NASA TECHNICAL NOTE

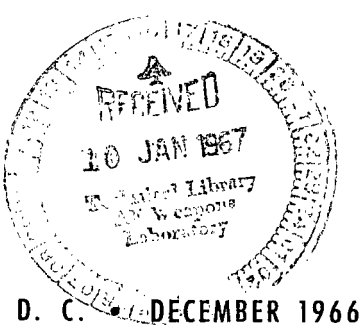
NASA TN D-3766

NASA TN D-3766

SOME NONPARAMETRIC TESTS FOR RANDOMNESS IN SEQUENCES

by Peter D. Argentiero and Robert H. Tolson

Langley Research Center
Langley Station, Hampton, Va.



NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • DECEMBER 1966

NASA TN D-3766

SOME NONPARAMETRIC TESTS FOR RANDOMNESS IN SEQUENCES

By Peter D. Argentiero and Robert H. Tolson

Langley Research Center
Langley Station, Hampton, Va.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – Price $1.00

SOME NONPARAMETRIC TESTS FOR RANDOMNESS IN SEQUENCES

By Peter D. Argentiero and Robert H. Tolson
Langley Research Center

SUMMARY

A rigorous definition of the concept of a random sequence is stated and statistical tests of the hypothesis that a given data sequence is random are discussed and tables are provided to facilitate the use of these tests. Certain tests are recommended for the detection of periodic nonrandomness and others are recommended for the detection of nonperiodic nonrandomness. A battery of tests designed specifically for the detection of periodic nonrandomness is constructed and examples of its use are given. The same is done for a battery of tests designed for the detection of nonperiodic nonrandomness. The question of the relationship between the confidence coefficient of a battery of tests and the confidence coefficients of the individual tests of the battery is considered.

INTRODUCTION

The determination of the best fit to a given set of data is a frequent objective of mathematical analysis. Examples of particular interest occur in orbit determination and in determination of gravitational field coefficients and other parameters from analysis of tracking data. After a "best fit" is obtained, it is important to examine the residuals or differences between the predicted data set and the observed data set. If these residuals do not form a random sequence, one might suspect that important parameters have been neglected or misrepresented in the analysis.

The object of this study is to discuss various methods of testing sequences for randomness and to provide examples of the application of these tests to given sequences. Intuitively speaking, a random sequence of numbers is a sequence in which the particular values of the elements are not a function of their position in the sequence. Even more loosely, a random sequence of numbers is a sequence in which nothing deterministic is taking place. However, if statistical tests are to be developed for deciding whether a given sequence is random, a considerably more rigorous definition of randomness is needed. To develop such a definition, assume that the ith element of a given sequence is a particular value of a random variable with which is associated in the usual fashion a probability density function $f_i(X)$. For the purposes of this paper the sequence is said to be random if for any value of i or j, $f_i(X) = f_j(X)$. From this definition it is clear

that the particular order in which the values of a random sequence present themselves was no more likely to have occurred than any other ordering. This notion that each permutation of a random sequence has the same probability of occurring is the basis of all the statistical tests for randomness which will be developed. It should also be mentioned that most of the tests mentioned are nonparametric in the sense that in the formulation of each test, no assumptions are made concerning the parameters of the probability density function used in the definition of a random sequence.

Several tests for randomness are discussed in this paper. Some are developed in detail and references are given for others. The use of several of these tests is then demonstrated by applications to given sequences and the results obtained are presented. The reader who is unfamiliar with statistical terminology is urged to consult one of the standard references in mathematical statistics such as reference 1.

SYMBOLS

A,B	sets
$A(\lambda),B(\lambda)$	terms defined by equations (13b) and (13c), respectively
a,b	elements of A and B, respectively
a_r,b_r	constants
C	term defined by equation (5)
$f(\ )$	function
g	term defined by equation (14)
I_a	set of all integers i, such that an element from A occupies ith position in a sequence
K_p	value of a normally distributed random variable with mean zero and variance unity and for which the probability of exceeding K_p is p
k,m,N,n,r	nonnegative integers
l	length of run

N_a number of elements in A

N_b number of elements in B

$\binom{n}{r} = \frac{n!}{(r)!\,(n - r)!}$

P probability

$\overline{P}$ confidence coefficient of a sequence of tests

R term defined by equation (9)

r_a number of runs of elements from set A

r_b number of runs of elements from set B

$S = \sum_{i \in I_a} i$

T an index

U number of runs in a sequence

U_r term defined by equation (10)

X_i ith element in a sequence

$Y_i = X_{i+k}$

λ, λ_r wavelengths

τ term defined by equation (8)

σ_r term defined by equation (11)

χ_T term defined by equation (12)

RUNS TEST

Several tests for randomness in sequences are based on the concept of runs and on the probability density function which will now be derived. Consider two sets of

elements A and B containing N_a and N_b elements, respectively. Let the elements of the sets form a sequence of $N_a + N_b$ points. Each maximal subsequence of elements of like kind is called a run. The two types of runs must alternate so that the number of runs is always one plus the number of unlike neighbors in the given sequence. For example, letting the symbol a represent a number from a set A and the symbol b represent a number from a set B, the sequence

a a b a b b a a b a b b

contains eight runs. If it is assumed that each permutation of the $N_a + N_b$ elements is equally likely to occur, the question is what is the probability of finding any specified number of runs in the sequence.

