Why do we Sometimes get Nonsense-Correlations between Time-Series?--A Study in Sampling and the Nature of Time-Series

Journal of the Royal Statistical Society, Volume 89, Issue 1 (Jan., 1926), 1-63.

Stable URL:
http://links.jstor.org/sici?sici=0952-8385%28192601%2989%3A1%3C1%3AWDWSGN%3E2.0.CO%3B2-L

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Journal of the Royal Statistical Society is published by Royal Statistical Society. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Journal of the Royal Statistical Society
©1926 Royal Statistical Society

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2002 JSTOR

http://www.jstor.org/
Sun Nov 3 11:53:08 2002

Vol. LXXXIX.] [Part I.

JOURNAL

OF THE ROYAL STATISTICAL SOCIETY.

JANUARY, 1926.

WHY DO WE SOMETIMES GET NONSENSE-CORRELATIONS BETWEEN TIME-SERIES ?—A STUDY IN SAMPLING AND THE NATURE OF TIME-SERIES.

THE PRESIDENTIAL ADDRESS OF MR. G. UDNY YULE, C.B.E., M.A., F.R.S., FOR THE SESSION 1925–26. DELIVERED TO THE ROYAL STATISTICAL SOCIETY, NOVEMBER 17, 1925.

		PAGE
Section I.—The problem		2
,, II.—The correlation between simultaneous segments of two variables that are simple harmonic functions of the time, of the same period but differing by a quarter-period in phase; and the frequency-distribution of correlations for random samples of such segments....		6
,, III.—Deductions from Section II: classification of empirical series		13
,, IV.—Experimental investigations		30
,, V.—Serial correlations for Sir William Beveridge's index-numbers of wheat prices in Western Europe; and for rainfall at Greenwich		41
Appendix I.—The correlations between segments of two sine-curves of the same period, etc.		54
,, II.—The relations between the serial correlations of a sum-series and of its difference series, when the series may be regarded as indefinitely long....		57

THE problem which I have chosen as the subject of my Address is one that puzzled me for many years. The lines of solution only occurred to me two or three years ago, and I thought that I could not do better than endeavour to work them out during the Session 1924–25—time and opportunity having hitherto been lacking—and utilize them for the present purpose. As often happens, the country

to be explored opened up so widely as one advanced, that two or three years would have been a happier allowance of time for preparation than one year: much has had to be left aside for further exploration. But the results obtained up to the present stage seem to be of a good deal of interest and of some value.

First, let me expound with a little more detail and illustration the brief statement of the problem in my title.

SECTION I.—*The problem.*

It is fairly familiar knowledge that we sometimes obtain between quantities varying with the time (time-variables) quite high correlations to which we cannot attach any physical significance whatever, although under the ordinary test the correlation would be held to be certainly "significant." As the occurrence of such "nonsense-correlations" makes one mistrust the serious arguments that are sometimes put forward on the basis of correlations between time-series—my readers can supply their own examples—it is important to clear up the problem how they arise and in what special cases. Fig. 1 gives a very good illustration. The full line shows the proportion of Church of England marriages to all marriages for the years 1866–1911 inclusive: the small circles give the standardized mortality per 1,000 persons for the same years. Evidently there is a very high correlation between the two figures for the same year: the correlation coefficient actually works out at $+ 0{\cdot}9512$.

Now I suppose it is possible, given a little ingenuity and good-will, to rationalize very nearly anything. And I can imagine some enthusiast arguing that the fall in the proportion of Church of England marriages is simply due to the Spread of Scientific Thinking since 1866, and the fall in mortality is also clearly to be ascribed to the Progress of Science; hence both variables are largely or mainly influenced by a common factor and consequently ought to be highly correlated. But most people would, I think, agree with me that the correlation is simply sheer nonsense; that it has no meaning whatever; that it is absurd to suppose that the two variables in question are in any sort of way, however indirect, causally related to one another.

And yet, if we apply the ordinary test of significance in the ordinary way, the result suggests that the correlation is certainly "significant"—that it lies far outside the probable limits of fluctuations of sampling. The standard error of a coefficient of correlation is $(1 - r^2)/\sqrt{n}$, where n is the number of observations: that is to say, if we have the values of the two variables x and y entered in their associated pairs on cards, if we take out at random a sample of n cards

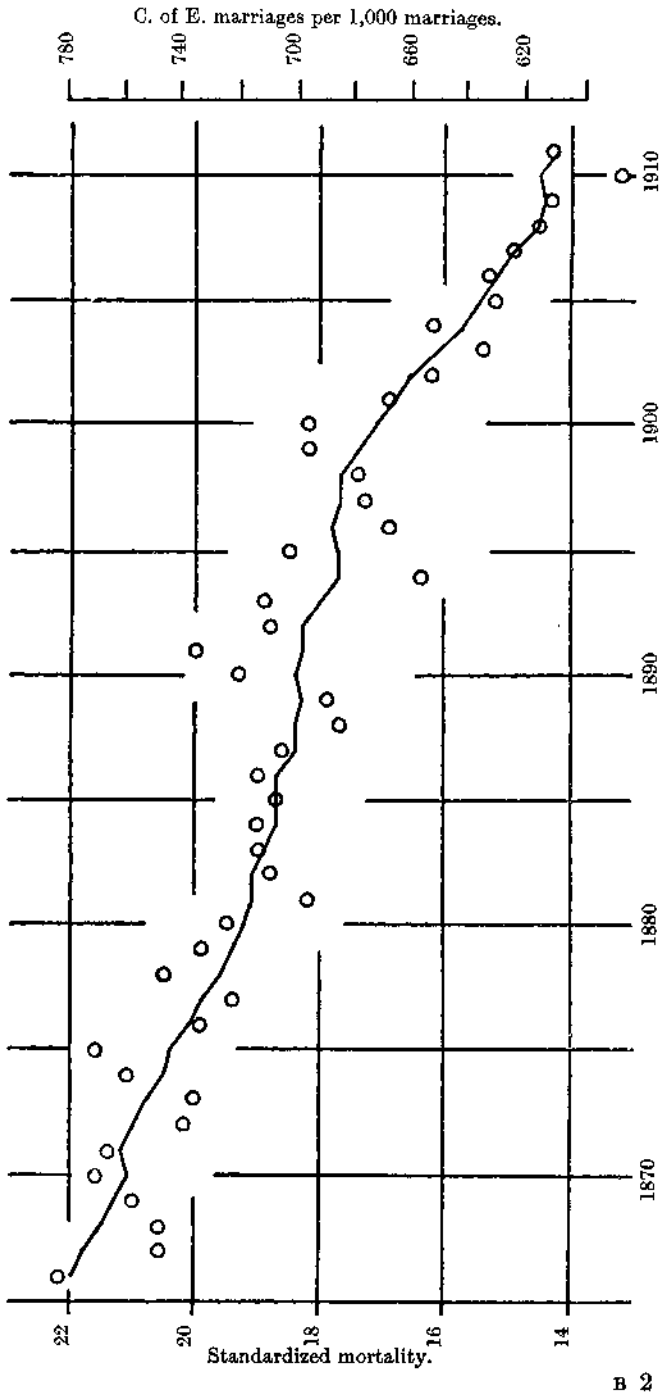

FIG. 1.—Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0{\cdot}9512$.

