RECENT DEVELOPMENTS IN STATISTICAL INFERENCE AND DATA ANALYSIS

K. Matusita Editor

NORTH-HOLLAND

RECENT DEVELOPMENTS IN STATISTICAL

Recent Developments in Statistical Inference and Data Analysis; K. Matusita, editor
© North-Holland Publishing Company, 1980

t. 1 (1973)1-23.
of the F-
PBIB design
t. Math.,Tokyo
l Volume of
ock designs
signs (Marcel
!, Submitted
An illustration
l experiments
ce, Jour. Mini-
a given area
gric. Sci. 4
f field plots,
esign with few
n Paper 417,
s, Biometrika
wheat nursery
rial experi-
d arrangements
trika 30 (1939)

PITMAN EFFICIENCIES OF SOME TWO-SAMPLE NONPARAMETRIC TESTS

J.S. Rao

University of California
Santa Barbara, USA

K.V. Mardia

Univeristy of Leeds
United Kingdom

This paper considers a group of two-sample nonparametric tests studied by Holst and Rao (1980) with special reference to their use in circular data analysis. The tests are based on the numbers of observations of one sample that fall in between the circular gaps formed by the other sample. Test statistics which are symmetric in these numbers have the property of invariance under changes in zero-direction and include the circular run test and Dixon's test. Relative efficiencies of various tests of this type have been evaluated and it is seen that the Dixon test has maximum asymptotic efficiency among these tests.

1. Introduction

In this paper, a group of two-sample nonparametric tests studied in Holst and Rao (1980) for the line, are considered with special reference to their use in circular data analysis. Let $X_1,\ldots X_m$ and $Y_1,\ldots Y_n$ be independent random samples from two continuous circular distribution functions F and G respectively, measured with respect to some arbitrary zero direction. The problem of interest is to test the hypothesis that these two parent populations are identical. For the purposes of this discussion, there is no loss of generality in assuming that the circle is of unit circumference and that the first sample is from the uniform distribution on the circle. This can be arranged through a probability integral transformation on the data, which does not alter the numbers $\{S_k\}$ defined in (1.2) or the statistics in (1.3). Thus from now on $F(y) = y,\ 0 \le y < 1$ and the null hypothesis to be tested is

$$H_0 : G(y) = y,\ 0 \le y < 1 . \tag{1.1}$$

Let $0 \le X_1' \le X_2' \le \ldots \le X_m' < 1$ be the order statistics from the first sample with respect to the given zero direction. The circular spacings corresponding the x-values are defined by

$$D_k = X_k' - X_{k-1}' ,\ k = 1,\ldots,m$$

where we put $X_0' = (X_m' - 1)$. Tests for the one sample problem of goodness of fit based on spacings are discussed for instance in Sethuraman and Rao (1970) and Rao (1976). See also Mardia (1972, pp. 171-172, 187-190) for a discussion of circular spacings. Define

$$S_k = \text{number of } Y_j\text{'s in } [X'_{k-1}, X'_k),\ k = 2,\ldots m$$

(1.2)
$$\text{and } S_1 = n - \sum_2^m S_k .$$

These numbers $\{S_k\}$ will be referred to as the "spacing-frequencies". For various choices of a function $h(\cdot)$, test statistics of the type

(1.3)
$$T_{m,n} = \sum_{k=1}^{m} h(S_k)$$

are proposed here for testing the hypothesis H_0. These statistics based symmetrically on $\{S_k\}$ remain invariant under changes in zero-direction and hence are especially useful in circular data problems. The well known run test on the circle (cf. David and Barton (1962)) and the statistic suggested by Dixon (1940) are included in this group. The asymptotic theory and Pitman's asymptotic relative efficiencies (ARE's) of statistics of the type (1.3) is discussed here when

(1.4)
$$m,n \to \infty \text{ and } m/(m+n) \to \lambda ,\ 0 < \lambda < 1 .$$

Numerical evaluations of the relative efficiencies of various test statistics in this class are carried out and are summarized in Tables 2.1, 2.2, and 2.3.

