BIOMETRICS 41, 237–244
March 1985

A Runs Test Based on Run Lengths

Peter C. O'Brien[1] and Peter J. Dyck[2]

Section of Medical Research Statistics[1] and Department of Neurology[2],
Mayo Clinic, Rochester, Minnesota 55905, U.S.A.

SUMMARY

A procedure is proposed for testing the hypothesis that Bernoulli trials (successes and failures) are independent with common probability of success. Equivalently, the procedure may be used to test the hypothesis that the arrangement of a fixed number of successes and failures was determined randomly. The procedure is based on a weighted linear combination of the variances of run lengths of successes and failures. It is shown to have desirable asymptotic properties and to be generally more powerful than the usual runs test, while preserving computational simplicity.

1. Introduction

We consider a sequence of Bernoulli trials in which each trial results in either a success (S) or a failure (F). The null hypothesis of randomness is that the trials are independent with common probability of success, p. An alternative statement of the null hypothesis, obtained by conditioning on the observed number of successes, is that the successes and failures are randomly distributed (i.e., each permutation is equiprobable).

An example of the type of situation we are considering is the sequence S-S-F-S-F-F-F-F-S-S-S-F-F. The best-known procedure for testing randomness in this setting is the runs test, in which the null hypothesis is rejected if the total number of runs (six in the example) is too small (a clustered arrangement) or too large (a systematic arrangement).

As is widely appreciated, the runs test is relatively insensitive. O'Brien (1976) proposed a procedure based on the number of successes occurring between successive failures, before the first failure, and after the last failure. (Success was defined to be the most common outcome.) The corresponding definition of a run differs from the usual definition in that runs of length 0 are possible. Thus, in the previous example there would be eight success runs having lengths 2, 1, 0, 0, 0, 3, 0, 0. The proposed test statistic was based on the sample variance of run lengths ($s^2 = 1.357$ in the example). Notice that this procedure recognizes the sequence S-S-S-S-S-F-F-F-F-F-S-S-S-S-S-... as a clustered, rather than a systematic, arrangement. The appropriateness of this interpretation is arguable and may well depend on the circumstances of the particular study. In the application that originally motivated the procedure, such arrangements were very unlikely, but the interpretation of clustering would have been appropriate. In fact, relatively few failures were anticipated and the possibility of observing consecutive failures was remote. The procedure was found to enhance power significantly while largely preserving the computational simplicity of the runs test.

In this paper we anticipate applications in which both successes and failures may be sufficiently common to warrant evaluating run lengths of both. We have recently encountered such situations in our study of demyelination of nerve fibers. More generally, runs tests are used in a variety of contexts in which both outcomes occur with nearly equal frequency [e.g., runs above and below the median (Lehmann, 1975, p. 315)]. In these

Key words: Independence; Randomness; Runs.

situations, clustering of failures could well provide the more convincing evidence of a departure from randomness. However, the procedure based only on success run lengths fails to distinguish more than two failures occurring consecutively. For example, in the sequence S-S-F-S-F-F-F-F-S-S-S-F-F, the four consecutive F's enter the test statistic as three success runs of length 0, with no recognition of the fact that they occurred consecutively.

In the procedure which follows we will use the conventional definition of a run, in which runs of length 0 cannot occur. In terms of the example, there are thus three success runs and three failure runs, having lengths 2, 1, 3 and 1, 4, 2, respectively. The test we propose is based on a weighted linear combination of the variances of the run lengths ($s_S^2 = 1.000$ and $s_F^2 = 2.333$). Notice that here the sequence S-S-S-S-S-F-F-F-F-F-S-S-S-S-S-... is recognized as a systematic arrangement.

