A SIGNIFICANCE TEST FOR TIME SERIES ANALYSIS*

By W. Allen Wallis and Geoffrey H. Moore
Stanford University and Rutgers University
National Bureau of Economic Research

No known significance test is entirely appropriate to economic time series. One shortcoming of tests in common use is that they ignore sequential or temporal characteristics; that is, they take no account of order. The standard error of estimate, for example, implicitly throws all residuals into a single frequency distribution from which to estimate a variance. Furthermore, the usual tests cannot be applied when series are analyzed by moving averages, free-hand curves, or similar devices frequently resorted to in economics for want of more adequate tools. This paper presents a test of an opposite kind, one depending solely on order. Its principal advantages are speed and simplicity, absence of assumptions about the form of population, and freedom from dependence upon "mathematically efficient" methods, such as least squares. This test is based on sequences in direction of movement, that is, upon sequences of like sign in the differences between successive observations (or some derived quantities, e.g., residuals from a fitted curve). In essence, it tests the randomness of the distribution of these sequences by length.

Each point at which the series under analysis (either the original or a derived series) ceases to decline and starts to rise, or ceases to rise and starts to decline, is called a turning point. A turning point is a "peak" if it is a (relative) maximum or a "trough" if a (relative) minimum. The interval between consecutive turning points is called a "phase." A phase is an "expansion" or a "contraction" according to whether it starts from a trough and ends at a peak, or starts from a peak and ends at a trough. For the purposes of the present test, the incomplete phase preceding the first turning point and that following the last turning point are ignored. The length or duration of a phase is the number of intervals (hereafter referred to as "years," though they may represent any system of denoting sequence) between its initial and terminal turning points. Thus, a series of N observations may contain as few as zero or as many as $N-2$ turning points; and a phase may be as short as one year (when two consecutive observations are turning

* Presented (in slightly different wording) before the Nineteenth Annual Conference of the Pacific Coast Economic Association, Stanford University, December 28, 1940, and based on research carried out at the National Bureau of Economic Research, 1939–40, under Research Associateships provided by the Carnegie Corporation of New York. A fuller account of the method and its uses will be published soon by the National Bureau of Economic Research as the first of its new series of *Technical Papers*.

points) or as long as $N-3$ years (when only the second and penultimate observations are turning points).

The greater the number of consecutive rises in a series drawn at random from a stable population, the less is the probability of an additional rise; for the higher any observation may be the smaller is the chance of drawing one which exceeds it. To calculate the expected frequency distribution of phase durations, only one weak assumption need be made about the population from which the observations come, namely that the probability of two consecutive observations being identical is infinitesimal—a condition met by all continuous populations, hence by virtually all metric data.

Without further postulates about the form of the population, it is possible to conceive a mathematical transformation of it leading to a known population, but leaving unaltered the pattern of rises and falls of the original observations. For example, if each observation is replaced by its rank according to magnitude within the entire series, the ranks have exactly the same pattern of expansions and contractions as the original observations; and their distribution is simple and definite, each integer from 1 to N having a relative frequency of $1/N$. The distribution of phase durations expected among random arrangements of the digits 1 to N is, therefore, comparable with the distribution observed in any set of data. A little mathematical manipulation reveals that in random arrangements of N different items the expected number of completed phases of d is $\frac{2(d^2+3d+1)(N-d-2)}{(d+3)!}$. The expected mean duration of a phase is $\frac{3N-11.6194}{2N-7}$, essentially $1\frac{1}{2}$.