More generally for the circular problems, one could consider statistics of the type $T = T(S_1,\ldots,S_m)$ where T is a rotationally invariant function. But the theory is considerably more complex and (1.3) is an important special case of this. It may also be pointed out that a more general class of statistics considered in Holst and Rao (1980) viz. statistics of the type $\sum_{k=1}^{m} h_k(S_k)$ where possibly different functions $\{h_k(\cdot), k = 1,\ldots m\}$ are used, are not appropriate for the circle because they lack the rotational invariance property - even though they have higher asymptotic efficiencies compared to symmetric tests. See Holst and Rao (1980) for a detailed discussion of the asymptotic theory as it applies to the line. See also Govindarajulu and Sen (1966) and Govindarajulu (1977) for some related discussion of tests based linearly on these spacing-frequencies. Another possible approach may be to define "circular ranks". Since the zero direction as well as the sense of rotation (clockwise or anticlockwise) is arbitrary for the circle, there is no simple way to define ranks of the x-observations in the combined sample. This precludes the possibility of constructing simple rank tests for this problem as is done in linear rank theory (cf. Hajek and Sidak (1967)). Schach (1969) attempts to define equivalence classes of ranks on the circle but the approach does not seem very practical.

2. Asymptotic theory and efficiencies

In order to compute the asymptotic relative efficiency (ARE), one needs to obtain the asymptotic distribution of $T_{m,n}$ under an appropriate sequence of alternatives $A^{(m)}$ which converge to the null hypothesis. From the results in Holst and Rao (1980), an appropriate sequence in this case is of the form

$$A^{(m)} : G_m(y) = y + L_m(y)/m^{1/4} \ , \ 0 \le y < 1 \tag{2.1}$$

with $L_m(0) = L_m(1) = 0$. For these alternatives to be smooth and well behaved, we assume L_m is twice differentiable on $[0,1]$ and that there is a function $L(u)$, $0 \le u \le 1$ which is twice continuously differentiable and such that $L(0) = L(1) = 0$ and

$$\sup_{0 \le u \le 1} |L_m''(u) - L''(u)| = o(1) \ .$$

We shall write $L'(u) = \ell(u)$. Let η denote a geometric random variable with probability function

$$P(\eta = j) = \lambda(1-\lambda)^j \ , \ j = 0,1,2\ldots \tag{2.2}$$

where λ is the limit of $m/(m+n)$ as defined in (1.4).

The following result, Theorem 4.1 of Holst and Rao (1980), gives the asymptotic distribution of the statistic $T_{m,n}$ under the sequence of alternatives (2.1). Clearly, the asymptotic distribution of $T_{m,n}$ under the null hypothesis (1.1) is obtained simply by putting $\ell(u) \equiv 0$, $0 \le u \le 1$ in this result.

<u>THEOREM</u> (Holst and Rao)

Suppose there exist constants c_1, c_2 such that

$$h(j) < c_1(j^{c_2} + 1) \quad \text{for all } j. \tag{2.3}$$

Let $L_m(u)$ satisfy the conditions mentioned after (2.1) and let η be the geometric random variable defined in (2.2). Then the asymptotic distribution of

$$T^*_{m,n} = m^{-1/2} \sum_{k=1}^{m} [h(S_k) - Eh(\eta)] \tag{2.4}$$

is $N(\mu, \sigma^2)$ with

$$\mu = \left(\int_0^1 \ell^2(u)du\right)\left(\frac{\lambda^2}{2}\right)\mathrm{Cov}\left(h(\eta),\ \eta(\eta-1) - \frac{4\eta(1-\lambda)}{\lambda}\right) \tag{2.5}$$

and

$$\sigma^2 = \mathrm{Var}(h(\eta)) - \mathrm{Cov}^2(h(\eta), \eta)/\mathrm{Var}(\eta) \ . \tag{2.6}$$ ∎

This general result enables us to compute the asymptotic distributions both under the hypothesis as well as under the sequence of alternatives (2.1) for various choices of the function $h(\cdot)$. Under the present conditions, it is easy to see that the asymptotic "efficacy" of a test statistic T is (see Fraser (1957))

$$\text{efficacy} = \mu^4/\sigma^4 \tag{2.7}$$

where μ and σ^2 are as given in (2.5) and (2.6). The ARE of one test relative to another is simply the ratio of their efficacies. Since the term $(\int_0^1 \ell^2(u)du)^4$ appears in all the efficacies, the ARE's are indeed independent of which alternative sequence one considers. Thus we will calculate what we will call

$$\text{Modified efficacy} = \text{efficacy}/(\int_0^1 \ell^2(u)du)^4 \ . \tag{2.8}$$

The following four classes of test statistics will be considered. First let

$$h_1(x) = \begin{cases} 1 & \text{if} \quad x = r \\ 0 & \text{otherwise} \ . \end{cases} \tag{2.9}$$