Since each permutation is equally likely to occur, the probability of obtaining exactly U runs in the sequence is given by the number of permutations which provide exactly U runs divided by the total number of permutations $N_a + N_b$ things in which N_a objects are indistinguishable from each other and the N_b objects are indistinguishable from each other. The denominator of the probability in question is

$$\frac{(N_a + N_b)!}{N_a! N_b!} = \frac{N!}{N_a! N_b!} \tag{1}$$

Deriving the numerator of the probability ratio is more difficult. Let N_a and N_b be, respectively, the number of elements from sets A and B. Consider the number of runs from A first and notice that the runs from B serve only to separate the elements from A into r_a separate compartments. There are $r_a - 1$ runs of elements from B. (If a run of elements from B appears either at the end or at the beginning of the sequence, they do not have an influence on the number of runs of elements from A and hence are not counted.) Therefore, the number of ways in which r_a runs can be obtained is

$$\binom{N_a - 1}{r_a - 1} = \frac{(N_a - 1)!}{(r_a - 1)!(N_a - r_a)!} \tag{2}$$

The same argument applies to the elements from B; hence, the number of permutations giving N_b runs from B is

$$\binom{N_b - 1}{r_b - 1} = \frac{(N_b - 1)!}{(r_b - 1)!(N_b - r_b)!} \tag{3}$$

Suppose that the sequence begins with an element from A. Then for each permutation of the N_a elements, the N_b elements can be permutated in all possible ways in their runs

to give distinct permutations of the N_a elements and the N_b elements jointly. Consequently, the total number of permutations of the N_a elements and the N_b elements together and starting with an element from A is the product of equations (2) and (3), namely,

$$\binom{N_a - 1}{r_a - 1}\binom{N_b - 1}{r_b - 1} \tag{4}$$

The same relationship applies if the sequence begins with an element from B. Since the runs of elements from A and B alternate, either $r_a = r_b$ or $r_a = r_b \pm 1$. If $r_a = r_b + 1$, the sequence begins with a run from A. If $r_a = r_b - 1$, the sequence begins with a run from B. In either case there is no choice for the first run and the number of permutations is given by expression (4). If $r_a = r_b$, there are two possible choices for the first run. Hence the total number of permutations is given by

$$\left.\begin{aligned} &C\binom{N_a - 1}{r_b - 1}\binom{N_b - 1}{r_b - 1} \\ &C = 1\left(r_a + r_b = U \quad \text{is odd}\right) \\ &C = 2\left(r_a + r_b = U \quad \text{is even}\right) \end{aligned}\right\} \tag{5}$$

If U is even, $r_a = r_b = \frac{U}{2}$. If U is odd, there are two cases to consider; either $r_a = \frac{U+1}{2}$ and $r_b = \frac{U-1}{2}$ or $r_a = \frac{U-1}{2}$ and $r_b = \frac{U+1}{2}$. Combining this fact with expressions (5) and (1) yields the probability density function for U, the number of runs:

$$f(U) = \frac{2\binom{N_a - 1}{\frac{U}{2} - 1}\binom{N_b - 1}{\frac{U}{2} - 1} N_a! N_b!}{N!} \tag{6a}$$

for U an even integer and

$$f(U) = \frac{\left[\binom{N_a - 1}{\frac{U-1}{2} - 1}\binom{N_b - 1}{\frac{U+1}{2} - 1} + \binom{N_a - 1}{\frac{U+1}{2} - 1}\binom{N_b - 1}{\frac{U-1}{2} - 1}\right] N_a! N_b!}{N!} \tag{6b}$$

for U an odd integer and $f(U) = 0$ otherwise.

Extensive tables for integrated or summed values of this function for various values of N_a and N_b along with several examples of their use are given in reference 2. A more detailed derivation of equation (6) together with an abbreviated table of its integrated values is provided in reference 1 (pp. 293-299).

If $N_a = N_b \geq 20$ and if the confidence level P satisfies the condition $\frac{1}{4N_a} \leq P \leq 1 - \frac{1}{4N_a}$, approximate formulas for the critical number of runs U_p can be obtained. One such formula is given in reference 3, namely,

$$U_p = \frac{3}{2} + N_a - K_p\left(\frac{N_a^2 - N_a}{2}\right)^{1/2} \tag{7}$$

where K_p is the value of a normally distributed random variable with mean zero and variance unity for which the probability of exceeding K_p is p. Critical values of the number of runs for certain standard confidence levels and for certain numbers of elements $N_a = N_b$ as calculated from equation (7) are given in table I. Since equation (7) does not always give integral values, a conservative approach was utilized to obtain table I, that is, for $P < 0.5$ the largest integer less than or equal to the value of U_p from equation (7) was taken and for $P > 0.5$ the smallest integer greater than or equal to the value of U_p from equation (7) was taken.

