Ann. Inst. Statist. Math.
29 (1977), Part A, 287-293

WHITWORTH RUNS ON A CIRCLE

M. A. STEPHENS

(Received Feb. 2, 1970; revised June 15, 1976)

Summary

Suppose different classes of items, for example, beads of different colours. are placed in a circle. Two probability models have been proposed, which lead to different distributions of runs, i.e. sequences of one colour. Barton and David [3] have called these Whitworth runs and Jablonski runs, and have tabulated the distributions for small samples. Asano [1] has extended the tabulations for Jablonski runs. In this paper, Whitworth runs are examined, particularly some approximations to the distributions which avoid extensive tabulations. Some potential uses of Whitworth runs are also pointed out.

1. Introduction

1.1. In this article we discuss the distribution of Whitworth runs on a circle. These are runs which arise when samples are taken randomly from two or more identical distributions on the circle; this is equivalent to taking samples successively from the same distribution. Suppose the first sample gives points $P_1, P_2, \cdots$, on the circle, the second sample gives points $Q_1, Q_2, \cdots$, the third gives points $R_1, R_2, \cdots$, etc. A typical pattern, starting at an arbitrary origin, might be *RRPPRQ QQRRPPRQPRR*, for samples of size 5 for P, 4 for Q, and 8 for R. A *run* is defined as a sequence of the same letter; in the pattern above, arranged around a circle, there are 3 runs of P (of lengths 2, 2, and 1), 2 of Q (lengths 3 and 1) and 4 of R (lengths 1, 2, 1 and 4). In counting the runs on the circle, the first two R and the last two join to make one run of length four. It is this property which makes for a distribution of runs on a circle, different from the distribution of runs on a line. Runs obtained from samples drawn on the above model are called Whitworth runs after the man who earlier gave the distribution; Barton and David [3] also call them *runs in repeated sampling.*

Probabilities can be attached to Whitworth runs by considering patterns, or arrangements, on a line, such as the one given above and finding their probabilities on the hypothesis that the sets P, Q and R

come from the same distribution; these arrangements are then placed round a circle and the runs counted. A line arrangement with T runs gives $T-1$ or T runs on a circle, depending on whether the first and last letters on the line are the same or not. Several arrangements on the line will often become indistinguishable on the circle, since rotations will not be distinguishable; thus *PQQPPQ*, *QQPPQP*, and four similar patterns obtained by cyclic permutation, all become indistinguishable on a circle. With three P and three Q, there are four distinguishable patterns: the one above, and those obtained by *PPQQPQ*, *QQQPPP*, *QPQPQP* and their cyclic permutations; these give, on the model of repeated sampling above, 2, 4 or 6 runs with probabilities .3, .6, .1 respectively. If, in a different model, the four *distinguishable arrangements* on the circle are considered equally likely, the probabilities of 2, 4 or 6 runs become .25, .5, .25 respectively. Runs provided on this second model are called *Jablonski* runs; the distribution of Jablonski runs was discussed and tabulated by Barton and David [3] and extended by Asano [1].

1.2. In the next sections, the distribution of Whitworth runs is discussed with a view to testing the hypothesis H_0: that k populations of points on a circle come from identical populations. For small samples, in Section 2, we use tables prepared for the line by Swed and Eisenhart [5] and quoted, for example, in Owen [4]; these refer to two samples of sizes M, N, both $M, N \leqq 20$. For three or four samples, the total number of observations not exceeding twelve, we use tables prepared by Barton and David [3]. Since a runs test is a quick and convenient technique especially suited to large samples, we also examine, in Section 3, two approximations, for large samples, given by Barton and David [3]. These are shown to be very good even for relatively small samples, and can be used for values beyond those given in the exact tables; this will be valuable when several samples are involved. In the next sections the tests have size α; since discrete probabilities are involved, H_0, when true, is rejected with probability as close as possible to α, but less than α. The tests given are one-sided tests (the null hypothesis will be rejected for too few runs) but can be easily adapted to two-sided tests. The tables of Swed and Eisenhart [5] will be referred to as SE, and those of Barton and David [3] as BD.

2. Exact tests

2.1. *Case* (a). *An exact two-sample test of* H_0

Suppose the two sample sizes are M, N ($M, N \leqq 20$). The test consists of the following steps:

1. Count the number of runs T on the circle.
2. Find $p=\Pr(\boldsymbol{U}\leqq T+1)$ from the SE Table 1.
3. If $p<\alpha$, reject H_0.

Case (a) *Critical region.* It is often useful to establish a critical region for T, such that H_0 is rejected if T falls within it. This may be done from the SE Table 1 by finding the largest odd number u_1 so that the probability $p_1=\Pr(\boldsymbol{U}\leqq u_1)$ is less than or equal to α. The critical region is then $T\leqq u_1-1$.

