Sankhyā : The Indian Journal of Statistics
1976, Volume **38**, Series B. Pt. 4, pp. 329–338.

SOME TESTS BASED ON ARC-LENGTHS FOR THE CIRCLE

By J. S. RAO

University of California, Santa Barbara

SUMMARY. Tests for circular populations based on sample arc-lengths, i.e., distances between consecutive observations on the circumference, are considered for the following two problems (i) testing goodness of fit, and (ii) testing whether two circular populations are identical. For problem (i) if $D_1, \ldots, D_n$ are the sample arc-lengths made by a random sample of size n, we discuss in Sections 1 and 2 some tests based symmetrically on $\{D_i\}$ and comment on their asymptotic efficiencies. Section 3 deals with comparison of two circular populations for which we suggest a test called the V-test which is shown to be asymptotically superior to the standard Run test for the circle (see for instance David and Barton (1962)).

I. Introduction

We deal here with observations which are directions in two dimensions. The sample frame in this case can be conveniently taken to be the circumference of a unit circle, each point on it representing a direction. A Circular Distribution (CD) is a probability distribution concentrated on this circumference and a set of independent, identically distributed random variables from such a distribution will be referred to as a random sample from the CD. They can be represented as angles $\alpha_1, \ldots, \alpha_n$ $(0 \leqslant \alpha_i < 2\pi)$ with reference to a particular zero-direction and with a fixed sense of rotation (clockwise or anticlockwise) as positive. Since the choice of a zero-direction and the sense of rotation are arbitrary, we need decision procedures which are invariant under such choices.

The goodness of fit problem on the circle is to test whether a random sample $(\alpha_1, \ldots, \alpha_n)$ comes from a specified CD $F_0(\alpha)$, $0 \leqslant \alpha < 2\pi$. In the more nonparametric cases, we can make a probability integral transformation and the problem then is to test if $\{x_i = F_0(\alpha_i), i = 1, \ldots, n\}$ can be considered as a random sample from a uniform distribution on the circle. Broadly speaking, the test procedures available for this purpose on the line may be grouped into three categories, viz., (a) methods utilizing the empirical distribution functions, (b) methods based on χ^2 and (c) those based on sample spacings, i.e., differences between successive order statistics. Since the linear methods are not directly applicable to the circular data, Kuiper (1960) and Watson (1961) suggested origin-invariant modifications of the first group of techniques, while Ajne (1968) and Rao (1972b) considered what may be termed as modifications of χ^2 methods. Rao (1969) also suggested adapting the methods based on spacings for the circle and the aim of the present paper is to focus attention on this third group of tests. An important point that should be mentioned in this connection, is that these circular spacings or arc langths form a maximal invariant for the problem and hence any invariant statistic, even one obtained by modifying methods (a) and (b), is a function of these arc lengths. Thus all the known invariant procedures suggested for the circle are based on arc-lengths, which thus play a much more prominent role in the case of the circle than do the spacings on the line.

Let $\alpha_1, \ldots, \alpha_n$ be a random sample from a CD on the basis of which we wish to test the hypothesis of uniformity, $H_0 : f(\alpha) = 1/2\pi$ where $f(\cdot)$ denotes the density function. Let $\alpha_{(1)} \leqslant \ldots \leqslant \alpha_{(n)}$ be the order statistics, obtained by arranging $\alpha_1, \ldots, \alpha_n$ in ascending order. Define the sample arc lengths

$$D_i = \alpha_{(i)} - \alpha_{(i-1)}, \quad i = 1, \ldots, n \qquad \ldots \quad (1.1)$$

where we take $\alpha_{(0)} = \alpha_{(n)} - 2\pi$. These $\{D_i, i = 1, \ldots, n\}$ represent the distances between successive observations on the circumference and remain invariant under the choice of zero-direction or the sense of rotation. The joint distribution of the arc-lengths under the hypothesis of uniformity can easily seen to be (see Rao, 1969; Mardia, 1972) a $(n-1)$-variate Dirichlet distribution $D(1, 1, \ldots 1; 1)$ (see Wilks, 1963) with density

$$f_{D_1, \ldots, D_{n-1}}(d_1, \ldots, d_{n-1}) = (n-1)!/(2\pi)^{n-1}$$

$$\text{for} \qquad 0 \leqslant d_i \leqslant 2\pi, \ \sum_{i=1}^{n-1} d_i \leqslant 2\pi. \qquad \ldots \quad (1.2)$$