To test the randomness of a series with respect to phase durations, the first step is to list in order the signs of the differences between successive items. Thus the sequence 0, 2, 1, 5, 7, 9, 8, 7, 9, 8 becomes $+, -, +, +, +, -, -, +, -$. The signs are, of course, one fewer than the observations. The second step is to make a frequency distribution of the lengths of runs in the signs. There are four completed runs in the example just given (the first and last being ignored as incomplete), of lengths, 1, 3, 2, and 1. The frequency distribution thus shows two phases of one year, one of two years, and one of three years. In case consecutive items are equal but it can be assumed that sufficiently refined measurement would reveal at least a slight difference (an assumption valid whenever the test is applicable), the phase lengths are tabulated separately for each possible sequence of signs of differences between tied items; and the resultant distributions are averaged, each being weighted by the probability of securing that distribution if each difference observed as zero is equally likely to be positive or negative. Third, the expected frequency for each length of phase is calculated from the formula above, taking as N the number of items in the sequence being tested—in this case, 10. Next, the observed and expected frequency distributions are compared by computing chi-square in the usual way for testing goodness-of-fit: that is, by squaring the differences between actual frequencies and corresponding theoretical frequencies, dividing these squares by the respective theoretical frequencies, and summing the resultant ratios. In nearly all applications of the present test, avoidance of expected frequencies that are too small necessitates restricting the distribution of phase durations to three frequency classes, namely one year's duration, two years' duration, and over two years' duration, the theoretical frequencies for these classes being $5(N-3)/12$, $11(N-4)/60$, and $(4N-21)/60$, respectively.

The sum of the three ratios of squared deviations to expectations is, then, similar to chi-square for two degrees of freedom, one degree of freedom being lost because a single linear constraint is imposed on the theoretical frequencies by taking the value of N from the observations. It is advisable, however, to distinguish chi-square for phase durations by a subscript p (denoting phase), because it does not quite conform to the Pearsonian distribution function ordinarily associated with the symbol χ^2. The phase lengths in a single series are not entirely independent of one another; as a result, very large and very small values of χ_p^2 are a little more likely than is shown by the χ^2 distribution, and the mean and variance of χ_p^2 generally exceed those of χ^2. We have not determined the sampling distribution of χ_p^2 mathematically, but have secured empirically a substitute that appears satisfactory. In the first place, a recursion formula enabled us to calculate the exact distribution of χ_p^2 for small values of N. Table I gives the exact probability of obtaining a value as large as or larger than each possible value of χ_p^2 for values of N from 6 to 12, inclusive. As a second step toward determining the sampling distribution, an empirical distribution of χ_p^2 was secured from 700 random series, 200 for $N=25$, 300 for $N=50$, and 200 for $N=75$. The three distributions for separate values of N were not homogeneous with one another nor with the exact distribution for $N=12$; but the differences among them were unimportant for the present purposes, occurring chiefly at the higher probabilities rather than at the tail (the important region for a test of significance), and representing

TABLE I

DISTRIBUTIONS OF χ_p^2: EXACT FOR $N=6$ TO 12, AND APPROXIMATE FOR LARGER VALUES OF N

(P represents the probability that an observed χ_p^2 will equal or exceed the specified value)

$N=6$

χ_p^2	P
.467	1.000
.867	.869
1.194	.675
1.667	.453
2.394	.367
2.867	.222
19.667	.053

$N=7$

χ_p^2	P
.552	1.000
.733	.789
.752	.703
.933	.536
1.733	.493
2.152	.370
2.333	.302
3.933	.277
5.606	.169
7.504	.117
8.904	.055

$N=8$

χ_p^2	P
.284	1.000
.684	.843
.844	.665
.920	.590
1.320	.560
1.480	.506
2.364	.495
2.680	.471
2.935	.392
3.000	.299
4.375	.293
4.455	.235
4.935	.194
5.000	.133
5.819	.064
6.455	.033

$N=9$

χ_p^2	P
.358	1.000
1.158	.798
1.267	.631
1.630	.605
2.067	.489
2.430	.452
2.758	.381
3.158	.374
3.267	.321
3.667	.215
4.030	.164
4.067	.144
4.758	.110
5.667	.078
6.067	.064
7.485	.020
15.666	.005

$N=10$

χ_p^2	P
.328	1.000
.614	.941
.728	.917
1.055	.813
1.341	.693
1.419	.606
1.585	.601
1.705	.594
1.772	.592
1.814	.526
1.819	.419
2.313	.407
2.577	.374
2.676	.327
2.743	.327
2.863	.274
2.905	.242
2.977	.220
3.242	.181
3.834	.179
3.970	.165
4.333	.158
4.400	.158
4.676	.139
4.858	.107
5.128	.072
5.491	.059
6.515	.054
7.133	.042
11.308	.014
12.965	.006