(small compared with the total of cards available) and work out the correlation for this sample, take another sample in the same way, and so on—then the correlation coefficients for the samples will fluctuate round the correlation r for the aggregate of cards with a standard deviation $(1 - r^2)/\sqrt{n}$. For the assigned value of r, viz., 0·9512... and 46 observations, the standard error so calculated is only 0·0140, and on this basis we would judge that we could probably trust the coefficient within 2 or 3 units in the second place of decimals. But we might ask ourselves a different question, and one more germane perhaps to the present enquiry. If we took samples of 46 observations at random from a record in which the correlation for the entire aggregate was zero, would there be any appreciable chance of our getting such a correlation as 0·9512 merely by the chances of sampling? In this case the standard error would be $1/\sqrt{46}$, or 0·1474. the observed correlation is 6·45 times this, and the odds would be many millions to one against such a value occurring " by chance "—odds so great that the event may be written down as for all practical purposes impossible. On the ordinary test applied in the ordinary way we seem compelled to regard the correlation as having *some* meaning.

Now it has been said that to interpret such correlations as implying causation is to ignore the common influence of the time-factor. While there is a sense—a special and definite sense—in which this may perhaps be said to cover the explanation, as will appear in the sequel, to my own mind the phrase has never been intellectually satisfying. I cannot regard time *per se* as a causal factor; and the words only suggest that there is some third quantity varying with the time to which the changes in both the observed variables are due—as in the argument of the imaginary rationalist above. But what one feels about such a correlation is, not that it must be interpreted in terms of some very indirect catena of causation, but that it has no meaning at all; that in non-technical terms it is simply a fluke, and if we had or could have experience of the two variables over a very much longer period of time we would not find any appreciable correlation between them. But to argue like this is, in technical terms, to imply that the observed correlation *is* only a fluctuation of sampling, whatever the ordinary formula for the standard error may seem to imply: we are arguing that the result given by the ordinary formula is not merely wrong, but very badly wrong.

When we find that a theoretical formula applied to a particular case gives results which common sense judges to be incorrect, it is generally as well to examine the particular assumptions from which

it was deduced, and see which of them are inapplicable to the case in point. In obtaining the formula for the standard error we assume, to speak as before in terms of drawing cards from a record : (1) that we are drawing throughout from the same aggregate and not taking one sample from one aggregate, a second sample from another aggregate, and so on ; (2) that every card in each sample is also drawn from the same aggregate, in such a way that the 1st, 2nd, . . . nth cards in any sample are each equally likely to be drawn from any part of the aggregate, not the first card from one batch, the second from another, and so on ; (3) that the magnitude of x drawn on, say, the second card of the sample is quite independent of that on the first card, and so on for all other pairs in the sample ; and similarly for y ; there must be no tendency for a high value of x on the first card drawn to imply that the value of x on the second card will also probably be high ; (4) in order to reduce the formula to the very simple form given, we have also to make certain assumptions as to the form of the frequency-distribution in the correlation table for the aggregate from which the samples are taken.

In the particular case considered and in many similar cases there are two of these assumptions—leaving aside the fourth as comparatively a minor matter—which quite obviously do not apply, namely, the related assumptions (2) and (3). Our data necessarily refer to a *continuous* series of years, and the changes in both variables are, more or less, continuous. The proportion of marriages celebrated in the Established Church falls without a break for years together ; only a few plateaus and little peaks here and there interrupt the fall. The death-rate, it is true, shows much larger and more irregular fluctuations from year to year, but there is again a steady tendency to fall throughout the period ; only one rate (the last) in the first half of the years chosen, 1866–88, is below the average, only five in 1889–1911 are above it. Neither series, obviously, in the least resembles a random series as required by assumption (3).*

But can this breach of the assumed conditions render the usual formula so wholly inapplicable as it seems to be ? May it not merely imply, the reader may be inclined to question, some comparatively slight modification ? Even if the standard error by the usual formula were doubled, this would still leave the correlation

* The point that the usual formula for the standard error simply does not apply when we are dealing with correlations between time-series, has been made by Professor Persons ; *cf.* his chapter on Time-Series in the *Handbook of Mathematical Statistics*, ed. by H. L. Rietz, p. 162. *Cf.* also Professor Secrist's remarks in the chapter on Time-Series of the new edition of his *Introduction to Statistical Methods* (1925), pp. 464–65.

almost certainly significant. The special case considered in the next section will suffice to show that when the successive x's and y's in a sample no longer form a random series, but a series in which successive terms are closely related to one another, the usual conceptions to which we are accustomed fail totally and entirely to apply.

SECTION II.—*The correlation between simultaneous segments of two variables that are simple harmonic functions of the time, of the same period but differing by a quarter-period in phase; and the frequency-distribution of correlations for random samples of such segments.*

To clarify our ideas, let us consider a case in which each of our variables is some simple mathematical function of the time. A very general form of function to take would be the polynomial

$$y = a + bt + ct^2 + dt^3 + \ldots$$

But this is an inconvenient function for our present purpose, since it compels us to choose particular arbitrary values for the parameters a, b, c, etc.; nor is it a natural function to take as representing the changes in, say, some economic variable, over a long period of time, since y becomes infinite with t. A simple harmonic function of the time will be much better adapted to our purpose. Suppose, then, that the upper curve in Fig. 2 represents the changes in the first variable over some long period of time, say, many centuries—some period very much longer than any for which we are likely to

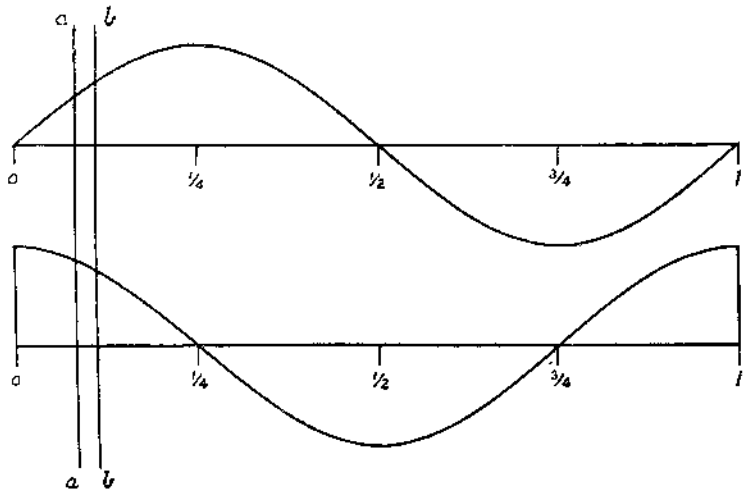

FIG. 2.—Two sine curves differing by a quarter-period in phase, and consequently uncorrelated when the correlation is taken over a whole period.

have statistics. Further, suppose that the lower curve, which is precisely similar to the first, except that it differs by a quarter-period in phase, represents the course of the second variable. Then it is evident that if we are given the two curves over a whole period, or any number of whole periods, the correlation between them is zero, for positive deviations in the one occur equally frequently with positive and with negative deviations in the other. But in actual fact, if the whole period 0 to 1 represents many centuries of time, our statistics will cover no more than some very short interval of the whole period, such as that enclosed between the two verticals *aa*, *bb*. This interval is so short that the segments of the two curves enclosed between *aa*, *bb*, are very nearly straight lines, the upper one rising, the lower one falling: the correlation between the corresponding observations will therefore be something very closely approaching -1.

