Detecting Patterns by One-Sample Runs Test: Paradox, Explanation, and a New Omnibus Procedure

Author(s): Dean P. McKenzie, Patrick Onghena, Robert Hogenraad, Colin Martindale and Andrew J. MacKinnon

Source: *The Journal of Experimental Education*, Vol. 67, No. 2 (Winter, 1999), pp. 167-179
Published by: Taylor & Francis, Ltd.
Stable URL: https://www.jstor.org/stable/20152591
Accessed: 17-07-2019 09:26 UTC

The Journal of Experimental Education, *1999*, *67*(2), 167–179

Detecting Patterns by One-Sample Runs Test: Paradox, Explanation, and a New Omnibus Procedure

DEAN P. McKENZIE
Monash University & Mental Health Research Institute
Melbourne, Victoria, Australia

PATRICK ONGHENA
Katholieke Universiteit Leuven
Leuven, Belgium

ROBERT HOGENRAAD
Université Catholique de Louvain
Louvain-la-Neuve, Belgium

COLIN MARTINDALE
University of Maine

ANDREW J. MACKINNON
Monash University & Mental Health Research Institute
Melbourne, Victoria, Australia

ABSTRACT. The standard nonparametric 1-sample runs test, which is based on the total number of runs, has recently been found to give anomalous results in the case of data consisting solely of runs of length 2. Furthermore, tests based on the total number of runs provide little information as to the length of the individual runs. An explanation for that anomaly is provided, and tests based on the length of runs are described. Such tests can be successfully applied to runs of length 2 but usually use asymptotic methods of assessing statistical significance. In addition, such tests generally impose arbitrary limits as to the maximum run length. A new procedure that allows the maximum run length to be empirically determined is described. In that procedure, one uses a Monte Carlo permutation test to ascertain statistical significance. The new Monte Carlo omnibus length-of-runs procedure is illustrated with examples from suicide research and from psycholinguistics.

RESEARCHERS ARE OFTEN CONCERNED with determining whether observed sequences contain patterns. One form of pattern is a run, or repetition of like elements in a series of observations. As defined by Siegel and Castellan

(1988), "A run is defined as a succession of identical symbols which are followed and preceded by different symbols or by no symbols at all" (p. 58). Thus, in a series of coin flips (H = heads, T = tails) HTTTTTHTT, there are two runs of length one for heads (the first H is preceded by no symbol and followed by a different symbol T; the second H is both preceded and followed by a T), one run of length two for tails (preceded by an H and followed by no symbol at all), and one run of length five for tails (both preceded and followed by an H).

The authors of several reviews (e.g., Hazell, 1993; Jobes, Berman, O'Carroll, Eastgard, & Knickmeyer, 1996; Phillips, 1985) have examined whether occurrences of suicide occur randomly over time. The alternative hypothesis is that media reports of suicide induce copycat or imitative behaviors, leading to suicides occurring in clumps or runs. Phillips (1974) termed such a process the *Werther effect*, after Johann Wolfgang von Goethe's novella *Die Leiden des jungen Werthers* (The sorrows of young Werther), which ends with the protagonist committing suicide. Publication of the novella in 1774 was rumored to have been followed by a large number of suicides in Europe (Phillips, 1974, 1985).

To test the hypothesis described above, Church and Phillips (1984) searched for clumps or patterns in a particular form of suicide—asphyxiation by means of a plastic bag. If the series is binary (e.g., a suicide did or did not occur during a particular time segment), then one can generally use the standard one-sample runs test to examine clumps or runs. That test, originally developed by von Bortkiewicz (1917), Mood (1940), and Stevens (1939), among others, is based on the total number of runs of length one or higher (e.g., Siegel & Castellan, 1988).

The one-sample runs test is concerned with whether there are fewer or more runs than would be expected by chance. To proceed with the opening example, Church and Phillips (1984) analyzed suicides in Kingston-on-Hull, England, between 1971 and 1981. They first divided the series of observations into sixty 2-month segments. A plastic bag asphyxiation suicide did not occur in 51 segments and did occur in 9 segments.