$N=11$

χ_p^2	P
.479	1.000
.579	.980
.817	.934
.917	.844
.979	.730
1.088	.723
1.279	.655
1.317	.576
1.588	.537
1.700	.473
1.800	.472
2.079	.468
2.200	.467
2.309	.466
2.409	.440
2.417	.403
2.500	.392
2.579	.384
2.688	.304
2.809	.274
3.026	.261
3.109	.230
3.213	.201
3.300	.147
3.779	.147
3.800	.147
3.909	.133
4.117	.128
4.313	.126
4.388	.099
4.726	.091
5.000	.077
5.609	.077
5.700	.076
6.013	.055
8.200	.050
8.635	.032
9.468	.022
9.735	.018
10.214	.009
11.435	.004

$N=12$

χ_p^2	P
0.615	1.000
0.661	.984
0.748	.896
0.794	.891
0.837	.850
0.971	.786
1.015	.720
1.061	.685
1.415	.585
1.461	.583
1.637	.569
1.683	.533
1.933	.487
1.948	.486
2.067	.428
2.156	.427
2.203	.407
2.289	.344
2.333	.333
2.556	.331
2.615	.303
2.661	.303
2.733	.300
2.837	.300
2.870	.287
2.883	.246
2.956	.216
3.267	.211
3.415	.207
3.489	.149
3.933	.127
4.070	.127
4.156	.114
4.348	.113
4.394	.113
4.571	.112
4.616	.109
4.733	.101
5.667	.092
5.803	.092
5.889	.090
6.025	.090
6.733	.085
6.842	.072
6.956	.060
7.504	.050
7.622	.041
8.576	.029
8.822	.026
9.237	.019
9.267	.014
10.556	.003
19.667	.000

$N>12$

χ_p^2	P
5.448	*.10*
5.50	.098
5.674	*.09*
5.75	.087
5.927	*.08*
6.00	.077
6.163	*.07*
6.25	.069
6.50	.061
6.541	*.06*
6.75	.054
6.898	*.05*
7.00	.048
7.25	.043
7.401	*.04*
7.50	.038
7.75	.034
8.00	.030
8.009	*.03*
8.25	.027
8.50	.024
8.75	.021
8.836	*.02*
9.00	.019
9.25	.017
9.50	.015
9.75	.013
10.00	.012
10.25	.010
10.312	*.01*
10.50	.009
10.75	.008
11.00	.007
11.25	.006
11.50	.006
11.755	*.005*
12.00	.004
13.00	.003
14.00	.002
15.085	*.001*

local irregularities rather than basic differences in form. It appears, therefore, that when N is as large as 12 a single sampling distribution of χ_p^2 is sufficient.

The mean of the 700 values of χ_p^2 is 2.3049, and the variance 5.0458. As an approach to the distribution of χ_p^2, it seems reasonable simply to reduce χ_p^2 by approximately one-seventh and refer it to the χ^2 distribution for two degrees of freedom, which has a mean of two and tables for which are readily available; and such a comparison does indicate good conformity. That the variance of the observed values is less than that of $(7/6)\chi^2$ for two degrees of freedom suggests, however, that a more satisfactory fit at the tails can be secured by using a distribution having a variance of 5, e.g., χ^2 for "two and one-half degrees of freedom." For χ_p^2 above about 5.5 and P below about .10, the agreement of this distribution with the observations is very satisfactory. In the main body of the distribution the function whose mean value is equated to the sample mean gives a somewhat better fit.

In practice, therefore, the procedure for interpreting χ_p^2, assumed always to be calculated from three frequency classes, is as follows: If χ_p^2 is less than 6.3 (the point of intersection between the ogives of $(7/6)\chi^2$ for two degrees of freedom and χ^2 for two and one-half degrees of freedom), reduce it by one-seventh and refer to the usual χ^2 tables for two degrees of freedom. This procedure is satisfactory for all values of χ_p^2, but for values above 6.3 somewhat more accurate results are apparently secured by referring χ_p^2 to the last column of Table I, which gives the distribution of χ^2 for two and one-half degrees of freedom. When $N<13$ the exact distributions should, of course, be used.