Suppose the interval to become infinitesimally short so that the segments of the two curves may be taken as strictly linear, and let us trace the changes in the correlation coefficient as the centre of the interval moves across the figure from left to right. If the centre of he interval is placed at 0, the correlation must be zero, since the segment of the lower curve is horizontal and the values of the second variable are therefore the same for all values of the first. But as soon as the centre of the interval moves just to the right of 0, the segment of the upper curve is rising and that of the lower curve falling, so that the correlation becomes -1. This value is maintained until the centre of the interval passes over the point $t = \frac{1}{4}$, when the correlation rises abruptly again to zero, as in Fig. 3. As soon

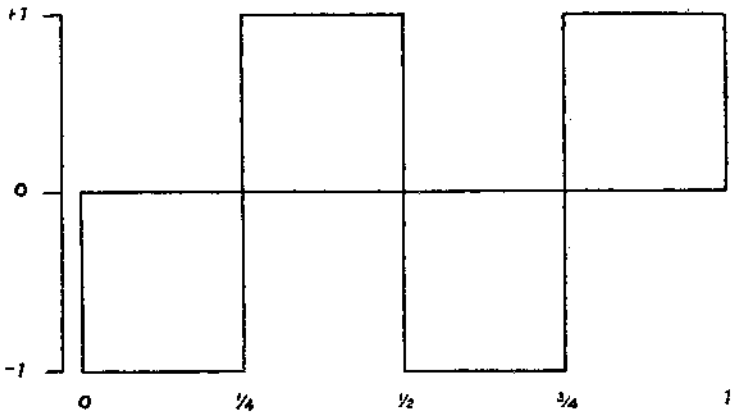

FIG. 3.—Variation of the correlation between two simultaneous infinitesimal elements of the harmonic curves of Fig. 2, as the centre of the element is moved across from left to right.

as the centre of the interval has passed this point the segments of both curves are falling, and the correlation is therefore + 1. This value is maintained until the centre of the interval reaches the half-period, when the cycle repeats itself: Fig. 3 shows the complete course of affairs.

It is quite possible to imagine that our experience covers no more than a practically infinitesimal interval out of the whole period supposed, and the centre of that interval—the mid-point of our experience—will be equally likely to fall at any point between the times 0 and 1. If this is so. what will be the frequency-distribution of correlations for a series of such chance experiences? Evidently, from Fig. 3, + 1 and — 1 are the only values of the correlation that occur with finite frequency, and each of these values holds good over one-half of the entire range on which the centre of the interval may fall. Hence the frequency-distribution has burst outwards, as it were, into an ordinate at + 1 and an equal ordinate at — 1: no intermediate values of r are possible.

If the interval over which we had experience, instead of being infinitesimal, covered just an entire period, the correlation would be zero: *i.e.*, the frequency-distribution of values of r on taking a series of random samples each of the length of a whole period would be simply an ordinate at zero. How, then, does the frequency-distribution for the first case pass into the frequency-distribution for the second case, as the length of the sample interval is gradually increased from something infinitesimally small up to the length of a period?

To solve this problem, it is first of all necessary to calculate curves like Fig. 3, showing, for any length of interval chosen, the values of r as the centre of the interval passes across the curves of Fig. 2 from left to right. As the curves are symmetrical, however, and repeat themselves, it is only necessary to carry out the calculations for one-eighth of the whole period. Fig. 4 shows such curves (the vertical scale being reversed as compared with Fig. 3 for convenience) when the interval is one-tenth, three-tenths, five-tenths, seven-tenths, and nine-tenths respectively of the period: the formulæ and method of calculation will be found in Appendix I. The first effect of lengthening the interval from something infinitesimally small up to 0·1 of a period is only slightly to round off the corners of the rectangles of Fig. 3, and quite slightly to decrease the maximum correlation attainable; it is not until the sample-interval becomes as large as half the period, or thereabouts, that the contours of the curve round off and the maximum undergoes a rather sudden drop. To obtain from any one of these curves the frequency-

distribution of values of r that would be given by placing the centre of the interval at random, it being equally likely to fall at any epoch in the whole period, we mark off along the base the abscissæ at which r attains, say, the values 0, 0·1, 0·2, 0·3, 0·4. . . .

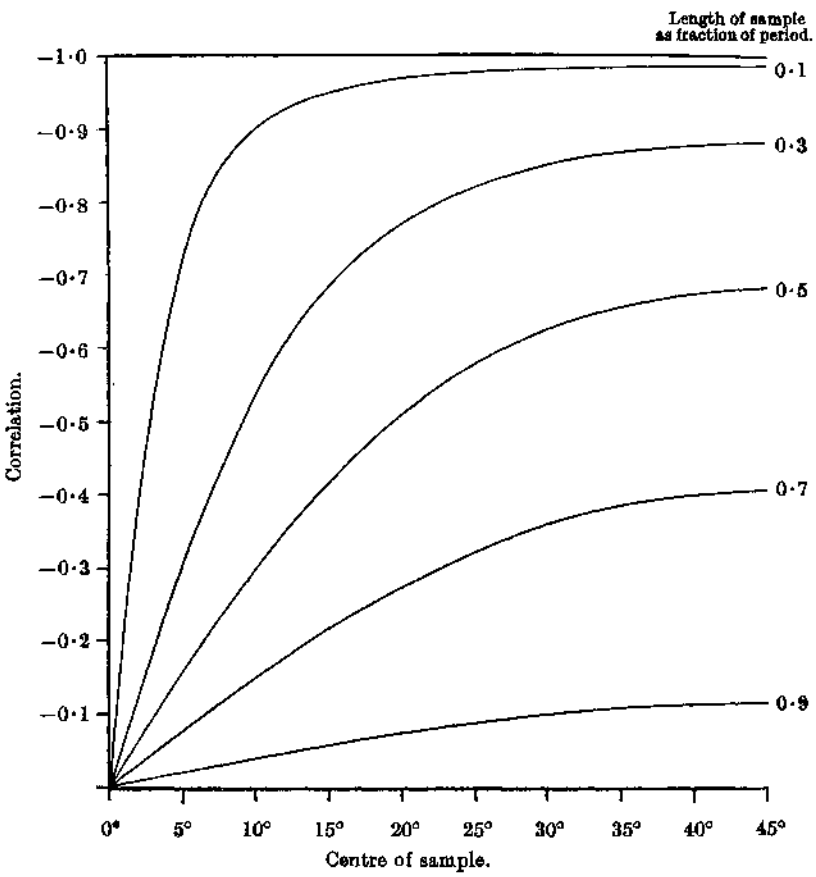

FIG. 4.—Variation of the correlation between two simultaneous finite elements of the harmonic curves of Fig. 2, when the length of the element is 0·1, 0·3, . . . , 0·9 of the period, as the centre of the element is moved across from left to right; only one-eighth of the whole period shown.