These D_i's are interchangeable (or exchangeable) random variables under H_0 and have the same distribution as the spacings from a sample of size $(n-1)$ from the uniform distribution on the line segment $[0, 2\pi]$. This suggests we could apply any of the spacings tests on the line (see e.g. Pyke, 1965), to the circular case with minor modifications. More common among these are tests using symmetric spacings statistics of the form

$$T_n = \sum_{i=1}^{n} h(nD_i)/n \qquad \ldots \quad (1.3)$$

where for instance we may take $h(x) = x^r$ $(r > 0)$, $\frac{1}{2}|x-1|$, $\log x$, to get different tests. Also of interest are tests based on ordered spacings or linear combinations of spacings. Among the latter, $D_{(n)} = \max\limits_{1 \leqslant i \leqslant n} D_i$ is of special interest on the circle since $(2\pi - D_{(n)})$, which is the length of the shortest arc containing all the sample points, is called the "circular range" and has been investigated in Rao (1969). For further literature on spacings, see Pyke (1965) and the references contained there.

To discuss the asymptotic relative efficiencies (ARE's) of various spacings tests of the form (1.3), Sethuraman and Rao (1970) consider a sequence of alternate densities

$$A_n : g_n(x) = 1/2\pi + l(x)/n^{\delta}, \quad 0 \leqslant x < 2\pi \quad (\delta \geqslant 1/4) \qquad \ldots \quad (1.4)$$

which converges to the uniform density on he circle as $n \to \infty$. Here $l(\cdot)$ is assumed to be square integrable and continuously differentiable on $[0, 2\pi]$. Sethuraman and Rao (1970) obtain the limiting distributions of various spacings statistics of the form (1.3) both under H_0 and A_n using ideas of weak convergence of the empirical process of the spacings. This weak convergence approach gives an elegant and unified method for deriving the asymptotic distributions of the spacings statistics unlike the sundry and often complex methods devised earlier separately for each case. We quote two results from this paper since they will be useful later in Section 3. Define the empirical distribution function of $\{nD_i, \, i = 1, \ldots, n\}$ by

$$H_n(x) = \sum_{i=1}^{n} I(nD_i; x)/n \quad \text{for} \quad x \geqslant 0$$

where $I(z; x) = 1$ if $z \leqslant x$, and $= 0$ if $z > x$.

Let $$F_n(2\pi x) = 1 - e^{-x} + e^{-x}(x - x^2/2)\Big(\int_0^{2\pi} l^2(p)dp\Big)/n^{2\delta}. \qquad \ldots \quad (1.5)$$

Then we have

Theorem 1 : (Sethuraman and Rao, 1970). *Under the alternatives* A_n, *the process* $\{\zeta_n(x) = \sqrt{n}[H_n(x) - F_n(x)], \; x \geqslant 0\}$ *converges weakly to a Gaussian process* $\{\zeta(x), \; x \geqslant 0\}$ *in* $D[0, \infty]$ *with mean function zero and covariance kernel* $K(2\pi s, 2\pi t) = e^{-t}(1 - e^{-s} - ste^{-s})$ *for* $0 \leqslant s \leqslant t \leqslant \infty$.