A simpler test of the same nature may be based on the fact that in a random sequence of N observations (where N is not too small—not less than 10, say) the total number of completed phases is normally distributed about a mean of $(2N-7)/3$ with variance of $(16N-29)/90$. (In using this test, the difference between the observed and expected numbers of phases should be reduced in absolute value by one-half unit, in order to allow for discontinuity.) This test of the total number of phases, which is essentially equivalent to a test of the mean phase duration, is normally less sensitive than the χ_p^2 test, which takes account of the lengths of the phases, though the superiority of the χ_p^2 test in this respect is limited by the necessity of confining the frequency distribution to three classes. Advantages of the test of the total number of phases are that it is even simpler to apply than the χ_p^2 test, that its sampling distribution is known exactly and is readily available, and that it is adaptable to cases where the hypothesis alternative to the null hypothesis is either that the phases are abnormally long or that they are abnormally short.

Application of the χ_p^2 test to an economic problem may be illustrated by an analysis of sweet potato production, yield per acre, and acreage harvested in the United States, 1868–1937, as recorded on page 243 of *Agricultural Statistics, 1939.* The frequency distributions of phase durations in these series have been compared with those to be expected in a random sequence. From the values of χ_p^2 and their corresponding probabilities, it appears that the fluctuations in production conform with what would be expected in a random series; while of the two components of total production, yield per acre conforms well and acreage harvested does not conform at all.

The figures on total production do not, of course, constitute a random series, for there is a marked upward trend in the data. In general, the method here presented is not very sensitive to primary trend. The removal of trend from a series, or its introduction into a trendless series, can change the sign of the difference between consecutive items only if the trend factor for a single year is greater than the difference between the items in trend-adjusted form. If, therefore, the residuals from trend are such that their first differences are rarely as small as the trend factor for a single year, as is frequently the case in economic time series, the distribution of phase lengths will not be much affected by the presence or absence of trend. Another factor tending to minimize the effect of trend on the test is that expansions are lengthened and contractions shortened if the trend is upward, and vice versa if it is downward, leaving the total number of phases of a given duration relatively unaffected. In such cases the existence of trend may be revealed by separate distributions for expansions and contractions. In the case of sweet potato production, both distributions conform well to the chance distribution and to one another; but for acreage the two distributions differ markedly, suggesting that the non-randomness evidenced in the acreage series may be at least partly attributable to trend.

Lack of sensitivity to primary trend is a limitation of the technique from the point of view of detecting the existence of such a trend. On the other hand, it is not difficult to determine by other methods whether a primary trend exists—the rank correlation between the variate and the date often affords a convenient test. And for determining whether the systematic variation contains secondary components, e.g., cyclical or seasonal variations, it is a decided advantage of the present method that it frequently gives satisfactory results regardless of the presence of trend, thus avoiding the complexities of trend elimination. It is possible, of course, for secondary fluctuations also to be concealed if their year to year magnitude is definitely less than year to year random movements; this is not so likely as in the case of a primary trend, but is a real possibility in the case of gradual movements—e.g., long waves.

A second example illustrates the use of χ_p^2 as a criterion of the fit of moving averages, and for selecting the proper period for a moving average. If a moving average, or any other curve, describes adequately the systematic variation in a series, the residuals should constitute a random sequence. If the period is too long, waves or cycles may appear in the residuals, and if it is too short the residuals will cluster too closely about the line. To illustrate this application, ten moving averages having spans from 2 to 11 years have been fitted to the data on sweet potato acreage, and the residuals tested for randomness. Each moving average uses equal weights; the necessity of centering each average at an observation, however, means an implicit increase of one year in the span of averages based on an even number of points, with the first and last observations receiving half weight. The last two columns of Table II show the values of χ_p^2, and the corresponding probabilities, obtained by testing the residuals for randomness. The moving averages based on even numbers of years give notably better results than those based on the corresponding odd numbers of items; the tapering of the weight diagram implicit in the even averages evidently improves the fit. Another striking feature is that the probabilities first rise and then decline. Thus, the odd averages attain a maximum probability of .24 at seven years while the even averages give the best result at six years, when the probability is .61. Had other weight diagrams been tested still better results might have been obtained.

