples go to infinity at the same rate T
T_2 appear to have about the same power. Also, whenever $m<<n$
statistics seem to be equally powerful. However, when $m>>n$ th
symmetric test based on T_2 has greater power than the symmetric
based on T_1.

We thank Dr. M.R. Chernick for his help in computations and h
valuable criticism.

REFERENCES

Dixon, W.J. (1940) A criterion for testing the hypothesis that tw
samples are from the same population, Ann. Math. Statistics 11, 1

Feller, W. (1968) An introduction to probability theory and its appl
tions, Vol. I, (3rd Ed.) John Wiley & Sons

Holst, L. (1979) Two conditional limit theorems with applications,
Statistic. 7, 551-557.

Holst, L. and Ross, J.S. (1980) Asymptotic theory for some famil
two-sample nonparametric statistics, Sankhya 42, Sec. A., 1-28

TABLE 2.1

m, n	5%		G Power		10%		G Powe
	T_1	T_2	T_1	T_2	T_1	T_2	T_1
(10, 10)	39	66	.055	.11	31	60	.145
(10, 20)	129	152	.125	.115	105	132	.325
(10, 50)	745	762	.13	.13	667	682	.215
(10, 100)	2639	2650	.27	.27	2451	2464	.33
(20, 10)	23	140	.120	.135	21	128	.20
(20, 20)	73	142	.190	.195	65	128	.290
(20, 50)	385	436	.210	.195	339	374	.39
(20, 100)	1441	1470	.305	.32	1239	1268	.52
(20, 200)	5263	5286	.32	.32	4865	4886	.41
(50, 10)	15	752	.135	.120	15	658	.135
(50, 20)	45	416	.125	.27	43	378	.175
(50, 50)	173	366	.37	.325	165	330	.49
(100, 10)	15	2856	.025	.11	13	2660	.115
(100, 20)	33	1420	.075	.295	31	1326	.185
(200, 20)	27	5182	.095	.320	25	4904	.170
(200, 200)							