The expected number of runs of both suicides and nonsuicides can be calculated as (Lentner, 1982, Formula 894)

$$(2mn/N) + 1, \tag{1}$$

where m represents the number of occurrences of one category (e.g., suicides), n represents the number of occurrences of the other category, and $N = m + n$. The expected number of runs was therefore 16.3. The number of runs (of both suicides and nonsuicides) found by Church and Phillips (1984) was 11. They found that the exact probability that 11 or fewer runs would occur was .0142. In other words, there were fewer runs than would be expected under the null hypothesis. Such a finding may reflect clumping or clustering of that particular form of suicide.

The runs test is implemented in general statistics packages such as SPSS (SPSS Inc., 1997) and Stata (Stata Corporation, 1997); exact significance tests are available in MULTRUNS (Koppen & Verhelst, 1986), SPSS Exact Tests (SPSS Inc. 1997), and StatXact (Cytel Software Corporation, 1995). In addition to the suicide study cited previously (Church & Phillips, 1984), the runs test has been used as an adjunct to data smoothing (Bradley, Steil, & Bergman, 1995) and in the fields of genetics (Feil, Zhou, Maynard-Smith, & Spratt, 1996), image processing (Chuang & Huang, 1993), narrative analysis (Dorval, 1986; Frautschi & Thoiron, 1993), and neurophysiology (Rahamimoff, Edry-Schiller, Rubin-Fraenkel, Butkovich, & Ginsbury, 1995). The runs test has also been applied in the fields of psychopharmacology (Rump, Siekmann, Dreja, & Kalff, 1997), regression diagnostics (Mendenhall, Reinmuth, & Beaver, 1993; but see Huitema, McKean, & Zhao, 1996), research productivity (Katerndahl, 1996), response validity (Cliffe, 1992; Huba, 1986), and the testing of random number generators (Barry, 1996; Onghena, 1993; Tries, 1997).

The Runs Test: An Anomaly

Mogull (1994) recently described an anomaly in the standard one-sample runs test. When used with both exact tabled (Lentner, 1982; Swed & Eisenhart, 1943) and asymptotic (Mood, 1940; Siegel & Castellan, 1988) statistical significance values, the one-sample runs test is not able to reject the null hypothesis for a data series consisting exclusively of runs of length two. An example of such a series is HHTTHHTTHHTTHHTTHHTTHHTT, where *H* denotes heads and *T* denotes tails. Mogull (1994) provided several hypothetical examples of such runs, including students' copying answers from other students' examination papers.

In terms of the suicide example, it is possible that two time periods in which suicides did not occur were immediately followed by two time periods in which suicides did occur. However, we preferred to use a hypothetical example based on the simple tossing of coins. That activity still commonly appears in discussions of elementary probability (e.g., Konold, 1995).

Although betting on the outcome of a single coin is an informal activity at best, Australian casinos offer gamblers the traditional game of two-up (Hill, 1995), which involves the tossing of two coins. The main participant, known as the spinner, seeks to throw three consecutive pairs of heads. Other players bet on both coins turning up heads or both coins turning up tails. The series of 24 coin tosses shown previously is clearly not random, something that could be exploited by our hypothetical gamblers. However, an exact runs test, performed using SPSS Exact Tests (SPSS Inc., 1997), established that the probability of obtaining 12 runs in 24 observations or coin tosses is .8421, with a 95% confidence interval (Lentner, 1982, p. 165) of 7 to 19 runs. On the basis of a standard one-sample runs test, therefore, the above sequence cannot be distinguished from a random process.

Mogull (1994) has also shown that the runs test is not able to reject the null hypothesis of no pattern, even if a data series consisting of one million runs of length two is analyzed. The power of the runs test for detecting patterns in runs of length two actually decreases for large samples. When applied to runs of length one (e.g., HTHT . . .), three (e.g., HHHTTT . . .), and four (e.g., HHHHTTTT . . .), however, the runs test was successfully able to reject the null hypothesis.

Because Mogull (1994) did not provide an explanation for the above anomaly, in the present article we examined formulas for the expected total number of runs and the expected number of runs of length two.

In the case of equal numbers of heads and tails, the exact expected total number of runs (*tr*) is given by (e.g., Lentner, 1982, Formula 894)

$$tr = (2n^2/N) + 1, \qquad (2)$$

where n is the number of either heads or tails and N is the total numbers of heads and tails, as defined earlier. The above formula can be further simplified as follows:

$$tr = (N/2) + 1. \qquad (3)$$

The maximum number of runs of length two is equal to $N/2$, and so it can be seen that the observed ($N/2$) and the expected ($N/2$) + 1 values are about equal, giving rise to the anomaly. For no other run length is the expected number of runs of length (L)—N/L— equal, or very close, to the total expected number of runs. That anomaly can be explained by the fact that the standard one-sample runs test is sensitive only to an aberrant number of runs. Other patterns in the data, such as those introduced by systematic runs of length two, may not be able to be detected. In the present study, we were concerned with the development of a type of runs test that is able to detect runs of length two as well as other lengths. We validated such a test on random, as well as actual, data.