The resulting statistic $T_{m,n}$ represents the number of runs of length r in the second sample. In particular if $r = 0$, the quantity $2(m - T_{m,n})$ is the number of circular runs and has the same ARE as $T_{m,n}$ with $r = 0$. For a discussion of runs on the circle, see David and Barton (1962, pp. 94-95, 132-136) and Mardia (1972, p. 203). For the choice $h_1(x)$ as in (2.9) equations (2.5) and (2.6) yield

$$\mu = \left(\int_0^1 \ell^2(u)du\right)\left(\frac{\lambda^2}{2}\right) P_r \{r(r-1) + 2((1-\lambda)/\lambda)^2 - 4r(1-\lambda)/\lambda\}$$

$$\sigma^2 = P_r(1-P_r) - P_r^2 (r - (1-\lambda)/\lambda)^2\lambda^2/(1-\lambda)$$

where

$$P_r = P(\eta = r) = \lambda(1-\lambda)^r \ .$$

In particular when $r = 0$, the modified efficacy defined in (2.8) gives $\lambda^2(1-\lambda)^4$. In Table 2.1 this modified efficacy is tabulated for various values of r and λ. It can be clearly seen that the case $r = 0$ (which corresponds to the run test) performs best among this group of tests.

Table 2.1

Modified efficacies for different values of r and λ derived from $h_1(x)$ i.e., based on the number of S_k equal to r

λ	r=0	r=1	r=2	r=3	r=4
.05	.0020	.0012	.0006	.0003	.0002
.10	.0067	.0018	.0004	.0001	.0000
.15	.0118	.0012	.0001	.0000	.0000
.20	.0164	.0005	.0000	.0000	.0002
.25	.0198	.0001	.0000	.0003	.0008
.30	.0216	.0000	.0001	.0008	.0011
.35	.0219	.0000	.0005	.0013	.0010
.40	.0207	.0000	.0011	.0013	.0005
.45	.0185	.0002	.0015	.0009	.0001
.50	.0156	.0004	.0015	.0004	.0000
.60	.0092	.0011	.0006	.0000	.0000
.70	.0040	.0011	.0000	.0000	.0000
.80	.0010	.0005	.0000	.0000	.0000
.90	.0001	.0001	.0000	.0000	.0000

Consider another class of tests given by

(2.10) $$h_2(x) = x^{\alpha}, \alpha > -1/2, \neq 0 \text{ or } 1 .$$

When $\alpha = 2$, the resulting statistic corresponds to the one proposed by Dixon (1940). With $\alpha = 2$, the statistic is also equivalent to

(2.11) $$\sum_{k=1}^{m} (S_k - \frac{n}{m})^2$$

which has a clear significance if one observes that under the null hypothesis $E(S_k) = n/m$. Simple closed expressions for μ and σ^2 are not available for this case unlike for the function $h_1(x)$. But numerical evaluation of the expressions in (2.5), (2.6) and (2.7) yields Table 2.2 which gives values of the modified efficacies of this group of tests for different values of α and λ.

Table 2.2

Modified efficacies for different values of α and λ

for the statistic $\sum_{k=1}^{m} S_k^{\alpha}$ corresponding to the case $h_2(x) = x^{\alpha}$

λ	α=.5	α=1.5	α=2.0	α=2.5	α=3	α=4
.05	.2978	.7608	.8145	.7695	.6576	.3726
.10	.2128	.6112	.6561	.6192	.5279	.2976
.15	.1621	.4850	.5220	.4921	.4187	.2348
.20	.1278	.3797	.4096	.3857	.3275	.1828
.25	.1022	.2928	.3164	.2977	.2522	.1401
.30	.0819	.2220	.2401	.2257	.1909	.1055
.35	.0651	.1650	.1785	.1677	.1416	.0779
.40	.0509	.1198	.1296	.1217	.1026	.0562
.45	.0389	.0847	.0915	.0859	.0723	.0395
.50	.0288	.0580	.0625	.0587	.0494	.0269
.60	.0139	.0239	.0256	.0241	.0203	.0110
.70	.0052	.0077	.0081	.0077	.0065	.0035
.80	.0012	.0015	.0016	.0015	.0013	.0007
.90	.0001	.0001	.0001	.0001	.0001	.0001

By inspection, it is clear that $h_2(x)$ with $\alpha = 2$ has the maximum efficacy for all values of λ. Instead of (2.11) one may consider the statistic based on absolute deviations

(2.12) $$\sum_{k=1}^{m} |S_k - \frac{n}{m}|$$

which, because of (1.4) is asymptotically equivalent to the choice of the function

(2.13) $$h_3(x) = |x - (1 - \lambda)/\lambda| .$$

The corresponding goodness of fit test based on spacings has been discussed by Rao (1969) and the ideas are reproduced in Mardia (1972, pp. 187-190). See also Rao (1976). Finally for the choice

(2.14) $$h_4(x) = \log (1 + x)$$

one gets a two-sample test that is analogous to a test proposed by Darling (1952) for the goodness of fit problem. Again simple closed expressions are not available for μ and σ^2 for these two cases. Using expressions (2.5), (2.6), and (2.7), numerical evaluations can be obtained to the desired degree of accuracy. The modified efficacies for the cases $h_3(x)$ and $h_4(x)$ defined by (2.13) and (2.14) are tabulated in Table 2.3 .