If $h(\cdot)$ is a function such that for $y \in D[0, \infty]$, the mapping $y \to \int_0^{\infty} h(x)dy(x)$ is continuous with probability one under ζ, then

$$T_n = \sum_{i=1}^{n} h(nD_i)/n$$

$$= \int_0^{\infty} h(x)dH_n(x)$$

can be considered as a continuous functional of the process ζ_n, and we have

Theorem 2 : (Sethuraman and Rao, 1970). *Under the sequence of alternatives A_n, the random variable*

$$T_n^* = \sqrt{n}(T_n - \int_0^\infty h(x)dF_n(x))$$

has a limiting normal distribution with mean zero and variance

$$\sigma^2 = \int_0^\infty \int_0^\infty h'(s)h'(t)K(s,t)ds\,dt.$$

For more details and remarks on the efficiencies see Sethuraman and Rao (1970) and for details on empirical spacings process, see Rao and Sethuraman (1975).

2. A spacings test of uniformity

Among tests of the form (1.3), the statistic

$$U_n = \tfrac{1}{2} \sum_{i=1}^{n} |D_i - 2\pi/n|$$

$$= \sum_{i=1}^{n} \max(D_i - 2\pi/n, 0) \qquad \ldots \quad (2.1)$$

which corresponds to taking $h(x) = 1/2|x-1|$, has a particularly nice interpretation for the circle. Suppose we place n arcs of fixed length $2\pi/n$ each, starting with each of the samplė points on the circumference. The circumference would be completely covered by these fixed arcs only when the sample points are equispaced on the circumference. The uncovered portion of the circumference contributed by the i-th observation is given by max $(D_i - 2\pi/n, 0)$ whereas U_n gives the total uncovered portion of the circumference (or equivalently the extent to which the fixed arcs overlap each other). U_n takes values in the interval $[0, 2\pi(1-1/n)]$. Large values of U_n indicate clustering of the sample points or evidence for rejecting the hypothesis of uniformity. Stevens (1939) considers the probability that the circumference is completely covered when n arcs of arbitrary lengths are randomly placed on the circumference of a unit circle. In our case, the probability of complete coverage is zero and we can in fact, find the distribution of the uncovered portion, U_n. The exact distribution of U_n can be found using the contour integral for the characteristic function developed in Darling (1953) and the density of U_n is seen to be (see Darling, 1953)

$$f_n(u) = (n-1)! \sum_{j=1}^{n-1} \binom{n}{j} (u/2\pi)^{n-j-1} g_j(nu)/(n-j-1)!\, n^{j-1}, \ \ldots \quad (2.2)$$

$$\text{for} \quad 0 < u < 2\pi(1-1/n)$$

where $g_j(x)$ is the density of the sum of j independent uniform random variables on $[0, 2\pi)$ and is given by

$$g_j(x) = [1/(j-1)!(2\pi)] \sum_{k=0}^{\infty} (-1)^k \binom{j}{k} \left\langle x/2\pi - k \right\rangle^{j-1} \qquad \ldots \quad (2.3)$$

with the notation $\langle x \rangle = x$ if $x > 0$ and $= 0$ if $x \leqslant 0$. In Table 1 we give the upper percentage points for the statistic U_n for testing the hypothesis of uniformity on the circle. If for a given sample size n and level α, the calculated value of U_n exceeds the tabulated critical point $U_0(\alpha, n)$, we reject the hypothesis of uniformity. The critical points have been given in terms of degrees for ready applicability.

TABLE 1. CRITICAL VALUES FOR THE TEST STATISTIC U IN DEGREES

n	$\alpha = 0.01$	0.05	0.10
4	$U =$ 221.0	186.5	171.7
5	212.0	183.6	168.8
6	206.0	180.7	166.3
7	202.7	177.8	164.9
8	198.4	175.7	163.4
9	195.1	173.5	162.4
10	192.2	172.1	161.3
11	189.7	170.7	160.2
12	187.6	169.2	159.2
13	185.8	167.8	158.4
14	184.0	166.7	157.7
15	182.2	165.6	157.0
16	180.7	164.9	156.6
17	179.6	164.2	155.9
18	178.2	163.1	155.2
19	177.1	162.4	154.8
20	176.0	161.6	154.4
25	171.9	158.9	152.7
30	168.8	156.7	151.4
35	166.4	155.0	150.3
40	164.4	153.6	149.5
45	162.7	152.4	148.7
50	161.2	151.4	148.1
100	152.8	146.8	143.7
200	146.8	142.6	140.4

The following numerical example illustrates the use of the U_n test :

Example : In an experiment on homing orientations in pigeons, 10 birds were released singly at 25 km west of their loft. Field glass observation yielded the vanishing points of each departing bird as 20, 35, 350, 120, 85, 345, 80, 320, 280 and 85 degrees. It is required to know whether the birds have a preferred orientation of flight.