In order to distinguish between the variance-based procedures, we shall refer to them as one-sample (successes) and two-sample (successes and failures) procedures. Our statement of the two-sample procedure requires the following notation. For successes:

$$n_S = \text{Number of successes},$$

$$r_S = \text{Number of success runs},$$

$$s_S^2 = \text{Sample variance of success run lengths},$$

$$c_S = (r^2 - 1)(r + 2)(r + 3)/[2r(n - r - 1)(n + 1)],$$

$$v_S = cn(n - r)/[r(r + 1)],$$

where, for notational convenience, the subscript S has been deleted in defining c_S and v_S. The notation for failures is defined in similar fashion. The test statistic is

$$\chi_f^2 = c_S s_S^2 + c_F s_F^2$$

where $f = v_S + v_F$. We will reject H_0 in favor of clustering if χ_f^2 exceeds the $1 - \alpha$ percentile of the chi-square distribution with f degrees of freedom (df). (In practice, it is necessary either to round off or to interpolate between tabled integer values of f.)

The accuracy of the chi-square approximation is evaluated for small samples in §4. However, aside from the recommendations provided there, it is apparent from the previous specifications that r_S, r_F, $n_S - r_S$, and $n_F - r_F$ must all exceed 1 for the test to be defined. This condition is assumed in the remainder of the paper.

2. Application to Schwann Cell Disease

The original motivation for the one-sample procedure related to a study of demyelination in human and experimental neuropathy. A few biologic facts and observations are given here to provide background. The myelinated fiber, whose pathologic alteration with disease is being considered, is composed of two different cells: a very elongated process of the nerve cell called the axon and Schwann cells related to the axon. During early development, Schwann cells line up in rows along the axon, and each one wraps itself around a different segment of the same axon and not around other axons. Each Schwann cell forms a segment or internode of myelin. The place where the myelin of two internodes abut is called a node of Ranvier.

In segmental demyelination a portion of an internode of myelin disappears, leaving only a bare axon. We observed that the pattern of demyelination in peripheral neuropathy took two forms. In the one type (e.g., lead neuropathy), the demyelination appeared to be randomly distributed among Schwann cells, possibly providing a reason for thinking that demyelination related to acquired or intrinsic dysfunction of Schwann cells. In another

Table 1
Observed run lengths of myelinated and demyelinated internodes

	Frequency			Frequency	
Run length	Myelinated	Demyelinated	Run length	Myelinated	Demyelinated
1	29	33	7	1	0
2	10	17	8	1	0
3	8	6	9	0	0
4	3	0	10	0	0
5	1	1	11	0	0
6	1	0	12	2	0

$n_S = 135$	$n_F = 90$	
$r_S = 56$	$r_F = 57$	
$s_S^2 = 5.919$	$s_F^2 = .6767$	
$c_S = 9.030$	$c_F = 34.636$	
$v_S = 30.171$	$v_F = 31.116$	
$\chi_f^2 = 76.89$	$\text{df} = 61.29$	$P = .086$

type (e.g., Friedreich's ataxia), the demyelination appeared to be clustered to the Schwann cells related to certain axons. The clustered distribution might therefore provide evidence that demyelination related to an axonal abnormality since the axon was common to the Schwann cells affected. Evidence that one type of segmental demyelination may be due to a primary dysfunction of Schwann cells and another to dysfunction of the axon has now come from other types of studies and has been confirmed experimentally (Dyck et al., 1984). The need to provide validation for the claim that demyelination was clustered in one type and random in the other type led to the approaches discussed here.

We illustrate these considerations with the following example. Listed in Table 1 are observed run lengths of myelinated and demyelinated internodes. Although there is suggestive evidence that the hypothesis of a random distribution may be rejected in favor of multiple clustering ($P = .086$), we note that the usual runs test (counting only the number of runs) provides no evidence of a departure from randomness ($P = .571$, one-sided test).

3. Theoretical Considerations

The derivation that the test procedure controls the size of the test in large samples consists of first conditioning on n_S, n_F, r_S, and r_F. One may then show (see Appendix 1) that, conditionally, (i) $c_S s_S^2$ and $c_F s_F^2$ are independent, (ii) they follow approximately a chi-square distribution, and (iii) the approximations are quite accurate with moderate sample sizes and exact asymptotically. Thus, $c_S s_S^2$ and $c_F s_F^2$ each provide an independent test statistic, raising the question of how best to combine them. In order to assure accurate control over the size of the test, we have chosen to weight the two test statistics equally, thus obtaining approximately a chi-square statistic with $v_S + v_F$ degrees of freedom, with the approximation becoming exact asymptotically. We note that unequal weighting would result in the summation of variates having approximate gamma distributions but with unequal scale parameters. As a consequence, the sum may not be well approximated by a gamma distribution.