Alternatively, SE Table 2, giving critical regions for the straight line, for M, $N\leqq 20$, may be used as follows. For given M, N, and usual values of α, SE Table 2 gives u_α such that, for the lower tail,

$$\Pr(\boldsymbol{U}\leqq u_\alpha)\leqq\alpha \qquad \text{and} \qquad \Pr(\boldsymbol{U}\geqq u_\alpha+1)>\alpha\,;$$

and for the upper tail,

$$\Pr(\boldsymbol{U}\geqq u_\alpha)\leqq\alpha \qquad \text{and} \qquad \Pr(\boldsymbol{U}\geqq u_\alpha-1)>\alpha\,.$$

If u_α is even, the lower tail critical region on the circle is $T\leqq u_\alpha$; if u_α is odd, the critical region is $T\leqq u_\alpha-1$.

2.2. *Case* (b). *An exact test of H_0, for three or four samples*

Suppose the k sample sizes are r_1, r_2, r_3, and possibly r_4, and r, the total number of observations, satisfies $r\leqq 12$. The test is as follows.

1. Count the number of runs T on the circle.
2. From BD Table 1b, ($k=3$) or Table 1c ($k=4$), choose the horizontal line in which the sample sizes r_1, r_2, r_3, (r_4), are found under the heading "Partition". Add the table entries to find the number of arrangements giving T runs or less, and divide by the multinomial term for the line. This gives $p=\Pr(\boldsymbol{T}\leqq T)$.
3. If $p<\alpha$, reject H_0.

Case (b) *Critical region.* This is given by $T\leqq T_1$, where T_1 is the largest integer for which $p\leqq\alpha$, where $p=\Pr(T\leqq T_1)$.

3. Approximate tests

3.1. When the sample sizes are beyond the ranges quoted for the exact tests, the SE and BD exact tables cannot be used. An approximate test is given below for use in these situations. The test is based on two approximations to the distribution of $\boldsymbol{T}$, suggested by Barton and David [3]. They are the circular analogues to approximations to the distribution of $\boldsymbol{U}$ which these authors investigated in detail in Barton and David [2]. Their notation is followed as much as possible.

For sample sizes r_i, $i=1, 2, \cdots, k$, define

$$F_2=\sum_{i=1}^{k} r_i(r_i-1) \qquad \text{and} \qquad F_3=\sum_{i=1}^{k} r_i(r_i-1)(r_i-2)\,.$$

The mean and variance of T are then given by

$$\mu=r-\frac{F_2}{r-1} \tag{1}$$

and

$$\sigma^2=\frac{1}{(r-1)(r-2)}\left(\frac{F_2^2}{r-1}+F_2(r-4)-2F_3\right) \tag{2}$$

where, as before, $\sum_{i=1}^{k} r_i=r$. Two special cases often occur, for which simpler expressions can be found for μ and σ^2.

Case (a). Two samples, sizes M, N.

$$\mu=\frac{2MN}{M+N-1}\,; \qquad \sigma^2=\frac{\mu^2-2\mu}{M+N-2}=\frac{4MN(MN+1-M-N)}{(M+N-1)^2(M+N-2)}\,. \tag{3}$$

Case (b). k samples of equal size s.

$$\mu=\frac{ks^2(k-1)}{ks-1}\,; \qquad \sigma^2=\frac{\mu(s-1)}{ks-1}\,. \tag{4}$$

3.2. *Approximate distributions*

Approximation 1. Suppose z has a normal distribution with mean μ and variance σ^2 given above, written $N(\mu, \sigma^2)$. As $r\to\infty$, k remaining fixed, the distribution of T may be approximated by z. With only two samples, T must be even, and when the continuity correction is introduced $\Pr(T<T)$ is approximately given by $\Pr(z<T+1)$. For three or more samples, $\Pr(T<T)\simeq\Pr(z<T+1/2)$.

Approximation 2. A binomial expansion $(q+p)^n$ may be fitted to the distribution of T, with n and p chosen so that the mean is μ and the variance is σ^2; $(q=1-p)$. This can lead to non-integral n; but in the special case of k equal sample sizes s, the values of n, p and q are

$$n=ks=r\,; \qquad p=\frac{(k-1)s}{ks-1}\,; \qquad q=\frac{s-1}{ks-1}\,. \tag{5}$$

With this approximation, $\Pr(T=T)$ is given approximately by the coefficient of p^T in $(q+p)^n$. If the coefficient of p^i is c_i,

$$\Pr(T\leqq T)\simeq\sum_{i=0}^{T} c_i=p_\alpha\,, \qquad \text{say}\,. \tag{6}$$

Examination of the approximations. Tables 1 and 2 show, for sev-

Table 1 Comparison of true and approximate cumulative distribution functions of T for two samples

Sample sizes	No. of runs	2	4	6	8	10	12	14	16	18	20
6, 6	True	.0130	.1753	.6082	.9329	.9870	1.0000				
	Approx. 1	.0199	.1851	.6039	.9226	.9951	.9999				
15	True	.0013	.0374	.2722	.7417	1.0000					
5	Approx. 1	.0017	.0359	.2890	.7541	.9733					
20	True	.0000	.0000	.0000	.0001	.0009	.0075	.0380	.1301	.3143	.5619
20	Approx. 1	.0000	.0000	.0000	.0001	.0013	.0087	.0406	.1332	.3161	.5612

eral combinations of sample size, the values of the true cumulative distribution functions and those obtained from the approximations. The binomial approximation has been examined only for 3 or 4 equal samples. It is clear that both approximations give good results, even for very unequal samples. For most cases, the normal approximation will be good enough to use for significance tests at the usual levels, provided borderline judgements are treated with reserve. Then it might be useful to use the binomial approximation, which evidently gives greater accuracy. The test will therefore be as follows.