If these points are t_0, t_1, t_2, t_3, t_4, etc., the frequencies of correlations between the limits 0–0·1, 0·1–0·2, 0·2–0·3, 0·3–0·4, etc., are proportional to t_1-t_0, t_2-t_1, t_3-t_2, t_4-t_3, and so on. Graphic work would suffice to give a rough result, actually an algebraic interpolation formula was used (Appendix I). Inspection of the curves of Fig. 4

shows, however, what the form of the frequency-distributions must be, for evidently the steeper the curve in Fig. 4 the lower is the frequency. The maximum frequency must therefore always coincide with the maximum correlation attainable, where the curve is flat. Consequently all the curves must be U-shaped and, of course, symmetrical: the five distributions corresponding to the curves of Fig. 4 are shown in Figs. 5 to 9.

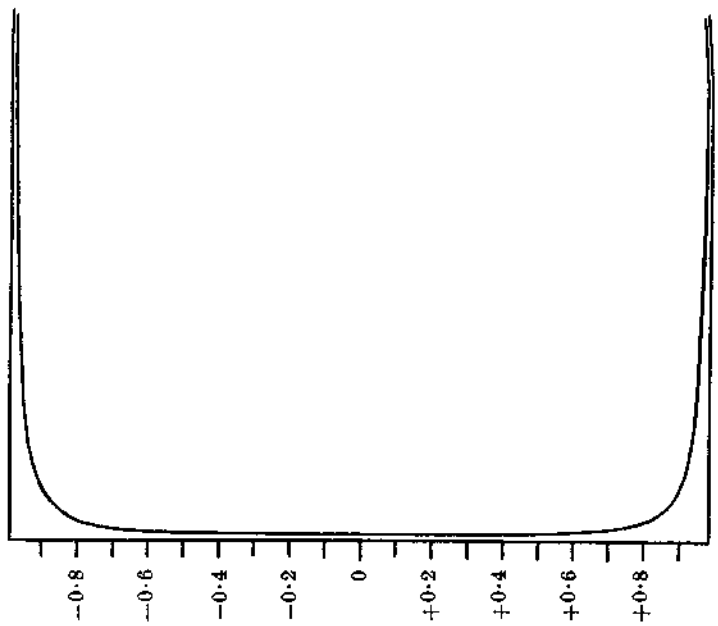

FIG. 5.—Frequency-distribution of correlations between simultaneous elements of the harmonic curves of Fig. 2, when the length of the element is 0·1 of the period. The following Figs. 6 to 9 show the change of form as the length of the element is increased from 0·1 to 0·9 by steps of 0·2.

The answer to our question, how the distribution of isolated frequencies at + 1 and − 1 closes up to the distribution of an isolated clump of frequency at zero, is then that the distribution first of all becomes a U-shaped distribution with limits not far from + 1 and − 1, and that these limits, at first gradually and then more rapidly, close in on zero; but *the distribution always remains U-shaped, and values of the correlation as far as possible removed from the true value* (*zero*) *always remain the most frequent.*

The result is in complete contrast with what we expect in sampling under the conditions usually assumed, when the successive values of either variable drawn for the sample are independent of one

another. In that case the values of r in successive samples may differ widely, but the mode tends to coincide with the "true" value in the aggregate from which the sample is drawn—zero in the present illustration. Here the values in the samples tend to diverge

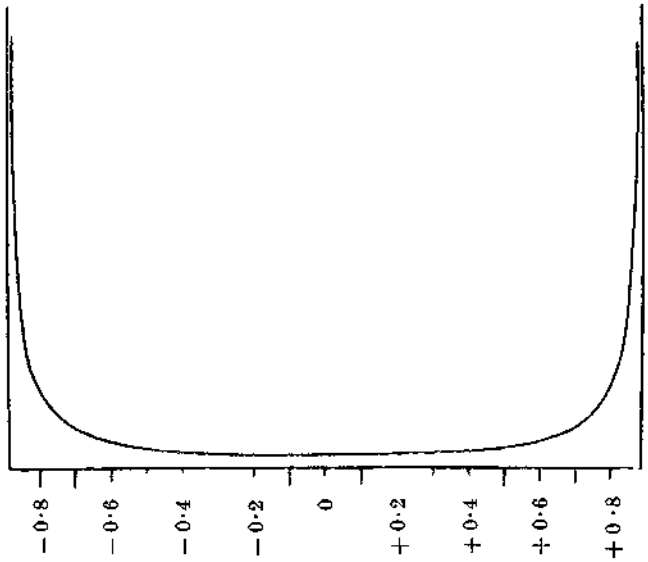

FIG. 6.—*Cf.* Fig. 5. Length of element, 0·3 of the period.

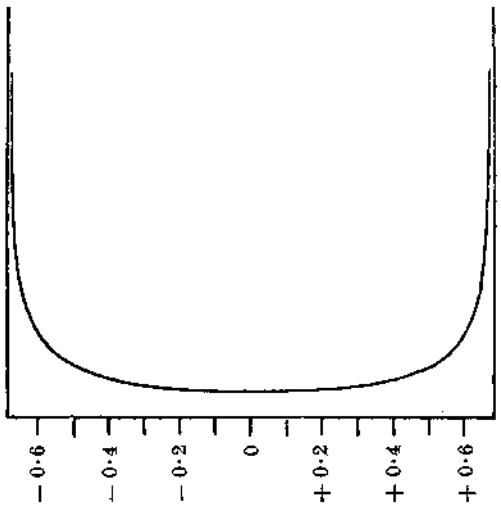

FIG. 7.—*Cf.* Fig. 5. Length of element, 0·5 of the period.

as widely as possible, in both directions, from the truth. We must evidently divest ourselves, in such a case, from all our preconceptions based on sampling under fundamentally different conditions. And evidently the result *suggests*—it cannot do more—the answer to the problem with which we started. We tend—it suggests—to get "nonsense-correlations" between time-series, *in some cases*, because *some* time-series are *in some way* analogous to the harmonic series that we have taken as illustration, and our available samples must be regarded as very small samples, if not practically infinitesimal, when compared with the length required to give the true correlation.

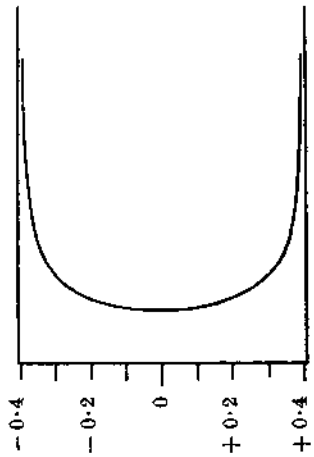

FIG. 8.—*Cf.* Fig. 5. Length of element, 0·7 of the period.

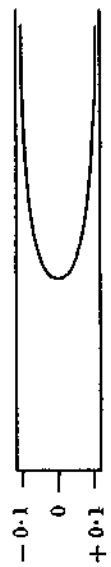

FIG. 9.—*Cf.* Fig. 5. Length of element, 0·9 of the period.