Length of Runs

Although the test based on the total number of runs is arguably the most well known, there are several other forms of one-sample runs tests (e.g., Agin & Godbole, 1992; Gibbons & Chakraborti, 1992; Lentner, 1982; Potthoff, 1989). For instance, Levene and Wolfowitz (1944) and Mood (1940) described tests involving specific run lengths.

In the running example involving 24 observations, there were 12 observed (o) runs of length two. The expected (e) mean number of runs of length (x) can be calculated by the asymptotic or large sample formula provided by Lentner (1982, Formula 889), Harrison and Tamaschke (1993), Mood (1940), and von Bortkiewicz (1917):

$$e = Np^xq^2, \qquad (4)$$

where N is the total number of heads and tails (e.g., 24), p is the proportion of heads (e.g., .50), q is the proportion of tails (e.g., .50), and x is the run length (e.g., 2).

The asymptotic variance (v) (e.g., Lentner, 1982, Formula 889) can be estimated by

$$v = e\{1 - e/N[(x^2/p + 2/q) - (x + 1)^2]\}. \quad (5)$$

The above formulas are intended for runs of either heads or tails, but not both. Because the number of heads and tails in the running example are equal, both e and v were multiplied by 2. Note that the doubling of the variances incorrectly assumes that runs of heads and runs of tails are independent, which they are not. For example, if there were 12 runs of heads observed in 24 coin tosses, there must also be 12 runs of tails, if heads and tails occur with equal probability. The variance estimate and the statistical significance test in which it is used here are therefore only approximations and were used only for the sake of the present example. The z statistic obtained using the formula (e.g., Lentner, 1982; Formula 885)

$$z = (o - e)/\sqrt{v} \quad (6)$$

was 3.0984, with a one-tailed probability of .0010. Thus, an approximate test based on runs of a specific length was correctly able to reject the null hypothesis.

Omnibus Tests of Run Lengths

Application of a test concerned with a specific run length normally requires the researcher to choose the run length in advance. However, Wallis and Moore (1941) introduced an omnibus test based on an examination of several run lengths. Such a test was originally intended to be used with data that were measured on at least an ordinal scale, although later variations of the test (Knuth, 1981) require an interval scale. In terms of the latter reference, observations that are higher in value than the preceding observation form part of a run up and observations that are lower than the preceding observation form part of a run down. In the case of the previous example of 12 runs of length two, there were 12 runs up of length one (12 occurrences of an *H* followed by a *T*) and 6 runs up of length two (six occurrences of *H* followed by *TT*).

Wallis and Moore (1941) used the standard asymptotic chi-square test (Siegel & Castellan, 1988) to establish statistical significance. To avoid the problem of small expected cell frequencies (Mehta, 1994), they required the maximum run length of either runs up or runs down to be three. Runs of greater length were truncated to a length of three. Even with that constraint, Wallis and Moore found that the standard chi-square test was overly liberal in rejecting the null hypothesis as a result of the lack of independence between run lengths. In other words, if two runs of length 12 were observed in 24 observations, all other run lengths would have zero probability of occurrence.

Gustafson, Dawson, Nielsen, and Caeli (1994) and Harrison and Tamaschke (1993) described similar omnibus tests based on run lengths. The latter test requires the maximum run length to be six, and the former test allows the maximum run length to be determined by the data. Knuth (1981) also described an omnibus test that allows the maximum run length to equal the longest run that is found in a particular set of data. That test, in which one attempts to take the correlations between different run lengths into account, was implemented in the DRUNS/RUNS subroutine of IMSL (1991). However, the latter implementation was specifically intended for continuous data, and the maximum run length was required to be specified in advance. Finally, Knuth (1981) intended that test to be applied to large samples, consisting of at least several thousand observations.