3.3. *An approximate test of* H_0

The test is for use with

Table 2 Comparison of true and approximate cumulative distribution functions of T for 3 or 4 samples

Sample sizes	No. of runs	3	4	5	6	7	8	9	10	11	12
444	True	.00069	.0038	.0225	.0786	.2095	.4262	.6776	.8771	.9664	1.0000
	Approx. 1	.00030	.0031	.0181	.0748	.2132	.4415	.6918	.8747	.9639	.9930
	Approx. 2	.00079	.0050	.0231	.0794	.2080	.4223	.6763	.8796	.9781	1.0000
642	True	.0017	.0095	.0494	.1515	.3593	.6212	.8463	.9632	.9978	1.0000
	Approx. 1	.0012	.0089	.0453	.1549	.3675	.6325	.9451	.9547	.9911	.9988
831	True	.0121	.0667	.2364	.5333	.7879	1.0000				
	Approx. 1	.0090	.0620	.2380	.5448	.8259	.9611				
3333	True		.0002	.0018	.0114	.0503	.1614	.3790	.6697	.9081	1.0000
	Approx. 1		.0000	.0006	.0065	.0414	.1619	.4059	.6951	.8959	.9776
	Approx. 2		.0003	.0022	.0122	.0508	.1595	.3767	.6700	.9100	1.0000
5421	True		.0009	.0078	.0433	.1515	.3719	.6558	.8792	.9805	1.0000
	Approx. 1		.0007	.0067	.0403	.1534	.3833	.6659	.8757	.9699	.9954
7311	True		.0046	.0409	.1742	.4470	.7424	.9546	1.0000		
	Approx. 1		.0048	.0389	.1741	.4552	.7620	.9380	.9910		

(a) two samples, when either M or N is greater than 20, or
(b) more than two samples, when r, the total number of observations, exceeds 12. The test, of significance level α, consists of the following steps:
   1. Find μ, σ^2 from equations (1) and (2), or (3) or (4).
   2. Count the number of runs T on the circle.
   3. Suppose z is $N(\mu, \sigma^2)$, with μ, σ^2 given by step 1. For *two samples*, find $p_0 = \Pr(z \leqq T+1)$. For *more than two samples*, find $p_0 = \Pr(z \leqq T+0.5)$. The probability $\Pr(T \leqq T)$ is then approximately p_0.
   4. If $p_0 < \alpha$, reject H_0.

If the sample sizes are small and equal, and if p_0 is very near α (Tables 1 and 2 suggest $|p_0 - \alpha| < 0.005$) the binomial approximation might be used. In this case, n, p, q are found from equation (5), and p_α from (6). If $p_\alpha < \alpha$, H_0 is rejected.

3.4. *Critical region for T*

This is constructed as follows: Let $z = N(\mu, \sigma^2)$ as above. For *two samples*, find the largest odd number u such that $p \leqq \alpha$, where $p = \Pr(z \leqq u)$. The critical region is then $T \leqq u-1$. For *more than two samples*, find the largest integer v such that $p \leqq \alpha$, where $p = \Pr(z \leqq v+0.5)$. The critical region is $T \leqq v$. If the binomial approximation is used, suppose s is an integer such that, if $p_1 = \sum_{i=1}^{s} c_i$ and $p_2 = p_1 + c_{s+1}$ (with c_i as in Section 3.2), $p_2 \geqq \alpha \geqq p_1$. The critical region is then $T \leqq s$.

Use of SE Table 3. This table gives the critical regions for U, for various values of α, for *two equal samples*, sizes up to 100. The critical regions for T are found by following the steps in Section 2.1. The SE Table has been itself constructed from the corresponding normal approximation for the distribution of U.

Acknowledgements

The author is grateful for support for this research from the National Research Council of Canada, and the U.S. Office of Naval Research.

McMaster University

References

[1] Asano, Chooichiro (1965). Runs test for a circular distribution and a table of probabilities, *Ann. Inst. Statist. Math.*, 17, 331-346.
[2] Barton, D. E. and David, F. N. (1957). Multiple runs, *Biometrika*, 44, 168-176.

[3] Barton, D. E. and David, F. N. (1958). Runs in a ring, *Biometrika*, **45**, 572-578.
[4] Owen, D. B. (1962). *Handbook of Statistical Tables*, Addison-Wesley, Reading, Mass.
[5] Swed, F. S. and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives, *Ann. Math. Statist.*, **14**, 66-87.