TABLE I.- CRITICAL VALUES OF U

$N_a = N_b$	U_p for a probability of –					
	0.005	0.01	0.025	0.975	0.99	0.995
20	13	14	15	28	29	30
30	21	22	23	40	41	42
40	30	31	32	51	52	53
50	38	39	41	62	64	65
60	47	48	50	73	75	76
70	56	57	59	84	86	87
80	65	66	68	94	97	98
90	74	75	78	105	108	109
100	83	85	87	116	118	120

THE MEDIAN TEST

An application of the concept of runs and of probability density function (6) is provided in what is known as the median test. (See ref. 2.) The procedure is as follows: Given a sequence to be tested for randomness, determine the median of the sequence. A run of length l is a sequence of values of length l such that all the l values are above the median or all the l values are below the median. If the null hypothesis is to be that the sequence is random, each permutation of its elements was as likely to have

occurred as another. Hence, the number of runs of the type just described is a random variable whose probability density function is given by equation (6). Thus, table I or the more elaborate tables provided in reference 2 can be used to test the null hypothesis that the sequence is random. For instance, suppose a sequence of length 100 is to be tested for randomness by using the median test. The numbers N_a and N_b are defined, respectively, as the number of elements above and below the median; thus, $N_a = N_b = 50$. If a confidence coefficient of 0.95 is chosen, consulting table I shows that with a probability of 0.95 the number of runs occurring should be between 41 and 62. If the number of runs is not between 41 and 62, with a confidence coefficient of 0.95, the hypothesis that the sequence is random can be rejected.

It should be clear that if a sequence tends to be periodic with a long period, several long runs would be introduced and since the sequence is finite, this condition would tend to produce fewer runs than would be expected in a random sequence. Conversely, if a sequence tends to have a very short period, too many runs would be expected. Thus the median test should be useful in detecting periodicities of long or short duration. This discussion leaves open the question of what, insofar as the median test is concerned, constitutes a periodicity of a long or a short period. In other words, is it true that random sequences exhibit a sort of periodicity of their own? In a sense the answer appears to be yes! If one divides the expected value of probability density function (eq. (6)) by the length of the sequence in question, the answer is invariably about 1/2. Thus a periodic sequence of period 4 points would exhibit about the same number of runs as a random sequence of the same length. In this crude sense, a random sequence can be said to have a natural period of around 4 points. Hence, the median test should not be expected to be sensitive to periodicities of this length.

A TWO-SAMPLE TEST

The concept of runs and the probability density function (eq. (6)) are useful in constructing nonparametric tests of a nonsequential nature. Consider the problem of determining when two different sets of numbers are sets of values of the same random variable. A test can be constructed in this manner. Suppose there are N_a elements in sample A and N_b elements in sample B. Arrange the $N_a + N_b$ elements in descending order of magnitude. If it is assumed as a null hypothesis that each set consists of values of the same random variable, the $N_a + N_b$ elements constitute a set of values from this random variable and hence all permutations of the elements should be equally likely to have occurred when ordered in descending magnitude. Thus, if a run of length l is defined as l elements in a row all from the same set, either A or B, then the probability density function for the number of runs of this type in the ordered set of $N_a + N_b$ elements is again given by equations (6). Thus, the tables in reference 2 can be used to test

the null hypothesis in the same way as with the median test. An example of this sort of test is given in reference 4 (page 263). It should be mentioned that it is also possible to derive the probability density function for the longest run of any particular sort in a sequence and base a test for randomness on this statistic. This possibility is discussed in reference 5.

WILCOXON'S TEST

Another method for testing the hypothesis that two sets of numbers are values of the same random variable is the Wilcoxon test. It differs from the two sample tests discussed previously in that it in no way relies on the concept of runs or on the probability density function (eq. (6)). Again, consider two sets of data A and B of N_a and N_b points, respectively. Arrange the $N_a + N_b$ numbers in descending order of magnitude. Let I_a be the set of all integers i, such that an element from A occupies the ith position in this sequence. If set A and set B consist of values from the same random variable, each set of N_a places occupied by the elements from A in the sequence is as likely to occur as any other set of N_a places. By making this assumption, a probability density function for the statistic $S = \sum_{i \in I_a} i$ may be derived. The derivation of this function along with tables of its integrated values are provided in reference 6. Examples of how to use these tables to test the hypothesis that two sets of numbers are values of the same random variable can be found in reference 4 (pp. 264-265).