A Monte Carlo Permutation Test

Existing omnibus length-of-runs tests either place arbitrary limits on the maximum run length or are intended to be used with continuous data, or both. We required a procedure that could be applied to binary data and would allow the maximum run length to be determined empirically. It is a straightforward matter to compute a chi-square statistic measuring the difference between the observed and expected numbers of runs of each length. Because of possible problems with small expected frequencies and dependence between run lengths (Wallis & Moore, 1941), however, we did not wish to rely on the chi-square distribution to assess statistical significance.

Good (1994) and McKenzie et al. (1996) described general nonparametric procedures for ascertaining the statistical significance of novel or modified test statistics. One such procedure is the Monte Carlo permutation test (e.g., Edgington, 1995; Good, 1994; Manly, 1997; Mehta, 1994). In permutation tests, one assumes that any pattern or ordering of data (e.g., HHTT . . . , HTHT . . . , HTTH . . .) is possible under the null hypothesis (Good, 1994). Monte Carlo permutation tests involve a sample of all the possible orderings being selected. The data must be randomly shuffled or permuted a large number of times (a value of 999 was used by McKenzie et al., 1996).

Monte Carlo permutation tests based on the total number of runs are implemented in SPSS Exact Tests (SPSS Inc., 1997) and StatXact (Cytel Software Corporation, 1995). With regard to the omnibus length-of-runs test, a chi-square statistic is computed for each Monte Carlo permutation. That statistic represents the difference between the observed and expected number of runs of each length. The formula for calculating the expected number is provided in Equation 4. As mentioned earlier, the formula is intended for the calculation of the expected number of runs of a single element. To calculate the number of runs of either element, one adds together the expected number of runs of heads and the expected number of runs of tails for each run length.

The number of times that the original chi-square is equalled or exceeded by the chi-squares for the randomly permuted samples is then recorded. That count, plus 1, is then divided by 1,000 (999 random permutations plus 1). The addition of 1 to the numerator and the denominator allows for the original permutation or ordering of the raw data (as discussed by Onghena & May, 1995). If the resulting value, expressed as a proportion, is less than or equal to .05, then the result is statistically significant at the .05 level.

With regard to the running example involving 24 coin tosses, only one (the obtained data series) permutation consisted of 12 runs of length two. The probability was therefore 1/1,000 or .001, which in this case is almost identical (to four-digit precision) to the probability obtained using the approximate test for runs of a specific length described above.

Validation

To examine the Type I (false positive) error rate of the omnibus length-of-runs test, one must perform a simulation study. For each of seven sample sizes (25, 50, 75, 100, 250, 500, and 1000), we used the RNUND (IMSL, 1991) pseudorandom uniform random number generator to generate 3,000 binary datasets. Simulations (and applications to real data) were carried out on an Intel Pentium 200 MMX microprocessor with 32 megabytes of RAM, with custom FORTRAN (Microsoft Corporation, 1995) software. We used two methods of assessing statistical significance of the chi-square goodness of fit—999 Monte Carlo permutations (MCP) and the standard asymptotic chi-square (ACS) test suggested by Gustafson et al. (1994) and Harrison and Tamaschke (1993).

Because the omnibus length-of-runs test is intended to be used for runs of both elements (heads and tails) and single elements (heads or tails), we performed two sets of simulations for each method of assessing statistical significance. In the first set, runs of heads and runs of tails were both counted. In the second set, only runs of heads were counted. We used a pseudorandom number generator to generate the data; therefore, approximately 5% of the results should be statistically significant at the .05 level.

Results

The Monte Carlo permutation length-of-runs test exhibited a satisfactory Type I error rate for all seven sample sizes, for runs of heads or tails as well as runs of heads only (see Table 1). However, the standard chi-square test was grossly liberal for all sample sizes, but particularly in the case of runs of either heads or tails, for sample sizes of 100 or more.

Runs of Length Two

We first applied the omnibus length-of-runs test (with 999 random permutations used throughout) to the ongoing example of 12 runs of length two. With the

TABLE 1
Type I Error Rates of the Monte Carlo Permutation (MCP) Length-of-Runs Test and the Asymptotic Chi-Square Test (ACS)

	MCP		ACS	
N	Heads or tails	Heads	Heads or tails	Heads
25	.049	.048	.172	.186
50	.052	.052	.186	.166
75	.046	.050	.178	.159
100	.046	.045	.195	.165
250	.051	.051	.194	.169
500	.055	.055	.207	.179
1,000	.052	.053	.209	.176

new test, applied to both runs of heads and runs of tails, there were no runs of length one (six expected) and 12 runs of length two (three expected). The overall result was statistically significant ($\chi^2 = 33.0$, $p = .009$). Note that because a permutation distribution was used, rather than the theoretical chi-square distribution, the concept of degrees of freedom is not relevant.