Table 2.3

Modified efficacies for different values of λ for the statistic $\sum_{k=1}^{m} |S_k - n/m|$ corresponding to $h_3(x)$ and $\sum_{k=1}^{m} \log(S_k + 1)$ corresponding to $h_4(x)$

λ	$\sum\|S_m - n/m\|$	$\sum \log(S_k + 1)$
.05	.2667	.2104
.10	.2137	.2084
.15	.1711	.1918
.20	.1304	.1691
.25	.0987	.1440
.30	.0781	.1190
.35	.0559	.0955
.40	.0431	.0742
.45	.0277	.0558
.50	.0156	.0404
.60	.0092	.0184
.70	.0040	.0064
.80	.0010	.0014
.90	.0001	.0001

Consideration of all the three tables leads to the conclusion that out of all the cases investigated, the test statistic (2.11) proposed by Dixon has the maximum efficiency. Indeed, Theorem 4.2 of Holst and Rao (1980) shows from theoretical considerations, that the Dixon's statistic is asymptotically optimal among all the symmetric test statistics. It may be noted finally that though the four classes of functions evaluated here have been used in connection with goodness of fit tests based on spacings (cf. for instance Sethuraman and Rao (1970)), only the special cases $h_1(x)$ with $r = 0$ and $h_2(x)$ with $\alpha = 2$ have been discussed so far in the literature for the two-sample problem.

It may also be observed from the tables, that the efficacies of all the tests decrease (except for $h_1(x)$ where it depends on the value of r) as λ increases from 0 to 1. This can be verified theoretically by considering the dominating terms in λ, in the expression (2.7) for efficacy. The following heuristic argument gives a further justification: Since the tests are based on the numbers of y's in between the x-spacings, it is desirable to have many more y's in relation to the x's i.e., n should be much greater than m. Otherwise many of these spacing-frequencies will be zero, which is not informative. Since λ is the limiting ratio of $m/(m+n)$, the larger efficacies thus correspond to smaller values of λ. Indeed one can assume without loss of generality that $\lambda < 1/2$ since otherwise the x's and y's can be relabelled to achieve this.

REFERENCES

[1] Darling, D.A., On a class of problems related to the random division of an interval. Ann. Math. Statist. 24, (1953), 239-253.

[2] David, F.N. and Barton, D.E., Combinatorial Chance (Griffin, London), (1962).

[3] Dixon, W.J., A criterion for testing the hypothesis that two samples are from the same population. Ann. Math. Statist. 11, (1940), 199-204.

[4] Fraser, D.A.S., Nonparametric Methods in Statistics (Wiley, New York), (1957).

[5] Govindarajulu, Z. and Sen, P.K., On a class of c-sample weighted rank-sum tests for location and scale, Ann. Inst. Statist. Math., 18, (1966), 87-105.

[6] Govindarajulu, Z., Asymptotic normality and efficiency of a class of test statistics, in Essays in Probability and Statistics, a volume in honor of J. Ogawa, Ed. S. Ikeda et al (Shinko Tsusho Co., Ltd., Tokyo (1977)), 535-558.

[7] Hajek, J. and Sidak, Z., Theory of Rank Tests, (Academic Press, New York (1967)).

[8] Holst, L. and Rao, J.S., Asymptotic theory for some families of two-sample nonparametric statistics. To appear in Sankhyā (1980).

[9] Mardia, K.V., Statistics of Directional Data. (Academic Press, London and New York (1972)).

[10] Rao, J.S., Some Contributions to the Analysis of Circular Data, Ph.D. Thesis. Indian Statist. Inst., Calcutta, (1969).

[11] Rao, J.S., Some tests based on arc-lengths for the circle, Sankhya, Ser. B. 38, (1976), 329-338.

[12] Rao, J.S. and Sethuraman, J., Pitman efficiencies of tests based on spacings, in Nonparametric Techniques in Statistical Inference. Cambridge University Press, (1970), 405-415.

[13] Schach, S., On a class of nonparametric two-sample tests for circular distributions. Ann. Math. Statist. 40, (1969), 1791-1800.