The arc lengths $\{D_i\}$ made by these observations on the circle are easily computed to be 15, 45, 5, 0, 35, 160, 40, 25, 5 and 30 degrees and $\frac{2\pi}{10} = 36°$. Therefore

$$U_{10} = B.\ \tfrac{1}{2} \sum_{i=0}^{10} |T_i - 36\ | = 137°.$$

This value of 137° for $n = 10$ is not significant even at the 10 per cent level of significance as the critical value is 161.3°. Thus we conclude that the data does not indicate a preferred direction of flight for these birds.

Rao (1972a) compares the Bahadur efficiency of the spacings test U_n with other known tests for uniformity using the circular normal alternatives. Pitman efficiency of U_n is compared with those of other spacings tests in Sethuraman and Rao (1970).

3. Two two-sample tests and their ARE's

Let $\alpha_1, \ldots, \alpha_m$ and $\beta_1, \ldots, \beta_n$ denote independent samples from two CD's with distribution functions F and G respectively. We wish to test if the two populations are identical, i.e., $H_0 : F = G$. We will consider two methods that make use of spacings. We can assume $m \geqslant n$ without loss of generality and let $\beta_{(1)} \leqslant \ldots \leqslant \beta_{(n)}$ be the order statistics for the second sample. Let

$$S_i = \text{number of } \alpha_j\text{'s in between } (\beta_{(i-1)}, \beta_{(i)}]\ i = 2, \ldots, n$$

and

$$S_1 = m - \sum_{i=2}^{n} S_i.$$

That is, these $\{S_i, i = 1, \ldots, n\}$ denote the number of observations of the first sample falling within the sample arcs made by the second sample. These numbers $\{S_i\}$ are clearly invariant under change of zero-direction or the sense of rotation and we consider two tests based on them. The first of these is the standard Run test for the circle. Since any non-zero S_i constitutes an α-run, the number of runs made by the first sample (or the α-runs) is given by

$$R_\alpha = \sum_{i=1}^{n} \delta(S_i)$$

where $\delta(x) = 0$ if $x = 0$ and $=1$ otherwise. Since the number of runs made by the second sample are exactly the same on a circle, the total number of runs made by the combined sample is simply

$$R_{m,n} = 2 \sum_{i=1}^{n} \delta(S_i). \qquad \ldots \quad (3.1)$$

The distribution of $R_{m,n}$ is available in e.g., David and Barton (1962). The other statistic, suggested by Dixon (1940) for the linear case, is

$$V_{m,n} = \sum_{i=1}^{n} S_i^2/n. \qquad \ldots \quad (3.2)$$

See Dixon (1940) for its mean and variance. Though its exact distribution is hard to get, the asymptotic normality is easily shown as we do in this section. Consistency properties of both these tests are discussed in Blum and Weiss (1957).

We now compute the Pitman's ARE of V against R and show that V is asymptotically more efficient. This comparison was attempted earlier by Blumenthal (1963). But unfortunately he uses there an incorrect result of Weiss (1957), which supposedly gives the distribution of V under general alternatives. See Pyke (1965), p. 417 for a comment on this error. However, since for computing the ARE we need the distribution of V only under a sequence of alternatives that converge to H_0, we can find this using Theorem 2 of the earlier section. The approach in this section is similar to that of Blumenthal (1963) and we establish the correctness of his final result.