Clusters of Suicides

The results of the length-of-runs test applied to the data on suicide by plastic bag asphyxiation (Church & Phillips, 1984) are given in Table 2. The test was applied only to runs of suicides, rather than nonsuicides. In addition, we used a Monte Carlo permutation version of the standard one-sample runs test similar to that described by Manly (1997).

We found a total of five runs of suicides. Probabilities for number of runs less than or equal to five and greater than or equal to five are reported by the computer program, in a similar fashion to StatXact (Cytel Software Corporation, 1995; see Table 1). Although statistically significant ($p = .018$ for total number of runs ≥ 5), that result tells us very little about the type of patterns in the data.

TABLE 2
Clusters of Suicide (Church & Phillips, 1984)

Run length	Observed *n*	Expected *n*
1	3.000	6.503
2	0.000	.975
3	2.000	.146

Note. $\chi^2 = 26.348$, $p = 0.01600$. Total runs = 5, $p = 0.99900$ (≤ 5), 0.018 (≥ 5).

A test based on the total number of runs does not by itself provide much information as to whether runs tend to be long or short. By examining the results of the length-of-runs test, one can see that fewer isolated occurrences of suicide (3) than would be expected by chance (6.503) were found. On the other hand, we observed a greater number of runs of three suicides (2) than would be expected (.146) by chance. Such a result was statistically significant ($\chi^2 = 26.348$, $p = .016$). Our finding offers positive support for the original observation of Church and Phillips (1984)—that there appears to be clustering of suicides within the Kingston-on-Hull data. However, the length of the time intervals during which suicides occur may have bearing on the results (e.g., Kirch & Lester, 1986).

Vowels and Consonants

Current psycholinguistic (e.g., Martindale, 1991) and speech processing (e.g., Markvitz, 1995) research is concerned with the understanding of graphemes or phonemes rather than vowels and consonants. However, because the latter elements are binary, one can readily analyze them by using the length-of-runs test. In two papers published in *The American Journal of Psychology,* Newman (1951a, 1951b) analyzed patterns of vowels and consonants in short segments (1,000 characters) of biblical text in various languages. Bishop, Fienberg, and Holland (1975) later used Markov chains (e.g., Gottman & Roy, 1990) to reanalyze Newman's data. Interestingly, Markov (1913/1955) used counts of the vowels and consonants found in a sample of Russian verse as a demonstration of the statistical method that now bears his name (Gani, 1985). Bishop et al. (1975) established the presence of a third-order Markov chain for King James English. In other words, the probability of a written character being a vowel or a consonant is dependent on whether the preceding three characters are vowels or consonants.

For purposes of demonstration, we applied the Monte Carlo omnibus length-of-runs and standard one-sample, or total-number-of-runs, tests to the first 1,000 characters (not including the preface, chapter title, or opening quotation) of a machine readable (World Library, Inc., 1994) edition of a recent American English text (Stannard, 1980). For the sake of convenience, we adopted the early cryptanalytic (Gaines, 1956) convention of treating the letter *y* as a vowel. The proportion of vowels was .40, identical to that found in the analysis of larger amounts of English text (Gaines, 1956). The results for runs of vowels are given in Table 3.

There were a total of 325 runs found in the data ($p = .001$, for total number of runs ≥ 325). By examining the length-of-runs test, one can see that single vowels (runs of length one) were more common than would be expected under the null hypothesis (observed = 261, expected = 144.353). On the other hand, vowel combinations (in particular triplets, or runs of length three) were less common

TABLE 3
Clusters of Vowels in a Sample of American English Text

Run length	Observed n	Expected n
1	261	144.353
2	57	57.308
3	6	22.751
4	1	9.032

Note. $\chi^2 = 113.737$, $p = 0.00900$. Total runs = 325, $p = 1.00000$ (≤ 325), 0.001 (≥ 325).

than would be expected if the series were random. It should be noted that the above analysis, as well as those described by Bishop et al. (1975) and Newman (1951a, 1951b), did not delineate character sequences into individual words. Runs of three, or even four, vowels were therefore possible (e.g., a particular word may end in two vowels, and the very next word may begin with a vowel).