SERIAL CORRELATION

If a sequence of numbers is random, no particular correlation would be expected to exist between a value in the sequence and the value, for example, k places in front of it. That is, if X_i is defined to be the value in the ith position in the sequence and Y_i, to be X_{i+k}, the correlation between X_i and Y_i should not be significant. Based on this notion, a nonparametric method of testing sequences for randomness can be devised if it is assumed that all permutations of the sequence being considered are equally probable. Define the serial correlation coefficient with lag k to be the linear correlation coefficient between X_i and Y_i. Since there are N! possible permutations, there are N! possible values of the serial correlation coefficient to be computed. The ordered set of values obtained together with the relative frequencies of those values which are obtained more than once provides the distribution of the serial correlation coefficient. If the sequence being tested yields a large positive or negative value of the serial correlation coefficient, its randomness would be in doubt. To obtain a critical region for testing randomness, it would be necessary to find two values such that some small percentage

of the N! values of the serial correlation coefficient lies outside the interval determined by the two values.

It is clear from the preceding discussion that the computational difficulties of the outlined test are prohibitive. Hence, it is necessary to find an approximation for the probability density function of the serial correlation statistic when N is large.

Let $Y_i = X_{i+k}$ for $i = 1, 2, \ldots N-P$ and $Y_{n-k+1} = X_i$ for $i = 1, 2, \ldots k$. There will be N pairs of values in the calculation of the correlation between X_i and Y_i. The resulting correlation coefficient is called the circular form of the serial correlation. The serial correlation coefficient may be expressed in the form

$$\tau = \frac{\sum_{i=1}^{n} X_i Y_i - N\overline{X}\,\overline{Y}}{NS_xS_y} \tag{8}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1} X_i$, $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ and $\overline{Y}$ and S_y^2 are defined similarly. The statistics $\overline{X}$, $\overline{Y}$, S_x, and S_y are independent of the order of the sample values. Thus, the only quantity in equation (8) affected by permutations of the sequence is the sum

$$R = \sum_{i=1}^{n} X_i Y_i \tag{9}$$

instead of τ itself.

If it is assumed that the values of the sequence being tested constitute a set of values of a random variable with only low order moments, it can be shown that the random variable R possesses an approximately normal probability density function for large N. The details may be found in reference 7. In order to test the hypothesis of zero correlation, it suffices to know the mean and variance of R. The necessary values are

$$U_r = \frac{S_1^2 - S_2}{N - 1} \tag{10}$$

and

$$\sigma_r = \frac{S_2^2 - S_4}{n - 1} + \frac{S_1^4 - 4S_1^2S_2 + 4S_1S_3 + S_2^2 - 2S_4}{(n - 1)(n - 2)} \tag{11}$$

where

$$S_j = \sum_{i=1}^{n} (X_i)^j$$

The serial correlation statistic with lag k is useful in constructing tests to detect periodicities of around k data points. Since it is suspected that the runs test is not sensitive to periodicities of four data points, the serial correlation test could be run in conjunction with the runs test with the lag set at 4. This procedure would effectively correct the previously mentioned deficiency in the runs test. Otherwise, the lag is simply set at the periodicity which one suspects might occur. An example of the application of the serial correlation test is given in reference 1 (pp. 302-303).

THE PERIODOGRAM

One of the most common methods of detecting hidden periodicities in a sequence is the method of periodogram analysis. Assume that the values of a sequence can be written in the following manner:

$$\chi_T = \sum_{r=1}^{K} \left(a_r \cos \lambda_r T + b_r \sin \lambda_r T\right) + \epsilon_T \tag{12}$$

where ϵ_T is a value of a random variable with an as yet unspecified probability density function. It is wished to detect the periods $2\pi/\lambda_r$ that have been hidden by the random disturbances ϵ_T. For this purpose the following statistic has proved to be useful.

$$I_n(\lambda) = \frac{1}{4\pi}\left[A^2(\lambda) + B^2(\lambda)\right] \tag{13a}$$

where

$$A(\lambda) = \sqrt{\frac{2}{n}} \sum_{T=1}^{n} \chi_T \cos \lambda T \tag{13b}$$

$$B(\lambda) = \sqrt{\frac{2}{n}} \sum_{T=1}^{n} \chi_T \sin \lambda T \tag{13c}$$

$I_n(\lambda)$ is defined as the periodogram of the sequence. The maxima of $I_n(\lambda)$ correspond to $\lambda = \lambda_r$, $r = 1, 2, \ldots k$. Hence the peaks of the periodogram should correspond to

hidden periodicities of the sequence. The difficulty with the technique is that there is no rigorous way of deciding when the height of a peak is significant without making assumptions about the random variable which produced the values of ϵ_T. Thus, when used rigorously, the periodogram method is not a nonparametric test for randomness.