But what, it may be asked, is the frequency-distribution of values of r for small samples taken in the same way as for Figs. 5 to 9, if the correlation over a whole period is not zero? To answer this question by way of illustration I have taken two harmonic curves differing in phase by 60°, so that the correlation over a whole period is + 0·5, and have assumed the length of the samples to be one-fifth of the period. Fig. 10 shows the resulting frequency-distribution. It will be seen that it remains U-shaped, but has become asymmetrical. The limits are — 0·85055 and + 0·98221, and frequencies are much higher near the positive limit. Roundly

68 per cent. of the correlations are positive, 32 per cent. are negative, nearly 48 per cent. exceed + 0·9, only some 13 per cent. are less than — 0·8. We could only conjecture, in such a case, that the true correlation was positive, if we had a number of samples available, and noted that those giving a positive correlation were to those giving a negative correlation as about 2 to 1. Quite often,

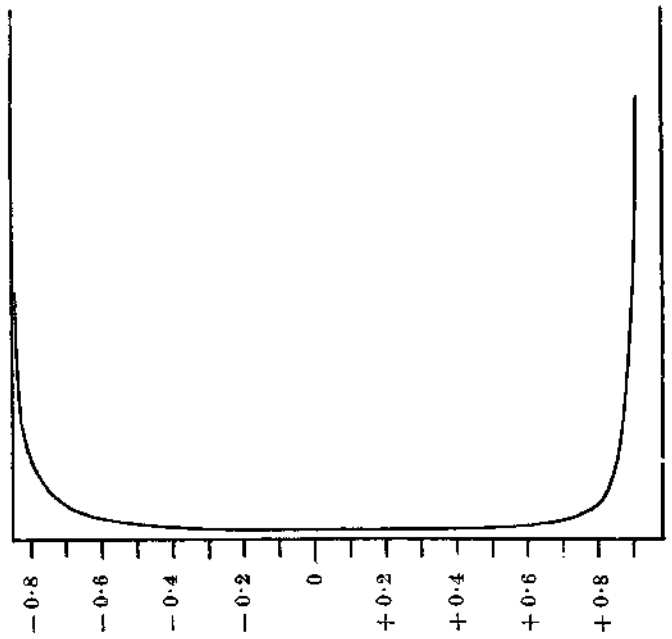

FIG. 10.—Frequency-distribution of correlations between two simultaneous elements of harmonic curves differing by 60° in phase (correlation over a whole period + 0·5) when the length of element is 0·2 of the period.

at about one trial in eight, a single sample might entirely mislead us by giving a high negative correlation exceeding 0·8. And, be it remembered, we have taken a fairly long sample, amounting to one-fifth of the period; if the complete period were something exceeding, say, 500 years, it is seldom that we would have such a sample at our disposal.

SECTION III.—*Deductions from Section II: classification of empirical series*

The work of Section II suggested that we tend, in some cases, to get meaningless correlations between time-series, because some time-series are in some way analogous to the harmonic series that

we took as illustration. The question has now to be answered, what is the precise analogy? What characteristics must two empirical series possess in order that small random samples, taken from them in the same way that we took the small samples from the sine-curves, may tend to give a U-shaped frequency-distribution for the resultant correlations?

The phenomenon is clearly related to the fact that a small segment of a sine-curve, taken at random, tends to be either *rising* or *falling*, not more or less level, and consequently tends to give high correlations of either sign with other segments taken at random. How can we secure such conditions in an empirical series? Will it suffice if, as in such series as might be represented by the curves of Fig. 2, successive terms of the series are highly correlated with one another? Thus, suppose the whole period in Fig. 2 is 360 years, so that one year corresponds to 1°. Then, if we take the product-sum over an entire period, the correlation between the value of the variable in one year and the value in the next is cos 1°, or 0·99985; between the value in one year and that in the next but one, cos 2°, or 0·99939, and so on (*cf.* Appendix I, equation 6), the correlations running

r_1	... ...	0·99985	
r_2	... ...	0·99939	
r_3	... ...	0·99863	
r_4	... ...	0·99756	
r_5	... ...	0·99169	
r_6	... ...	0·99452	
r_7	... ...	0·99255	
r_8	... ...	0·99027	
r_9	... ...	0·98769	
r_{10}	... ...	0·98481	

I propose to term such correlations, r_1 between u_s and u_{s+1}, r_2 between u_s and u_{s+2}, etc., where u_s is the value of the variable in year s, the *serial correlations* for the given series.

Now will it suffice to give us a U-shaped distribution of correlations for samples from two empirical series, if the serial correlations for both of them are high, and positive at least as far as r_{n-1} where n is the number of terms in the sample? This will imply that if the first term in a sample is considerably above the average of the sample, the next following terms will probably be above the average also, and some later terms must correspondingly be below the average to compensate for this excess: the graph of the sample will then tend to show a certain trend downwards from left to right. Conversely, if the first term is below average, the graph will tend to show an upward trend from left to right. Hence, generally, the graph of a random sample taken from such a series will tend to show not merely random fluctuations about a horizontal line, but a trend either upwards or downwards. The result must be that

if we take two such random samples, the correlation between them will tend to be markedly positive or markedly negative, according as the two trends are of the same or of opposite signs. This suggests that the frequency-distribution of correlations will be widely dispersed and possibly tend to be bimodal. But will it tend to the extreme of bimodality, a definite U-shape?

Is there not something more concealed in the assumption of a harmonic function for Fig. 2? When we take a small sample out of either of the curves, such as that between the verticals *aa*, *bb* of the figure, the sample does not tend to show a more or less *indefinite* upward or downward trend; it moves upward or downward with a clear unbroken sweep. This must imply something more: if the curve is going up from year s to year $s+1$, it tends to rise further from year $s+1$ to year $s+2$, which is to say, that *first differences are positively correlated with each other*, as well as the values of the variable. For the sine-curve, in fact, we know that the first differences form a curve of the same period as the original: the serial correlations for the *first differences* are therefore precisely the same as those for the values of the variable, given above. This is a very important additional property. It suggests that, for random samples from two empirical series to give a U-shaped distribution of correlations, each series should not merely exhibit positive values for the serial correlations up to r_{n-1}, but their difference series should also give positive serial correlations up to the limit of the sample.

Let us now endeavour to make these ideas a little more definite. The usual theory of sampling is concerned only with the simplest case, the *random series*, for which the serial correlations are zero. If we take a number of samples of n observations out of such a series, it is familiar that the correlation between the deviations of any two observations *from the mean of the sample* is $-1/(n-1)$. If, then, the first term of the sample is above the mean of the sample, there is no definite tendency for the sample as a whole to show a downward trend, excluding the first term itself; for *all* the remaining terms have an *equal*, and that only a slight tendency to be below the average. Thus, I took the 60 sets of 10 random terms each, forming the experimental series A_0 to F_0 of the next section, worked out the deviation of every term in each sample from the mean of that sample, and then separated the samples into two groups: (*a*) those in which the first deviation was positive, (*b*) those in which the first deviation was negative. I found 28 of the former and 32 of the latter. Taking each group separately, I averaged separately the deviations of the 1st, 2nd . . . 10th terms. The standard deviations of all the terms being the same, and the

correlation of every term with every other being $-1/9$, if we call the mean of the positive deviations of the first term 1000, the most probable deviation of each of the others is $-1000/9$ or -111, as in Table I, col. 2. The average of the series in which the first deviation was positive gave the result shown in col. 3: the figures run rather irregularly, as the fluctuations of sampling are large, but there is no consistent deviation from expectation and clearly no consistent trend in terms 2 to 10. The average of the series in which the first deviation was negative, reversing signs all through for readier comparability, gave the result shown in col. 4; and finally, combining the two sets by reversing sign in the totals of the series with first deviations negative and adding to the totals of the set with first deviations positive, we have the general average of col. 5. The figures of neither col. 3, nor col. 4, nor col. 5 show any definite trend in terms 2 to 10. Selection of the first term does not bias the remainder of the sample, or give it any trend or "tilt" either upwards or downwards; the remaining terms are still random in their order.