Although the decision to weight $c_S s_S^2$ and $c_F s_F^2$ equally was motivated primarily by size considerations, it would be of interest to determine weights which maximize power. We conjecture that optimal weights will depend on the specific alternatives to randomness envisaged, and this would seem to be a fertile area for further theoretical research. It is shown in Appendix 2 that the weights proposed here are fully efficient within a large class of alternatives.

4. Simulations

We conducted a Monte Carlo study to further evaluate the operating characteristics of the two-sample procedure relative to the one-sample test and the usual runs test. In each case, the alternative hypothesis specified clustering, so that the hypothesis of randomness was rejected when the observed number of runs was too small (runs test) or when the observed χ^2 was too large (one- and two-sample procedures). Three situations, all involving samples of size 200, were considered:

Randomness:

$$p_i = .5 \qquad i = 1, \ldots, 200$$

Systematic clustering:

$$p_i = \begin{cases} .6 - .01|50 - i| & i = 1, \ldots, 100 \\ .6 - .01|150 - i| & i = 101, \ldots, 200 \end{cases}$$

Nonsystematic clustering:

$$p_i = \begin{cases} .9 & 90 \leq i \leq 110 \\ .3 & \text{otherwise} \end{cases}$$

The first configuration was chosen to evaluate the ability of the three procedures to control the size of the test. One would expect a priori that the second configuration would be more advantageous to the one-sample procedure than to the two-sample procedure, an advantage which is most apparent in the extreme case of constant run lengths (referred to in the introduction). The advantage is reversed in the third configuration, where the evidence for nonrandomness occurs primarily in the less frequent outcome.

These expectations are confirmed in Table 2, which gives the proportion of times the hypothesis of randomness was rejected under each condition. Each entry is based on 1000 samples, with the same samples used for all three procedures. All methods provided reasonably good control over the size of the test, with the one- and two-sample procedures providing significantly greater power than the usual runs test.

In order to evaluate the adequacy of the chi-square approximation for the two-sample

Table 2
Observed rejection rates

		Level of test (α)			
Condition	Procedure	.10	.05	.025	.01
Randomness	Runs	.102	.062	.035	.013
	1-sample	.096	.052	.033	.013
	2-sample	.093	.051	.032	.017
Systematic clustering	Runs	.503	.361	.272	.172
	1-sample	.653	.552	.477	.370
	2-sample	.537	.433	.341	.246
Nonsystematic clustering	Runs	.401	.274	.179	.104
	1-sample	.548	.445	.386	.327
	2-sample	.714	.616	.545	.451

Table 3
Evaluation of chi-square approximation to the distribution of $c_S s_S^2 + c_F s_F^2$

	$n_S - r_S$			Percentile			
$r_S = r_F$	$= n_F - r_F$	$f =$ df	Distribution	90.0	95.0	97.5	99.0
5	5	10.18	Chi-square $[f]$	16.0	18.3	20.5	23.2
			Chi-square $[f + 1]$	17.3	19.7	21.9	24.7
			Observed distribution	16.88	19.57	21.64	24.18
10	5	16.45	Chi-square $[f]$	23.5	26.3	28.8	32.0
			Chi-square $[f + 1]$	24.8	27.6	30.2	33.4
			Observed distribution	25.59	27.07	31.28	34.72
10	10	14.86	Chi-square $[f]$	21.1	23.7	26.1	29.1
			Chi-square $[f + 1]$	22.3	25.0	27.5	30.6
			Observed distribution	21.97	24.75	27.24	34.50

procedure in small samples, we generated 1000 additional experiments for each of the sample size configurations shown in Table 3. From the theoretical considerations in Appendix 1, accuracy is seen to depend on r_S, r_F, $n_S - r_S$, and $n_F - r_F$. The simulations in Table 3 indicate that the approximation is quite adequate (except in extreme tail areas) when all four of these parameters are greater than or equal to 10. For values between 5 and 10, it appears that a better approximation is obtained by using the integer part of the degrees of freedom plus one, $[f + 1]$.