Although based on a comparatively small segment of one particular text, the above result tallies with Menzerath's (1950) finding that the most common vowel (*V*) consonant (*C*) sequences in English—CVC, CVCC, and CCVC—involve single vowels. The omnibus length-of-runs test was statistically significant ($\chi^2 = 113.737$, $p = .009$).

Discussion

Asymptotic chi-square significance tests have been found to give satisfactory results in the case of runs of particular elements (Koppen & Verhelst, 1986). However, the results of our simulation study clearly show that such methods should not be used with an omnibus test based on run lengths. Our results confirm the original findings of Wallis and Moore (1941) and call into question the validity of the length-of-runs procedures described by Gustafson et al. (1994) and Harrison and Tamaschke (1993).

The Monte Carlo omnibus length-of-runs procedure exhibited a satisfactory Type I error rate on simulated data. Unlike the standard one-sample or total-number-of-runs test, the above procedure was correctly able to reject the null hypothesis in the case of runs of length two. Although formal power studies are required, the length-of-runs test appears to give a smaller p value than does the total-number-of-runs test for the suicide cluster example, but a higher p value than the total-number-of-runs test in the case of the vowels and consonants example. In both of the above applications, as well as the runs-of-length-two example, the length-of-runs test provided information that was not supplied by the total-number-of-runs test.

We intend to apply both the omnibus length-of-runs test and the total-number-of-runs test to data concerned with forced-choice decision tasks for schizophrenic patients. On the basis of prior research (e.g., Lyon, Lyon, & Magnusson, 1994; Paulus, Geyer, & Braff, 1996), both very short ("chaotic" responses) and very long (perseverative responses) runs would be expected for an individual schizophrenic patient.

The present omnibus length-of-runs test is intended for use as an adjunct to the standard one-sample or total-number-of-runs test, as well as traditional (e.g., Knuth, 1981; Onghena, 1993) and newly developed (e.g., Davies, Dawson, Gustafson, & Pettit, 1995; Gentleman, 1994; Lyon et al., 1994; Paulus et al., 1996; Pincus & Kalman, 1997) tests of randomness. A Windows 95 microcomputer program for performing Monte Carlo permutation tests based on the length of runs, as well as the total number of runs, has been developed. That program may be obtained from the first author.

NOTE

The authors gratefully acknowledge Norman Verhelst for his comments made on an earlier version of this article.

Address correspondence to Dean P. McKenzie, Victorian Transcultural Psychiatry Unit, 2nd floor, Bolte Wing, St. Vincent's Hospital, Nicholson Street, Fitzroy, Victoria, Australia 3065. E-mail: deanm@vaxc.cc.monash.edu.au.

REFERENCES

Agin, M. A., & Godbole, A. P. (1992). A new exact runs test for randomness. In C. Page & R. LePage (Eds.), *Computing science and statistics: Proceedings of the 22nd symposium on the interface* (pp. 281–285). New York: Springer-Verlag.

Barry, T. M. (1996). Recommendations on the testing and use of pseudo-random number generators used in Monte Carlo analysis for risk assessment. *Risk Analysis, 16,* 93–105.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Bortkiewicz, L. von. (1917). *Die Iterationen.* Berlin: Springer.

Bradley, D. C., Steil, G. M., & Bergman, R. N. (1995). OOPSEG: A data smoothing program for quantitation and isolation of random measurement error. *Computer Methods and Programs in Biomedicine, 46,* 67–77.

Chuang, K. S., & Huang, H. K. (1993). Evaluation of inherent gray-level dynamic range in a digital image using the runs test and join-count statistics. *Medical Physics, 20,* 39–45.

Church, I. C., & Phillips, J. P. N. (1984). Suggestion and suicide by plastic bag asphyxia. *British Journal of Psychiatry, 144,* 100–101.

Cliffe, M. J. (1992). Symptom-validity testing of feigned sensory or memory deficits: A further elaboration for subjects who understand the rationale. *British Journal of Clinical Psychology, 31,* 207–209.

Cytel Software Corporation. (1995) StatXact-3 (Version 3.0.1) [Computer program]. Cambridge, MA: Author.