TABLE I.—Deviations from the mean of the sample in samples of 10 terms from a random series, averaging separately samples in which the first deviation is positive and samples in which the first deviation is negative: average of first deviations taken as + 1000.

Term.	Expectation.	Experimental results.		
		First term +.	First term −.	Together.
(1)	(2)	(3)	(4)	(5)
1	+ 1000	+ 1000	+ 1000	+ 1000
2	− 111	− 155	− 113	− 132
3	− 111	− 470	+ 25	− 206
4	− 111	− 15	− 105	− 63
5	− 111	− 452	− 136	− 284
6	− 111	+ 300	+ 87	+ 186
7	− 111	− 321	− 190	− 251
8	− 111	− 137	+ 171	+ 27
9	− 111	+ 449	− 389	+ 2
10	− 111	− 199	− 351	− 280

Now suppose we take from a series of random terms (with the mean zero) a sample of ten terms $a, b, c, d, e, f, g, h, k, l$, and form from it, by successive addition, a new series a, $a + b$, $a + b + c \ldots$ In this new series the terms are correlated with each other, since each term contains the term before, but the differences are random. Let us find the correlations between deviations of the terms from

the mean of the sample. For our special case of 10 terms the mean is

$$a+0\cdot9b+0\cdot8c+0\cdot7d+0\cdot6e+0\cdot5f+0\cdot4g+0\cdot3h+0\cdot2k+0\cdot1l.$$

The deviations of the successive terms from the mean are then as given in Table II. The standard deviation of each deviation in a

TABLE II.—Coefficients of the terms in the deviations from the mean of the sample, in a sample of 10 terms from a series with random differences $a, b, c \ldots l$.

Term.	(1) b	(2) c	(3) d	(4) e	(5) f	(6) g	(7) h	(8) k	(9) l	Coefficient of s.d.
1	−0·9	−0·8	−0·7	−0·6	−0·5	−0·4	−0·3	−0·2	−0·1	1·688
2	+0·1	−0·8	−0·7	−0·6	−0·5	−0·4	−0·3	−0·2	−0·1	1·432
3	+0·1	+0·2	−0·7	−0·6	−0·5	−0·4	−0·3	−0·2	−0·1	1·204
4	+0·1	+0·2	+0·3	−0·6	−0·5	−0·4	−0·3	−0·2	−0·1	1·025
5	+0·1	+0·2	+0·3	+0·4	−0·5	−0·4	−0·3	−0·2	−0·1	0·922
6	+0·1	+0·2	+0·3	+0·4	+0·5	−0·4	−0·3	−0·2	−0·1	0·922
7	+0·1	+0·2	+0·3	+0·4	+0·5	+0·6	−0·3	−0·2	−0·1	1·025
8	+0·1	+0·2	+0·3	+0·4	+0·5	+0·6	+0·7	−0·2	−0·1	1·204
9	+0·1	+0·2	+0·3	+0·4	+0·5	+0·6	+0·7	+0·8	−0·1	1·432
10	+0·1	+0·2	+0·3	+0·4	+0·5	+0·6	+0·7	+0·8	+0·9	1·688

TABLE III.—Correlations between deviations from the mean of the sample, in a sample of 10 terms from a series with random differences.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	+1·	+0·81	+0·57	+0·26	−0·10	−0·42	−0·61	−0·66	−0·64	−0·58
2	+0·81	+1·	+0·73	+0·37	−0·04	−0·42	−0·65	−0·73	−0·71	−0·64
3	+0·57	+0·73	+1·	+0·61	+0·14	−0·32	−0·61	−0·72	−0·73	−0·66
4	+0·26	+0·37	+0·61	+1·	+0·48	−0·05	−0·43	−0·61	−0·65	−0·61
5	−0·10	−0·04	+0·14	+0·48	+1·	+0·41	−0·05	−0·32	−0·42	−0·42
6	−0·42	−0·42	−0·32	−0·05	+0·41	+1·	+0·48	+0·14	−0·04	−0·10
7	−0·61	−0·65	−0·61	−0·43	−0·05	+0·48	+1·	+0·61	+0·37	+0·26
8	−0·66	−0·73	−0·72	−0·61	−0·32	+0·14	+0·61	+1·	+0·73	+0·57
9	−0·64	−0·71	−0·73	−0·65	−0·42	−0·04	+0·37	+0·73	+1·	+0·81
10	−0·58	−0·64	−0·66	−0·61	−0·42	−0·10	+0·26	+0·57	+0·81	+1·

series of such samples will be the square root of the sum of the squares of the numerical coefficients, multiplied by the standard deviation of the original random series $a, b, c \ldots x$; it will be seen from the column on the right of Table II that the end terms are the most variable, the central terms the least variable, and the standard deviations are symmetrical about the centre of the sample. The product-sum for any pair of terms will be the sum of the products of corresponding numerical coefficients in the same column, multiplied by the square of the s.d. of the series $a, b, c \ldots x$, and hence the correlation will be given by dividing the sum of

the products by the product of the s.d. coefficients on the right of Table II. The resulting coefficients of correlation are shown in Table III. It will be seen that for terms which are closely adjacent at either end of the sample they are fairly high and positive, but for terms at opposite ends moderately high and negative. Thus, taking the correlations of the first term with the others, the correlation between deviations 1 and 2 is $+0{\cdot}81$, but between 1 and 3 drops to $+0{\cdot}57$. Between 1 and 5 there is a small negative correlation, and this negative correlation reaches a maximum of $-0{\cdot}66$ between deviations 1 and 8. The negative correlation then falls away slightly and is only $-0{\cdot}58$ between the first and last deviations 1 and 10. Evidently the general effect of this arrangement of correlations must be, as already argued, to give the sample *as a whole* a *tendency* to be tilted one way or the other as the first term is above or below average. If the first term is, say, 1 unit above the mean of the sample, the mean deviations of the others will be given by their regressions on the first term, which can be found from the correlations and s.d.'s already given. Multiplied by 1000 these are shown in column 2 of Table IV, and it will be seen that they give a continuous descent from the + 1000 of term 1 to − 579 for term 10.

TABLE IV.—Deviations from the mean of the sample in samples of 10 terms from a series with random differences, averaging separately samples in which (*a*) first deviation is +, (*b*) first deviation is −, (*c*) last deviation is +, (*d*) last deviation is −. The average of first or last deviations, respectively, called + 1000.