Dixon (1940) showed that the chi-square approximation is adequate for the one-sample problem if the number of successes and success runs both exceed 10. The results in Table 3 indicate that analogous limits hold for the two-sample procedure when applied to both successes and failures. In turn, O'Brien (1976) found that the one-sample procedure using the chi-square approximation provided more accurate control over the size of the test than the usual runs test based on the normal approximation. Thus, it may be inferred that the same result holds vis-à-vis the two-sample procedure and the runs test.

5. Discussion

Our purpose has been to obtain a simple modification of the usual runs test with a view toward increasing efficiency by utilizing the information contained in the run lengths. The proposed procedure appears to provide adequate control over the size of the test in moderate sample sizes and provides improved efficiency, at very little additional computational expense.

Although the test may be motivated by theoretical efficiency considerations, it is proposed primarily as a simple omnibus procedure for use when the specific alternative to randomness is unknown. Conversely, when a specific alternative hypothesis can be stated, it is generally recognized that procedures which incorporate this additional information may provide significant gains in power.

In a somewhat different context, the proposed procedure may be viewed as the third in a sequence of procedures for testing randomness. O'Brien (1976) proposed a procedure based on the index of dispersion for testing whether events occurred randomly during continuous time (i.e., testing for a common exponential distribution). In the same paper, this (one-sample) procedure was generalized to discrete units of time (e.g., days), essentially testing for a common geometric distribution. Although the probability of an event occurring on consecutive days was considered to be positive, it was taken to be sufficiently small to warrant consideration only of interval lengths between successive occurrences. For the two-sample procedure proposed here, we envisage applications in which both successes and failures are sufficiently probable to warrant consideration of run lengths of both types of events.

The two-sample procedure may also be generalized in a straightforward manner to accommodate polychotomous outcomes. With the obvious modifications in notation, a test that each of K different types of letters are distributed randomly may be based on $\chi^2 = \Sigma c_i s_i^2$. That the test criterion has approximately a central chi-square distribution with Σv_i degrees of freedom under the null hypothesis follows directly from the fact that the individual statistics $c_i s_i^2$ $(i = 1, \ldots, K)$ are mutually independent and asymptotically chi-square with v_i degrees of freedom.

Résumé

On propose une méthode pour tester l'hypothèse que des essais de Bernoulli (succès et échecs) sont indépendants avec la même probabilité de succès. De manière équivalente, on peut utiliser cette méthode pour tester l'hypothèse que l'ordre dans lequel on obtient un nombre fixé de succès et d'échecs est aléatoire. La méthode s'appuie sur une combinaison linéaire pondérée des variances des longueurs de runs (suites continues) de succès, et d'échecs. On montre qu'elle possède des propriétés asymptotiques souhaitables, et qu'elle est en général plus puissante que le test de runs habituel, sans augmentation de la complexité des calculs.

References

Dixon, W. J. (1940). A criterion for testing that two samples are from the same population. *Annals of Mathematical Statistics* **11,** 199–204.

Dyck, P. J., Nukada, H., Lais, A. C., and Karnes, J. L. (1984). Permanent axotomy: A model of chronic neuronal degeneration preceded by axonal atrophy, myelin remodeling and degeneration. In *Peripheral Neuropathy, Vol.* 1, 2nd ed. Chapter 49, 1103–1138. P. J. Dyck, P. K. Thomas, E. H. Lambert, and R. Bunge (eds). Philadelphia: W. B. Saunders.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day.

O'Brien, P. C. (1976). A test for randomness. *Biometrics* **32,** 391–401.

Received March 1983; *revised September* 1983 *and April* 1984.