If there are an odd number (n = 2N+1) of ϵ_T values which are values of a random variable with a normal probability density function, a rigorous test for the significance of the height of the maximum peak can be based on the statistic

$$g = \frac{(S_r)_{max}}{\sum_{r=1}^{n} S_r} \tag{14}$$

where for $i = 1, 2, \ldots n$

$$S_i = \frac{2}{n}\left[\left(\sum_{j=1}^{n} \chi_j \cos\frac{2\pi ij}{n}\right)^2 + \left(\sum_{j=1}^{n} \chi_j \sin\frac{2\pi ij}{n}\right)^2\right]$$

The probability function for the statistic g is derived in reference 8. The functional form is

$$\text{Probability } (g > \chi) = \frac{n(1-x)^{n-1}}{1!} - \frac{(n-1)(n-2)^{n-1}}{2!} + \frac{n(n-1)(n-2)(1-3)^{n-1}}{3!} + \ldots \tag{15}$$

where the sum extends as long as the term of the form (1 - kx) is positive. Equation (15) was solved for the critical values of g for various confidence coefficients and various values of N. The results are presented in table II. For example, if N = 50 (that is, 101 data points) the probability that the statistic g will exceed 0.13135 is 0.05. If $g > 0.13135$, it can be assumed to a confidence level of 0.95 that the height of the peak corresponding to $(S_r)_{max}$ is significant and that the value of the period corresponding to $(S_r)_{max}$ represents a true periodicity in the sequence.

APPLICATION OF TESTS FOR RANDOMNESS

General Considerations

To test a sequence for randomness, one takes as the null hypothesis the proposition that the points of the sequence constitute a set of values of a random variable with continuous probability density function. But when confronted with a sequence to be tested for randomness, the scientist is seldom in a position to choose a form for the probability

TABLE II.- CRITICAL VALUES OF g

$$\left[\text{Probability}\,(g \leq X) = \frac{N(1-X)^{N-1}}{1!} - \frac{N(N-1)(1-2X)^{N-2}}{2!} \cdots + \frac{N(N-1)(N-k+1)(1-kX)^{k-1}}{k!} \right.$$

where k is the largest integer less than $1/X$]

N	X for a probability of −		
	0.01	0.02	0.05
10	0.53584	0.49868	0.44495
20	.32921	.30480	.27040
30	.24124	.22287	.19784
40	.19156	.17705	.15738
50	.15954	.14754	.13135
60	.11309	.12686	.13708
70	.09953	.11150	.12041
80	.08903	.09962	.10751
90	.08064	.09014	.09723
100	.07378	.08239	.08882

density function of the random variable from which the points came. For this reason most of the tests discussed previously were nonparametric in nature. In a nonparametric test the null hypothesis is formulated without resort to an assumption concerning the random variable which produced the data points. It is this feature which makes nonparametric tests extremely useful in testing sequences for randomness. The unfortunate aspect of nonparametric statistics is that it is virtually impossible to order nonparametric tests with respect to power and to decide in a given situation which is the best test to use. Thus, an element of arbitrariness enters into the choice of tests for the discovery of nonrandomness in a given sequence. But as will be seen, this arbitrariness need not be absolute. In subsequent sections two different types of nonrandom sequences are discussed and by means of examples, indications are given as to the most advantageous sort of nonparametric test to be used on each.

Periodic and Nonperiodic Sequences

Let X_i be the ith element in a sequence and let $f_i(X)$ be the probability density function of the random variable of which X_i is a value. Then a sequence is said to be periodic if there exists an m such that $f_i(X) = f_{i+m}(X)$ for all i. If m is the smallest number which satisfies this condition, the sequence is said to be periodic of period m.

If $m = 1$, the sequence is random and is not considered periodic. If the physical situation giving rise to a certain sequence of data points is repetitive in nature, one would expect nonrandomness in the sequence to be periodic in nature. If the physical situation which produced the sequence of data points is nonrepetitive, it would be more likely that nonrandomness in the sequence be nonperiodic.

Confidence Coefficients

If P is the confidence coefficient of a statistical test of a hypothesis, then $1 - P$ is the probability of the test rejecting the hypothesis when in fact it is true. If one wishes to use a battery of statistical tests in order to decide on the acceptance or rejection of a hypothesis, certain questions must be decided: First, what kind of outcome of the battery of tests should correspond to a rejection of the hypothesis, and second, what is the confidence coefficient of the battery of tests when used in this fashion?