Term.	Expectation.	Experimental results *a* and *b*.	Term.	Experimental results.	
				c and *d*.	Together.
(1)	(2)	(3)	(4)	(5)	(6)
1	+ 1000	+ 1000	10	+ 1000	+ 1000
2	+ 684	+ 681	9	+ 636	+ 658
3	+ 404	+ 367	8	+ 398	+ 383
4	+ 158	+ 144	7	+ 169	+ 157
5	− 53	− 98	6	− 56	− 76
6	− 228	− 300	5	− 217	− 257
7	− 368	− 361	4	− 459	− 411
8	− 474	− 286	3	− 516	− 404
9	− 544	− 528	2	− 545	− 537
10	− 579	− 619	1	− 411	− 512

This result was again checked by experiment. From the experiments described in the next section of the paper 60 sets of 10 terms each were available from series with random differences

(series A_1 to F_1). The deviations of the terms in each sample from the mean of that sample were worked out, and the samples were then arranged in two groups as before according as the first deviation was positive or negative, and totalled. Reversing signs in the second set, and combining with the first as for the last column of Table I, and then dividing through by the total of the first term, the figures of column 3 of Table IV were obtained. It will be seen that the agreement with expectation is very fair; the correlations are not high and fluctuations of sampling are large. But a second test can be made. Since the correlations and s.d.'s are symmetrical, exactly the same result is to be expected if we re-sort the samples according as the *last* term is positive or negative, and then take ratios on the average of the *last* term. The work was done in the same way, and the results are given in column 5 of Table IV. Combining the data on which columns 3 and 5 were based, the ratios of column 6 were obtained. Columns 3, 5 and 6 are all in fair agreement with expectation, and show exactly the same thing. In marked contrast with the random series, the sample from the series with random differences shows a clear tendency to tilt one way or the other as a whole; and hence one random sample from such a series will tend to give more or less marked correlations, either positive or negative, with another. But it must be remembered that this *tendency* of the sample to be tilted one way or the other as a whole *is* only a tendency; it is sufficiently clearly marked to attract attention during experimental work, but by no means stringent, as is evident from the moderate values of the correlations in Table III.

We have now to consider the third type of series that has suggested itself, the series in which not merely successive terms, but also successive first differences, are positively correlated. The simplest way in which to construct such a series seems to be to sum the last series again, term by term, *i.e.*, to form the second sum of the random series instead of the first. Taking 10 terms only, the second sum will run—

$$a$$
$$2a + b$$
$$3a + 2b + c$$
$$4a + 3b + 2c + d$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$10a + 9b + 8c + 7d + 6e + 5f + 4g + 3h + 2k + l,$$

and the mean is—

$$5{\cdot}5a + 4{\cdot}5b + 3{\cdot}6c + 2{\cdot}8d + 2{\cdot}1e + 1{\cdot}5f + g + 0{\cdot}6h + 0{\cdot}3k + 0{\cdot}1l.$$

The deviations and coefficients of the s.d.'s of the several terms are then as shown in Table V, and Table VI gives the correlations calculated in the same way as before. It will be seen that the standard deviations are now no longer symmetrical about the centre of the sample, the s.d. of term 10 being much larger than that of term 1; while the general arrangement of the correlations is similar to that of Table III, the correlations are much higher, and again they are not symmetrical with respect to the two ends of the sample. But the magnitude of the correlations is now *very* high. Between terms 1 and 2 there is a correlation of 0·992, and between terms 9 and 10 a correlation of 0·991. The maximum negative correlation is that between terms 2 and 8 or 3 and 9, and is −0·988. The tendency of the sample to "tilt" as a whole becomes now very clearly marked, so clear that it becomes quite evident on forming even a few experimental samples in this way.

TABLE V.—Coefficients of the terms in the deviations from the mean of the sample, in a sample of 10 terms from a series of which the second differences are random.

Term.	(1) *a*	(2) *b*	(3) *c*	(4) *d*	(5) *e*	(6) *f*	(7) *g*	(8) *h*	(9) *k*	(10) *l*	Coefficient of s.d.
1....	−4·5	−4·5	−3·6	−2·8	−2·1	−1·5	−1·0	−0·6	−0·3	−0·1	2·635
2....	−3·5	−3·5	−3·6	−2·8	−2·1	−1·5	−1·0	−0·6	−0·3	−0·1	2·311
3....	−2·5	−2·5	−2·6	−2·8	−2·1	−1·5	−1·0	−0·6	−0·3	−0·1	1·877
4....	−1·5	−1·5	−1·6	−1·8	−2·1	−1·5	−1·0	−0·6	−0·3	−0·1	1·357
5....	−0·5	−0·5	−0·6	−0·8	−1·1	−1·5	−1·0	−0·6	−0·3	−0·1	0·801
6....	+0·5	+0·5	+0·4	+0·2	−0·1	−0·5	−1·0	−0·6	−0·3	−0·1	0·492
7....	+1·5	+1·5	+1·4	+1·2	+0·9	+0·5	—	−0·6	−0·3	−0·1	0·971
8....	+2·5	+2·5	+2·4	+2·2	+1·9	+1·5	+1·0	+0·4	−0·3	−0·1	1·738
9....	+3·5	+3·5	+3·4	+3·2	+2·9	+2·5	+2·0	+1·4	+0·7	−0·1	2·597
10....	+4·5	+4·5	+4·4	+4·2	+3·9	+3·5	+3·0	+2·4	+1·7	+0·9	3·513

The experimental series with correlated differences were not as a fact formed in the way suggested, but the method used is equivalent in the present respect for samples of 10 observations (*cf.* Appendix II, under heading C, pp. 61–2). Of the 60 samples (series A_2 to F_2) only 6 gave first and last deviations of the same sign. The regressions obtained from Tables V and VI on the first term and the last respectively were used to obtain columns 2 and 5 of Table VII, and the experimental results are compared with these figures in columns 3 and 6 of the same table, which is analogous to Table IV. Given the first deviation, the last term should show a greater negative deviation, and in the experimental results it is greater, though not so much greater as it should be. Given the

TABLE VI.—Correlations between deviations from the mean of the sample, in a sample of 10 terms from a series of which the second differences are random.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	+ 1·	+ 0·992	+ 0·967	+ 0·907	+ 0·707	− 0·315	− 0·941	− 0·976	− 0·958	− 0·935
2	+ 0·992	+ 1·	+ 0·987	+ 0·938	+ 0·752	− 0·271	− 0·940	− 0·988	− 0·976	− 0·956
3	+ 0·967	+ 0·987	+ 1·	+ 0·974	+ 0·819	− 0·182	− 0·915	− 0·990	− 0·988	− 0·973
4	+ 0·907	+ 0·938	+ 0·974	+ 1·	+ 0·913	+ 0·012	− 0·841	− 0·962	− 0·981	− 0·977
5	+ 0·707	+ 0·752	+ 0·819	+ 0·913	+ 1·	+ 0·360	− 0·589	− 0·803	− 0·869	− 0·891
6	− 0·315	− 0·271	− 0·182	+ 0·012	+ 0·360	+ 1·	+ 0·507	+ 0·213	+ 0·072	− 0·005
7	− 0·941	− 0·940	− 0·915	− 0·841	− 0·589	+ 0·507	+ 1·	+ 0·938	+ 0·869	+ 0·819
8	− 0·976	− 0·988	− 0·990	− 0·962	− 0·803	+ 0·213	+ 0·938	+ 1·	+ 0·982	+ 0·955
9	− 0·958	− 0·976	− 0·988	− 0·981	− 0·869	+ 0·072	+ 0·869	+ 0·982	+ 1·	+ 0·991
10	− 0·935	− 0·956	− 0·973	− 0·977	− 0·891	− 0·005	+ 0·819	+ 0·955	+ 0·991	+ 1·

last deviation, on the other hand, the negative deviation of the first term should be considerably less, and in the experiment it is less, but not so much less as it should be. But the broad agreement with theory is evident; fluctuations of sampling from series to series are large as before.