It should be noted that a "better" result is not necessarily one in which the curve gives a closer fit, but one in which the residuals behave more like a series of independent, random observations, as judged by sequences in signs of first differences. The closest fits to the original observations are given by the shortest moving averages; but these describe not only the systematic variation but also a portion of the random fluctuations. If the moving average is either too short or too long, χ_p^2 will be significantly large; but in the former case its magnitude results from an excess of short phases and a deficiency of long ones, while in the latter case the reverse is true. Table II includes the actual frequency distributions of residuals from the ten curves.

In order to compare this new test with a more elaborate procedure frequently employed in time series analysis, a power series $y=a+bx+cx^2+dx^3+\cdots$ was fitted by the method of least squares to the series

TABLE II

FREQUENCY DISTRIBUTIONS OF PHASE DURATIONS IN RESIDUALS FROM MOVING AVERAGES FITTED TO SWEET POTATO ACREAGE HARVESTED UNITED STATES, 1868–1937

Span of moving average (years)		Duration of phase			Total (frequency)	χ_p^2	P
		One year (frequency)	Two years (frequency)	Over two years (frequency)			
2	Expected Observed	27.083 46	11.733 8	4.183 1	43 55	16.823	.0004
3	Expected Observed	27.083 46	11.733 8	4.183 1	43 55	16.823	.0004
4	Expected Observed	26.25 31.25	11.367 8.25	4.05 4.5	41.667 44	1.857	.45
5	Expected Observed	26.25 19.25	11.367 9	4.05 7.75	41.667 36	5.740	.09
6	Expected Observed	25.417 25.5	11 8.25	3.917 5.25	40.333 39	1.141	.61
7	Expected Observed	25.417 28.5	11 6	3.917 5.5	40.333 40	3.287	.24
8	Expected Observed	24.583 25	10.633 7	3.783 6	39 38	2.547	.34
9	Expected Observed	24.583 18.75	10.633 7.5	3.783 6.75	39 33	4.634	.14
10	Expected Observed	23.75 23	10.267 5.5	3.65 5.5	37.667 34	3.175	.26
11	Expected Observed	23.75 22	10.267 2	3.65 7	37.667 31	9.861	.01

on sweet potato acreage harvested. The calculations were carried as far as the ninth degree term, using the technique of orthogonal polynomials, but none beyond the third effected a significant reduction in the residual variance. According to the usual criterion, therefore, a third degree curve would be regarded as fitting adequately. The residuals from the third degree curve were then submitted to the present test. There were 24 one-year phases, 3 two-year phases, and 9 phases of more than two years, producing a χ_p^2 of 12.47, from which it is clear that the fit of the third degree polynomial is quite inadequate. The fault, of course, lies in inferring that a third degree power series gives an adequate fit because no other power series gives a significantly better fit as judged by the standard error of estimate.

χ_p^2 can also be used to test the independence of two variates, and in some circumstances is superior for this purpose to the rank correlation coefficient. The procedure is to arrange the pairs according to the order of magnitude of one variate and tabulate the distribution of phase durations in the other variate. If the two series are independent, the resulting value of χ_p^2 will not be significant. A difficulty, however, is that the conclusion occasionally depends upon which variate is chosen for arranging in order and which for counting the phase durations.

It is perhaps advisable to emphasize explicitly that the present test by no means utilizes all of the information in the data. In particular, it ignores the magnitude of the year to year fluctuations, treating the smallest as equivalent to the largest. A serial correlation coefficient computed from ranks may retrieve some of this information on magnitude. Another consideration in interpreting the test is that a set of phase durations which appears random when viewed only as a frequency distribution may not have been arranged at random in time. An additional point, obvious but worthy of mention, is that the time unit used may affect conclusions derived from the χ_p^2 test; for example, year to year movements may appear random and month to month movements non-random, or vice versa.