Suppose N tests of a certain hypothesis are to be used in sequence and suppose all tests are designed with a confidence coefficient of P. If it is decided that the hypothesis is to be rejected if k or more of the tests reject the hypothesis, and furthermore, if it is assumed that the tests are independent (this is usually a bad assumption but considerations of dependence are very difficult and can only serve to raise the confidence coefficient anyway), the confidence coefficient $\overline{P}$ of the battery reviewed as a single test is

$$\overline{P} = \sum_{r=K}^{n} \binom{n}{r} (1 - P)^r (P)^{n-r} \tag{16}$$

EXAMPLE OF A TEST FOR PERIODIC NONRANDOMNESS

A battery of seven tests has been chosen for the detection of periodic nonrandomness in sequences. The first is the median test on the total number of points in the sequence. The next six tests are of a serial correlation variety described previously. The lags are chosen to detect periodicities suspected to be present. This series of tests are applied to two sequences. The first sequence contains 200 points, all of which were chosen from a table of random numbers given in reference 4 (pp. 451-454). The numbers can be thought of as a set of values of a random variable whose probability density function is rectangular between 0 and 1000. These data are plotted in figure 1. The second sequence to be tested is obtained by imposing a sine wave of period 30 points and amplitude 300 on the data of figure 1. This new sequence is plotted in figure 2.

Each test is designed with a 0.99 confidence coefficient and the hypothesis of randomness is to be rejected if at least two of the seven tests reject the hypothesis. Under these conditions the confidence coefficient of the test is according to equation (16)

$$\overline{P} = 0.99$$

The lags chosen for the six serial correlation tests are 15, 30, 45, 60, 75, and 90. Since the sequence of figure 1 was constructed to be random, it would be hoped that the serial correlation statistic of equation (9) would be insignificant for all six values of the lag. But, since the sequence of figure 2 contains a sine wave of period 30 points, the serial correlation statistic would be expected to be significantly low for odd half-period lags of 15, 45, and 75 and significantly high for integer period lags of 30, 60, and 90.

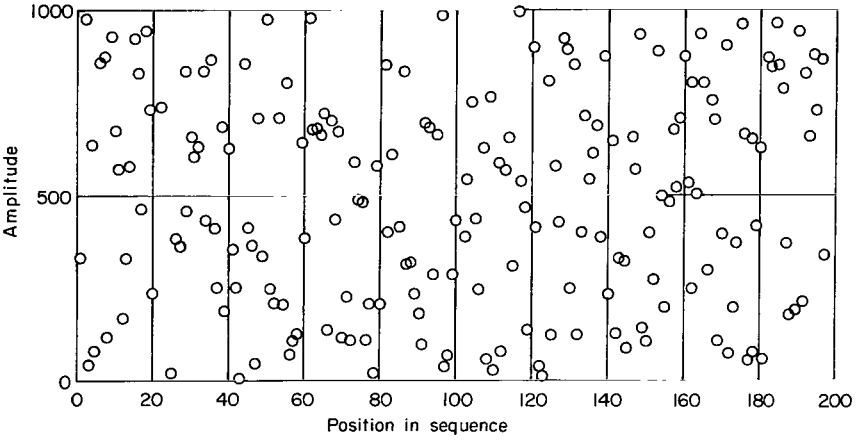


Figure 1.- Random data.

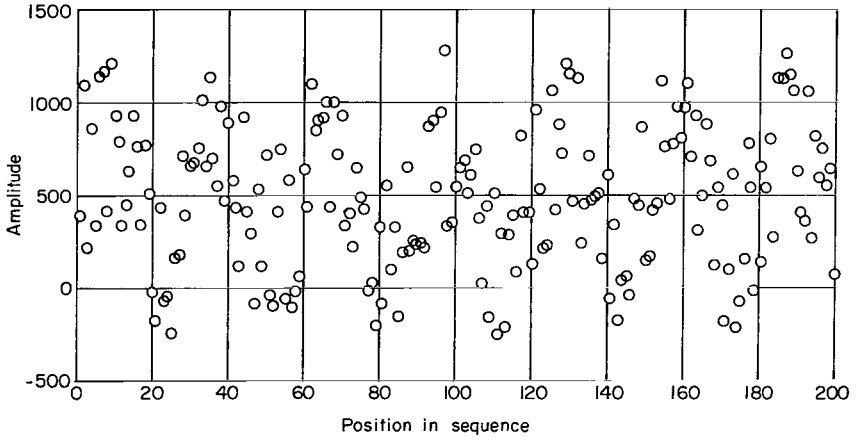


Figure 2.- Random data with sine wave of period 30 points and amplitude 300.

The results of these tests as performed on the random data of figure 1 are:

(a) the number of runs in the 200 data points is 111

(b) The serial correlation statistics are as follows:

Serial correlation statistic	Lag
0.4650×10^8	15
.4846	30
.4726	45
.4671	60
.4758	75
.4703	90

If table I is consulted, with a 0.99 confidence coefficient, a random sequence of 200 numbers should contain between 85 and 118 runs of the type used in the median test. Therefore, the sequence passes the median test.

If the approximation to the serial correlation statistic discussed in reference 4 (pp. 264-265) is used, the serial correlation statistic for this sequence with a 0.99 confidence coefficient should lie between 0.4435×10^8 and 0.5050×10^8. Clearly, the sequence passes the six serial correlation tests. The conclusion must be that the battery of tests has failed to detect any nonrandomness in the sequence.