Appendix 1

In this appendix we supply the details of the derivation of the null distribution of the test statistic, conditional on n_S, n_F, r_S, and r_F. Consider first the conditional distribution of s_S^2. Construct a new sequence of successes and failures as follows: (i) Omit all outcomes up to and including the first success, (ii) omit all failures (if any) occurring after the last success, and (iii) replace each remaining failure run and subsequent success with a single failure. In this new sequence, let n_F^* represent the number of failures; Y_i, the number of successes occurring after the $(i - 1)$th failure but prior to the ith failure $(i = 2, \ldots, n_F^*)$; Y_1, the number of successes preceding the first failure; and $Y^*_{n_F+1}$, the number of successes following the last failure. Notice that in the new sequence, $n_F^* = r_S - 1$ and the number of successes $(n_S^* = \Sigma Y_i)$ is given by $n_S^* = n_S - r_S$.

Let $\tilde{s}_S^2$ represent the sample variance of the Y values. Since the rule for obtaining the Y values had the effect of reducing the length of each success run by 1, the sample variance of run lengths in the new sequence is identically the same as in the old. That is, $\tilde{s}_S^2$ is identically equal to s_S^2, so that it is sufficient to consider the distribution of $\tilde{s}_S^2$.

To facilitate the correspondence between the notation here and that used by Dixon (1940) we note that, when expressed in Dixon's notation, s_S^2 has expectation

$$m(m + n + 1)/[(n + 1)(n + 2)]$$

and variance

$$4m(m - 1)(m + n + 1)(m + n + 2)/[n(n + 2)^2(n + 3)(n + 4)].$$

These expressions, become, respectively,

$$\mathrm{E}_S = (n_S - r_S)n_S/[r_S(r_S + 1)]$$

and

$$\mathrm{V_S} = 4(n_\mathrm{S} - r_\mathrm{S})(n_\mathrm{S} - r_\mathrm{S} - 1)n_\mathrm{S}(n_\mathrm{S} + 1)/[(r_\mathrm{S} - 1)(r_\mathrm{S} + 1)^2(r_\mathrm{S} + 2)(r_\mathrm{S} + 3)]$$

when $n^*_\mathrm{S} = n_\mathrm{S} - r_\mathrm{S}$ is substituted for Dixon's m and $n^*_\mathrm{F} = r_\mathrm{S} - 1$ is substituted for Dixon's n. (There are analogous expressions for $\mathrm{E_F}$ and $\mathrm{V_F}$.)

Following the development in O'Brien (1976) and Dixon (1940), we have the conditional distribution of $Y_1, \ldots, Y^*_{n_\mathrm{F}+1}$ given ΣY:

$$f_Y = \int \frac{(\Sigma Y_i)!}{\Pi Y_i!} (\Pi u_i^{Y_i}) \, \Pi du_i,$$

with integration over the generalized tetrahedron defined by $u_i > 0$, $i = 1, \ldots, n^*_\mathrm{F}$, and $\Sigma u_i \leqslant 1$. Furthermore, after obtaining the necessary correspondence in notation among O'Brien (1976), the new sequence of trials introduced in this section, and the original sequence in §1, we have (i) $c_\mathrm{S}\tilde{s}^2_\mathrm{S}$ has approximately a chi-square distribution, (ii) the approximation is quite accurate in small samples (r_S and $n_\mathrm{S} - r_\mathrm{S} \geqslant 10$), and (iii) the approximation is exact asymptotically. The same results are obtained for failure run lengths (denoted by X) using a completely analogous argument.

It should be noted that the same expression for f_Y obtains if we also condition on the individual failure run lengths, since the F's omitted in obtaining the new sequence do not enter the calculation for f_Y. Stated algebraically, $f_{Y|X} \equiv f_{YX}/f_X = f_Y$, so that $f_{YX} = f_X f_Y$. Thus, conditional on r_S, r_F, n_S, and n_F, the failure run lengths and success run lengths are statistically independent.

Appendix 2

In order to evaluate the efficiency of the proposed procedure, we first establish the following result.