Davies, N., Dawson, E., Gustafson, H., & Pettit, A. N. (1995). Testing for randomness in stream ciphers using the binary deriatives. *Statistics and Computing, 5,* 307–310.

Dorval, B. (1986). Analyzing runs of an event occurring in a series with other kinds of events. *Behavior Research Methods, Instruments and Computers, 18,* 10–16.

Edgington, E. S. (1995). *Randomization tests* (3rd ed). New York: Marcel Dekker.

Feil, E., Zhou, J., Maynard-Smith, J., & Spratt, B. G. (1996). A comparison of the nucleotide sequences of the adk and recA genes of pathogenic and commensal Neisseria species: Evidence for extensive interspecies recombination within adk. *Journal of Molecular Evolution, 43,* 631–640.

Frautschi, R., & Thoiron, P. (1993). Toward a narra-topography: A pilot study applied to Marguerite Duras' novel *Moderato Cantabile. Computers and the Humanities, 27,* 235–248.

Gaines, H. F. (1956). *Cryptanalysis: A study of ciphers and their solutions.* New York: Dover.

Gani, J. (1985). Literature and statistics. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 5, pp. 90–95). New York: Wiley.

Gentleman, J. F. (1994). The distribution of the frequency of subsequences in alphabetic sequences, as exemplified by deoxyribonucleic acid. *Applied Statistics, 43,* 401–409.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference* (3rd ed.). New York: Dekker.

Good, P. (1994). *Permutation tests for testing hypotheses.* New York: Springer-Verlag.

Gottman, J. M., & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers.* Cambridge: Cambridge University Press.

Gustafson, H., Dawson, E., Nielsen, L., & Caelli, W. (1994). A computer package for measuring the strength of encryption algorithms. *Computers and Security, 13,* 687–697.

Harrison, S. R., & Tamaschke, H. U. (1993). *Statistics for business, economics and management.* Sydney, Australia: Prentice-Hall.

Hazell, P. (1993). Adolescent suicide clusters: Evidence, mechanisms and prevention. *Australian and New Zealand Journal of Psychiatry, 27,* 653–665.

Hill, N. (1995). *Playing casino games to win.* Melbourne, Australia: Penguin.

Huba, G. J. (1986). The use of the runs test for assessing response validity in computer scored inventories. *Educational and Psychological Measurement, 46,* 929–932.

Huitema, B. E., McKean, J. W., & Zhao, J. (1996). The runs test for autocorrelated errors: Unacceptable properties. *Journal of Educational and Behavioral Statistics, 21,* 390–404.

IMSL. (1991). User's manual, IMSL Stat/Library (Version 2.0) [Computer program]. Houston, TX: Author.

Jobes, D. A., Berman, A. L., O'Carroll, P. W., Eastgard, S., & Knickmeyer, S. (1996). The Kurt Cobain suicide crisis: Perspectives from research, public health and the news media [with discussion]. *Suicide and Life Threatening Behavior, 26,* 260–271.

Katerndahl, D. A. (1996). Associations between departmental features and departmental scholarly activity. *Family Medicine, 29,* 119–127.

Kirch, M. R., & Lester, D. (1986). Clusters of suicide. *Psychological Reports, 59,* 1126.

Knuth, D. E. (1981). *The art of computer programming, Vol. 2: Semi-numerical algorithms* (2nd ed.). Reading, MA: Addison-Wesley.

Konold, C. (1995). Confessions of a coin flipper and would-be instructor. *The American Statistician, 49,* 203–209.

Koppen, M. G. M., & Verhelst, N. D. (1986). The exact runs test and some large sample approximations. *British Journal of Mathematical and Statistical Psychology, 39,* 168–182.

Lentner, C. (1982). (Ed.). *Geigy scientific tables: Vol. 2* (8th ed.). Basel, Switzerland: CIBA-GEIGY.

Levene, H., & Wolfowitz, J. (1944). The covariance matrix of runs up and down. *Annals of Mathematical Statistics, 15,* 58–69.

Lyon, M., Lyon, N., & Magnusson, M. S. (1994). The importance of temporal structure in analyzing schizophrenic behavior: Some theoretical and diagnostic implications. *Schizophrenia Research, 13,* 45–56.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology* (2nd ed.). London: Chapman & Hall.