Using the same tests on the data found in figure 2 yields the following results:

(a) The number of runs in the 200 data points is 89.

(b) The serial correlation statistics are as follows:

Serial correlation statistic	Lag
0.3898×10^8	15
.5842	30
.4287	45
.5362	60
.4625	75
.5079	90

The sequence passes the median test. Again, using the approximation discussed in reference 4 (pp. 264-265), the serial correlation statistic of this sequence and with a confidence coefficient of 0.99 should lie between 0.4427×10^8 and 0.5442×10^8. The sequence fails the first three serial correlation tests; therefore to a confidence coefficient of 0.99, the sequence is declared nonrandom. Noting further that the serial correlation statistics with lags 15 and 45 are significantly low and the serial correlation statistic with lag 30 is significantly high suggests a nonrandom periodicity of period 30 points. And this, of course, is the case.

EXAMPLE OF A TEST FOR NONPERIODIC NONRANDOMNESS

The series of tests to be used for the detection of nonperiodic nonrandomness is as follows. The points of the sequence are split into four parts. If the sequence has 200 points, part one consists of the first 50 points, part two of the second 50 points, etc. These sets are then compared pairwise by the two-sample test discussed previously. The two-sample test is a runs test but it involves a different type of run from the one used in the runs test of the first example. Nevertheless, the same probability density function is used in the construction of both types of runs test. Thus table I will again be useful. The two-sample test is an aid in deciding when a different type of statistical law is controlling the behavior of the second set of elements from that which is controlling the first set and so forth. Clearly, there will be six such tests. The seventh test will be a serial correlation test with lag 1. As was the case in the first example, each separate test is designed with a confidence coefficient of 0.99 and it is agreed to reject the hypothesis if at least two of the tests reject the hypothesis. The confidence coefficient $\overline{\overline{P}}$ of the test will again be 0.99.

Two sequences are tested. The first sequence to be tested is again the random data obtained from reference 8 and plotted in figure 1. The second sequence was obtained by adding a linear bias of 1.3 per point to the data of figure 1. This sequence is plotted in figure 3. Let A1 be the first fifty points, A2 the second, etc. Then the results of the tests applied to the first sequence are:

(a) The number of runs are as follows:

Runs between –	Number of runs
A1 and A2	52
A1 and A3	48
A1 and A4	49
A2 and A3	52
A2 and A4	49
A3 and A4	51

(b) The serial correlation statistic with lag 1 is 0.4616×10^8.

By consulting table I, it is found that with a confidence coefficient of 0.99, the number of runs between two sets of fifty points each from the same random sequence should lie between 39 and 64. The first six tests therefore do not reject the hypothesis of randomness. With a confidence coefficient of 0.99, the serial correlation statistic for this sequence should lie between 0.4435×10^8 and 0.5050×10^8. The sequence also passes the seventh test. Therefore, the battery of tests fails to detect any nonrandomness in the data displayed in figure 1.

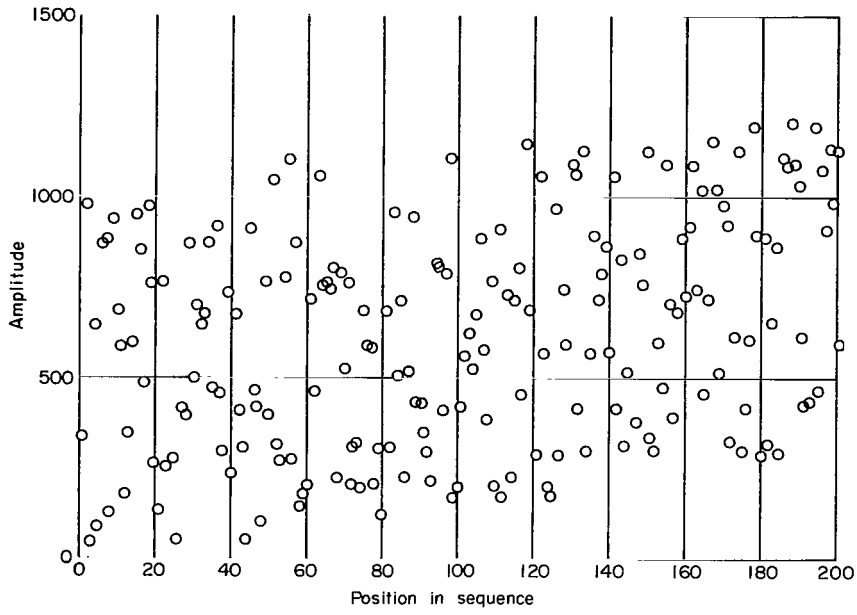


Figure 3.- Random data with linear bias of 1.3 per point.