Theorem. Let $X_{ij}(i = 1, 2; j = 1, \ldots, n)$ represent a sequence of mutually independent random variables with variance σ_i^2 and distribution function F_i. To test the null hypothesis H_0: $\sigma_i^2 = \sigma_{i0}^2$ against the alternative H_A: $\sigma_i^2 = \lambda\sigma_{i0}^2$ $(i = 1, 2)$, consider the family of conditional tests defined by

$$R = \{(w_1, w_2, c): \quad \Pr(w_1 s_1^2 + w_2 s_2^2 > c \,|\, H_0, \bar{X}_1, \bar{X}_2) = \alpha\}.$$

Let r_0 represent the member of R for which

$$(w_1, w_2, c) = (\ (n-1)/\sigma_{10}^2, \quad (n-1)/\sigma_{20}^2, \quad c_{0n}\ ),$$

where c_{0n} is chosen to maintain the size of the test.

1. If F_1 and F_2 are normal distributions, then r_0 is a uniformly most powerful (UMP) conditional test.
2. If F_1 and F_2 have finite fourth moments and r' is any other member of R, then the (Bahadur and Pitman) asymptotic relative efficiency (ARE) of r_0 relative to r' is greater than or equal to 1. Furthermore, $\lim_{n\to\infty} c_n/\chi^2_{2n-2} = 1$.

The proof of the first part of the theorem consists of noting that if the $\{X_{ij}\}$ are normally distributed, then the conditional distribution given the sample means is proportional to $\exp[-(n-1)\Sigma s_i^2/(2\sigma_i^2)]$. Thus, by direct application of the Neyman–Pearson lemma, the UMP test criterion is given by

$$\frac{L_0}{L_1} \propto \exp\left\{-n(n-1)\left[\Sigma\left(\frac{s_i^2}{2\sigma_{i0}^2}\right) - \Sigma\left(\frac{s_i^2}{2\lambda\sigma_{i0}^2}\right)\right]\right\}$$

$$\ln\left(\frac{L_0}{L_1}\right) \propto \left(\frac{s_1^2}{\sigma_{10}^2} + \frac{s_2^2}{\sigma_{20}^2}\right).$$

To establish the second part of the theorem we note that the conditions imposed on F_1 and F_2 are sufficient to ensure that s_1^2 and s_2^2 obey the central limit theorem (see appendix of O'Brien, 1976). Thus, the limiting distribution of the test criterion for any member of R is independent of the underlying distributions F_1 and F_2. It follows that the first conclusion of the theorem holds asymptotically for any F_1 and F_2 with finite fourth moments, so that no other member of R can have greater asymptotic efficiency than r_0.

Corollary. Consider conditional tests of the hypothesis of randomness as defined in §1, and let σ^2_{S0} and σ^2_{F0} represent the null variance of run lengths for successes and failures, respectively. For alternatives in which $\sigma^2_{SA} = \lambda\sigma^2_{S0}$ and $\sigma^2_{FA} = \lambda\sigma^2_{F0}$, the two-sample runs test has ARE $\geqslant 1$ relative to any other linear combination of sample variances.

Note first that asymptotically the run lengths (X_{ij}) follow independent geometric distributions under H_0. Asymptotically, therefore, since the n_S successes comprise r_S runs, the expected run length is given by $1/p_S = n_S/r_S$ and the variance of the run lengths is $(1 - p_S)/p^2_S = (n_S - r_S)n_S/r^2_S$. If we let $\sigma^2_{10} = (n_S - r_S)n_S/r^2_S$ and $\sigma^2_{20} = (n_F - r_F)n_F/r^2_F$, then the corresponding optimal criterion in the theorem is given by $w_S s^2_S + w_F s^2_S$, where w_S and w_F are any constants such that

$$\begin{aligned} w_S/w_F &= \sigma^2_{20}/\sigma^2_{10} \\ &\approx (n_F - r_F)n_F/[(n_S - r_S)n_S] \\ &\approx c_S/c_F. \end{aligned}$$

The approximations are exact asymptotically and consist only of noting that r_S and r_F can differ at most by 1 and their ratio equals 1 asymptotically.