Markov, A. A. (1955). *Example of a statistical investigation of the text of "Eugene Onegin" illustrating the dependence between samples in chain behavior* (M. Halle, Trans.). Unpublished manuscript, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge. (Original work published in 1913)

Markwitz, J. (1995). *Using speech recognition.* Englewood Cliffs, NJ: Prentice-Hall.

Martindale, C. (1991). *Cognitive psychology: A neural network approach.* Pacific Grove, CA: Brooks/Cole.

McKenzie, D. P., Mackinnon, A. J., Péladeau, N., Onghena, P., Bruce, P. C., Clarke, D. M., Harrigan, S., & McGorry, P. D. (1996). Comparing correlated kappas by resampling: Is one level of agreement significantly different from another? *Journal of Psychiatric Research, 30,* 483–492.

Mehta, C. R. (1994). The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research, 3,* 135–136.

Mendenhall, W., Reinmuth, J. E., & Beaver, R. (1993). *Statistics for management and economics* (7th ed.). Belmont, CA: Duxbury Press.

Menzerath, P. (1950). Typology of languages. *Journal of the Acoustical Society of America, 22,* 698–701.

Microsoft Corporation. (1995). FORTRAN PowerStation (Version 4.0) Professional edition [Computer program]. Redmond, WA: Author.

Mogull, R. G. (1994). The one-sample runs test: A category of exception. *Journal of Educational and Behavioral Statistics, 19,* 296–303.

Mood, A. M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics, 11,* 367–392.

Newman, E. B. (1951a). Computational methods useful in analyzing series of binary data. *American Journal of Psychology, 64,* 252–262.

Newman, E. B. (1951b). The pattern of vowels and consonants in various languages. *American Journal of Psychology, 64,* 369–375.

Onghena, P. (1993). A theoretical and empirical comparison of mainframe, microcomputer, and pocket calculator pseudorandom number generators. *Behavior Research Methods, Instruments and Computers, 25,* 384–395.

Onghena, P., & May, R. B. (1995). Pitfalls in computing and interpreting randomization test *p* values: A commentary on Chen & Dunlap. *Behavior Research Methods, Instruments and Computers, 27,* 408–411.

Paulus, M. P., Geyer, M. A., & Braff, D. L. (1996). Use of methods from chaos theory to quantify a fundamental dysfunction in the behavioral organization of schizophrenic patients. *American Journal of Psychiatry, 153,* 714–716.

Phillips, D. P. (1974). The influence of suggestion on suicide: Substantive and theoretical implications of the Werther effect. *American Sociological Review, 39,* 340–354.

Phillips, D. P. (1985). The Werther effect: Suicide and other forms of violence are contagious. *The Sciences, 7/8,* 32–39.

Pincus, S., & Kalman, R. E. (1997). Not all (possibly) "random" sequences are created equal. *Proceedings of the National Academy of Sciences USA, 94,* 3513–3518.

Pothoff, R. F. (1989). Run lengths, tests of. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Suppl., pp. 141–143). New York: Wiley.

Rahamimoff, R., Edry-Schiller, J., Rubin-Fraenkel, M., Butkovich, A., & Ginsbury, S. (1995). Oscillations in the activity of a potassium channel at the presynaptic nerve terminal. *Journal of Neurophysiology, 73,* 2448–2558.

Rump, A. F., Siekmann, U., Dreja, M., & Kalff, G. (1997). Caffeine pharmokinetics during hyperbaric hyperoxia in humans. *Aviation & Space Enviromental Medicine, 68,* 142–146.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

SPSS Inc. (1997). SPSS for Windows 95 (Version 7.5) [Computer program]. Chicago: Author.

Stannard, D. E. (1980). *Shrinking history: On Freud and the failure of psychohistory.* New York: Oxford University Press.

Stata Corporation. (1997). Stata statistical software: Release 5.0 [Computer program]. College Station, TX: Author.

Stevens, W. L. (1939). Distribution of groups in a sequence of alternatives. *Annals of Eugenics, 9,* 10–17.

Swed, F. S., & Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics, XIV,* 66–87.

Tries, M. A. (1997). Detection limits for samples that give rise to counting data with extra-Poisson variance. *Health Physics, 72,* 458–464.

Wallis, W. A., & Moore, G. H. (1941). A significance test for time series analysis. *Journal of the American Statistical Association, 36,* 401–409.

World Library, Inc. (1994). *Library of the future* (3rd ed.) [CD-ROM]. Irvine, CA: Author.