Applying the same set of tests to the sequence plotted in figure 3 yields the following results:

(a) The number of runs are as follows:

Runs between –	Number of runs
A1 and A2	42
A1 and A3	44
A1 and A4	36
A2 and A3	52
A2 and A4	34
A3 and A4	48

(b) The serial correlation statistic with lag 1 is 0.7670×10^8.

The number of runs between A1 and A4 and between A2 and A4 are significantly low. The serial correlation statistic for this sequence and with a confidence coefficient of 0.99 should lie between 0.7291×10^8 and 0.7968×10^8. The sequence passes the serial correlation test. But since two of the seven tests reject the randomness hypothesis, with a confidence coefficient of 0.99 the sequence must be judged to be nonrandom.

CONCLUDING REMARKS

When a test is designed for nonrandomness in sequences, it is important to decide what type of nonrandomness the test is expected to detect. The most important dichotomy on nonrandom sequences seems to be the one which differentiates between periodic and nonperiodic nonrandomness. For this reason, examples were given of tests designed to detect each of these types of nonrandomness. Experience indicates that the test described in the first example is the most effective in detecting periodic nonrandomness and the test described in the second example is the most effective in detecting nonperiodic nonrandomness. It is also worth mentioning that the tests to be used in the detection of nonrandomness should be chosen before the data to be tested is seen. Any other procedure can ruin the statistical rigor of the test and seriously bias the outcome.

The median test has been found to be an efficient tool in detecting periodicities of period 10 points or more. It is a flexible test in the sense that the statistician need not have a hint of the period of the nonrandom disturbance present in order to apply the test effectively. The penalty for this flexibility is a lack of precision. If it is known that a certain sequence has failed the runs test, one might suspect the presence of a nonrandom periodic disturbance. The test, however, gives no hint as to the period of the disturbance. The case with the serial correlation test is precisely the opposite. It is not effective unless the statistician has a suspicion of the period of a nonrandom periodicity present. But the serial correlation test has the ability to confirm almost positively any such suspicion.

The periodogram test has both the flexibility of the runs test and the precision of the serial correlation test. But the hypothesis it tests is not simply that a given sequence is random but that a sequence is random and that its elements are values of a normal random variable. Thus it cannot be applied rigorously to the large number of situations that nonparametric tests can.

Of the two types of nonrandomness discussed, nonperiodic nonrandomness is the more difficult to detect and classify. It was hoped that the runs test used in the first example would also be effective in detecting nonperiodic disturbances. Limited experience has not shown this to be the case. Also it should be noticed that the two-sample-runs test is designed to test the same statistical hypothesis as the Wilcoxon test, namely, that two samples are sets of values of the same random variable. Therefore there is no rigorous reason for using one test in preference to the other. But again experience has indicated that the two-sample-runs test is somewhat more effective in detecting nonrandomness. The serial correlation test with lag 1 has been found to be effective in detecting gradual drifts in the mean of the sequence. Drifts in the variance are more difficult to

detect. In fact, it appears as if unless the drift in variance is drastic, there is no nonparametric test which is completely adequate in discovering a changing variance.

Langley Research Center,
National Aeronautics and Space Administration,
Langley Station, Hampton, Va., August 19, 1966,
129-04-01-01-23.

REFERENCES

1. Hoel, Paul G.: Introduction to Mathematical Statistics. Second ed., John Wiley & Sons, Inc., c.1954.

2. Swed, Frieda S.; and Eisenhart, C.: Tables for Testing Randomness of Grouping in a Sequence of Alternatives. Ann. Math. Statist., vol. 14, 1943, pp. 66-87.

3. Walsh, John E.: Handbook of Nonparametric Statistics. D. Van Nostrand Co., Inc., c.1962.

4. Johnson, Norman L.; and Leone, Fred C.: Statistics and Experimental Design in Engineering and the Physical Sciences. Vol. I. John Wiley & Sons, Inc., c.1964.

5. Mosteller, Frederick: Note on an Application of Runs to Quality Control Charts. Ann. Math. Statist., vol. 12, 1941, pp. 228-232.

6. Siegel, Sidney; and Tukey, John W.: A Nonparametric Sum of Ranks Procedure for Relative Spread in Unpaired Samples. J. Amer. Statist. Assoc., vol. 55, 1960, pp. 429-445.

7. Wald, A.; and Wolfowitz, J.: An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation. Ann. Math. Statist., vol. 14, 1943, pp. 378-388.

8. Fisher, R. A.: Tests of Significance in Harmonic Analysis. Proc. Roy. Soc. (London), ser. A, vol. 125, 1929, pp. 54-59.

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Technical information generated in connection with a NASA contract or grant and released under NASA auspices.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

TECHNICAL REPRINTS: Information derived from NASA activities and initially published in the form of journal articles.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities but not necessarily reporting the results of individual NASA-programmed scientific efforts. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

Washington, D.C. 20546