Since the values $\{S_i\}$ remain the same under a probability integral transformation on the original observations, it is more convenient to consider

$$x_i = F(\alpha_i),\ i = 1, \ldots, m \text{ and } y_j = F(\beta_j),\ \ j = 1, \ldots, n \qquad \ldots \quad (3.3)$$

as the observations. These have values now in the unit interval with x's having a uniform distribution and the y's having density

$$g_1(y) = g(F^{-1}(y))/f(F^{-1}(y)),\quad 0 \leqslant y \leqslant 1,$$

where f, g denote the densities corresponding to F and G. The hypothesis H_0 is equivalent to $H_0' : g_1(y) = 1,\ 0 \leqslant y \leqslant 1$. We will again consider a sequence of alternatives

$$A_n : g_1(y) = 1 + l(y)/n^{1/4},\quad 0 \leqslant y \leqslant 1. \qquad \ldots \quad (3.4)$$

If

$$m, n \to \infty \text{ such that } m/n \to \lambda > 0 \qquad \ldots \quad (3.5)$$

then Wald and Wolfowitz (1940) show that under H_0,

$$\sqrt{n}(R/n-2\lambda/(1+\lambda))$$

is asymptotically normal with mean zero and variance $4\lambda^2/(1+\lambda)^3$. Also Blumenthal (1963) shows under the alternatives (3.4) the asymptotic distribution of

$$\sqrt{n}[R/n-2\lambda/(1+\lambda)+2\lambda^2(\int_0^1 l^2(p)dp)/(1+\lambda)^3\sqrt{n}]$$

is normal with mean zero and the same variance $4\lambda^2/(1+\lambda)^3$ as under H_0. We now need the distribution of V under H_0 as well as under A_n which we obtain using Theorem 2 of Section 1. Note that

$$\begin{aligned} V &= \sum_1^n S_i^2/n \\ &= m/n+2\sum_{i=1}^n \binom{S_i}{2}\Big/ n \\ &= m/n+2 \sum_{1\leq i_1 < i_2 \leq m} t(x_{i_1}, x_{i_2})/n \end{aligned}$$

where $t(x_i, x_j) = 1$ if (x_i, x_j) belong to the same y-arc, and $= 0$ otherwise. Clearly,

$$E(t(x_1,x_2) \mid Y) = P[t(x_1,x_2) = 1 \mid Y] = \sum_{i=1}^n (DY_i)^2$$

where DY_i denotes the length of the i-th Y-arc. Thus

$$\begin{aligned} E(V \mid Y) &= m/n+m(m-1)\sum_1^n (DY_i)^2/n \\ &= m/n+(m^2/n^2)\left[n \sum_{i=1}^n (DY_i)^2\right]-m/n \sum_{i=1}^n (DY_i)^2. \quad \ldots \quad (3.6) \end{aligned}$$

Under conditions (3.5), the asymptotic distribution of $E(V \mid Y)$ is the same as that of $(\lambda+\lambda^2 Z)$ where Z has the limiting distribution of $\left[n \sum_{i=1}^n (DY_i)^2 \right]$. This is so since the last term in (3.6) converges stochastically to zero. Now from Theorem 2 of Section 1 we have

$$\sqrt{n}\left[n \sum_{i=1}^n (DY_i)^2-2-(2\int_0^1 l^2(p)dp)/\sqrt{n}\right]$$

is asymptotically normal with mean zero and variance 4. Hence the asymptotic distribution of

$$\sqrt{n}\left[E(V\mid Y)-\lambda-2\lambda^2-2\lambda^2\left(\int_0^1 l^2(p)dp\right)/\sqrt{n}\right]$$

is normal with mean zero and variance $4\lambda^4$ under the alternatives (3.4). Also under these alternatives (3.4) Blumenthal's Theorem 3.1 gives that for every fixed Y, the conditional distribution of V given Y is asymptotic normal with mean $E(V\mid Y)$ and variance $4\lambda^2(1+2\lambda)$. These two results can be combined to conclude (see for instance Hajek and Sidak (1967) p. 195 or Theorem 2.1 of Blumenthal (1963))

$$\sqrt{n}[V-\lambda-2\lambda^2-2\lambda^2\int_0^1 l^2(p)dp/\sqrt{n}]$$

is asymptotically normal with mean zero and variance $4\lambda^2(1+\lambda)^2$. Clearly the distribution of V under H_0 is obtained by putting $l(x)\equiv 0$. Thus the efficacy of the Run test for testing H_0 against the sequence of alternatives A_n in (3.4) is given by

$$e_R = \lambda^2\left(\int_0^1 l^2(p)dp\right)^2/(1+\lambda)^3$$

while the efficacy of V is

$$e_V = \lambda^2\left(\int_0^1 l^2(p)dp\right)^2/(1+\lambda)^2.$$

Thus the ARE of the Run test against the V-test

$$e_{R,V} = 1/(1+\lambda),$$

independent of $l(x)$. This implies that the V-test is always more efficient than the Run test, the efficiency increasing with λ, the limiting proportion of observations in the two samples. This is to be expected in view of the fact that V-test utilizes more information than the R-test. While R-test merely counts the number of runs, the V-test also takes into account their magnitudes.

Acknowledgement

The author would like to thank Professor C. R. Rao of the Indian Statistical Institute for the advice and encouragement while this work was in progress. Thanks are also due to the editors for their remarks which improved the presentation.

References

AJNE, B. (1968) : A simple test for uniformity of a circular distribution. *Biometrika*, **55**, 343-54.

BLUM, J. R. and WEISS, L. (1957) : Consistency of certain two-sample tests. *Ann. Math. Statist.*, **28**, 242-246.

BLUMENTHAL, S. (1963) : The asymptotic normality of two test statistics associated with the two sample problem. *Ann. Math. Statist.*, **34**, 1513-1523.

DARLING, D. A. (1953) : On a class of problems related to the random division of an interval. *Ann. Math. Statist.*, **24**, 239-253.

DAVID, F. N. and BARTON, D. E. (1962) : *Combinatorial Chance*, Haffner Publ. Co. New York, 94-95, 132-136.

DIXON, W. J. (1940) : A criterion for testing the hypothesis that two samples are from the same population. *Ann. Math. Statist.*, **11**, 199-204.

HAJEK, J. and SIDAK, Z. (1967) : *Theory of Rank Tests*, Academic Press, New York, 195.

KUIPER, N. H. (1960) : Tests concerning random points on a circle. *Ned. Akad. Wetensch. Proc.* Ser. A, **63**, 38-47.

MARDIA, K. V. (1972) : *Statistics of Directional Data*, Academic Press, New York, 171-172.

PYKE, R. (1965) : Spacings. *Jour. Roy. Stat. Soc.* Ser. B, **27**, 395-449.

RAO, J. S. (1969) : Some contributions to the analysis of circular data. Unpublished Ph.D. thesis. Indian Statistical Institute, Calcutta.

——— (1972a) : Bahadur efficiencies of some tests for uniformity on the circle. *Ann. Math. Statist.*, **43**, 468-479.

——— (1972b) : Some variants of chi-square for testing uniformity on the circle. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **22**, 33-44.

RAO, J. S. and SETHURAMAN, J. (1975) : Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors. *Ann. Statist.*, **3**, 299-313.

SETHURAMAN, J. and RAO, J. S. (1970) : Pitman efficiencies of tests based on spacings, *Nonparametric Techniques in Statistical Inference*, Cambridge Univ. Press, 405-415.

STEVENS, W. L. (1939) : Solution to a geometrical problem in probability. *Ann. Eugenics*, **9**, 315-320.

WALD, A. and WOLFOWITZ, J. (1940) : On a test whether two samples are from the same population. *Ann. Math. Statist.*, **11**, 147-162.

WATSON, G. S. (1961) : Goodness of fit tests on a circle. *Biometrika*, **48**, 109-114.

WEISS, L. (1957) : The asymptotic power of certain tests of fit based on sample spacings. *Ann. Math. Statist.*, **28**, 783-786.

WILKS, S. S. (1963) : *Mathematical Statistics*, John Wiley, 177-182.

Paper received : January. 1975.

Revised : January, 1977.