TABLE VII.—Deviations from the mean of the sample, in samples of 10 terms from a series of which the second differences are random, averaging separately samples in which (*a*) first deviation is +, (*b*) first deviation is −, (*c*) last deviation is +, (*d*) last deviation is −. The average of first or last deviations respectively called 1000.

Term.	Expectation.	Experimental result *a* and *b*.	Term.	Expectation.	Experimental result *c* and *d*.
(1)	(2)	(3)	(4)	(5)	(6)
1	+ 1000	+ 1000	10	+ 1000	+ 1000
2	+ 870	+ 811	9	+ 733	+ 763
3	+ 689	+ 597	8	+ 473	+ 528
4	+ 467	+ 391	7	+ 226	+ 289
5	+ 215	+ 144	6	− 1	+ 49
6	− 59	− 107	5	− 203	− 173
7	− 347	− 360	4	− 377	− 376
8	− 644	− 607	3	− 520	− 542
9	− 945	− 829	2	− 629	− 697
10	− 1247	− 1040	1	− 702	− 841

Now this argument has led us to a remarkable result, which at first sight may seem paradoxical: namely, that for the present purpose we are really only concerned with the serial correlations for the *differences* of our given series, and not with the serial correlations of those series themselves. For if we take a long but finite series of random terms and sum it, the serial correlations for the sum-series are not determinate and will vary from one such series to another: and yet all such series evidently have the same characteristics from the present standpoint. And obviously again, if we form the second-sum of a long but finite series of random terms, the serial correlations for the second-sum are not determinate and will vary from one such series to another, and yet all such series, from the present standpoint, have the same characteristics. If in either case we make the series indefinitely long, all the serial correlations will tend towards unity, but the samples remain just the same as they were before, so evidently we cannot be concerned with the mere magnitude of the serial correlations themselves: they are dependent on the length of the series.

Let the serial correlations for the series itself be

$$1, r_1, r_2, r_3, r_4, \ldots . r_k,$$

and for the difference series

$$1, \rho_1, \rho_2, \rho_3, \rho_4, \ldots . \rho_k,$$

then it is shown in Appendix II that for a long series in which we may neglect the effect of the end-terms,

$$\rho_k = \frac{2r_k - r_{k+1} - r_{k-1}}{2(1-r_1)} = -\frac{1}{2(1-r_1)} \Delta^2(r_{k-1}).$$

If now we are given the ρ's, all that we know is the *form* of the function

$$r_k = \phi(k).$$

If the ρ's are all zero, or the sum-series is the sum of a random series, r_k is a linear function of k. If all that we know is that the ρ's are positive, all that we can say about the r's is that the graph of the r's to k as abscissa must give a curve that is concave downwards. If more definitely we know that the ρ's are a decreasing arithmetical series, the graph of the r's is a cubic parabola. If the ρ's form an oscillatory series, the graph of the r's must exhibit oscillations (*cf.* Fig. 19, p. 43).

The serial correlations up to r_{10} were worked out for three series of 100 terms with random differences, and the results are shown graphically in Fig. 11: the data will be found in Appendix II, Table A. The series A_1 and C_1 give very fair fits to straight lines: B_1 is rather more erratic—but it must be remembered that all are rather short series. It will be noted from the figure how greatly the actual magnitudes of the serial correlations differ for the three series: in A_1, r_{10} is $+0{\cdot}776$; in B_1, $+0{\cdot}242$; in C_1, $+0{\cdot}519$.

The serial correlations were also worked out for three series of 100 terms in which the difference correlations were a descending arithmetic series, and these results are shown in Fig. 12, the data being given in Table B of Appendix II. In this case the observed correlations for all three series lie fairly closely round cubics of the required type. Note again how largely the actual values of the serial correlations differ from series to series. It is the *form* of the curve alone which determines the values of the difference correlations. The fact that the concavity faces downwards indicates at once to the eye that the sign of the difference correlations is positive, but the eye alone can hardly judge what function ρ_k is of k.

Statistical series may evidently be classified by the nature of the serial correlations, and such a classification will be important from

the standpoint of the present enquiry. I suggest the following classification and technical terms :—

Random series.—Series for which all the serial correlations, in an indefinitely long series, are zero.

Conjunct series.—Series for which all the serial correlations are positive. We can readily imagine ideal cases for which, in an

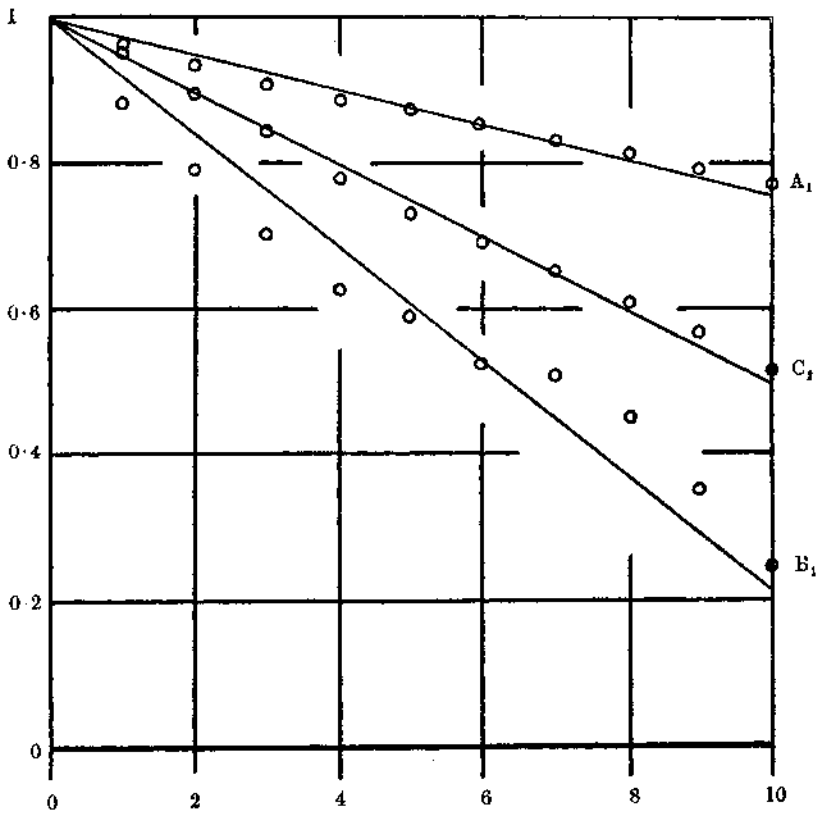

FIG. 11.—Serial correlations up to r_{10} for three experimental series (of 100 terms) with random differences.

indefinitely long series, r_k is positive for all values of k, but in a finite series r_k decreases with k and becomes negative. For practical specification we are only concerned with a finite number of serial correlations, and may speak of a series as "conjunct up to r_k." If, for example, some statistical variable is strictly periodic with a period of 1,000 years, annual data concerning it form, properly speaking, a periodic series. But if we have data for no more than a

century or two we may only recognize it as a conjunct series, "conjunct up to r_{50}" or so.

Disjunct series.—Series for which the serial correlations are all negative. The ideal case is possible (*cf.* Appendix II, sub-head D, pp. 62–3), but the conditions of consistence imply stringent limitations on the values of the correlations. For the random series ρ_1, for

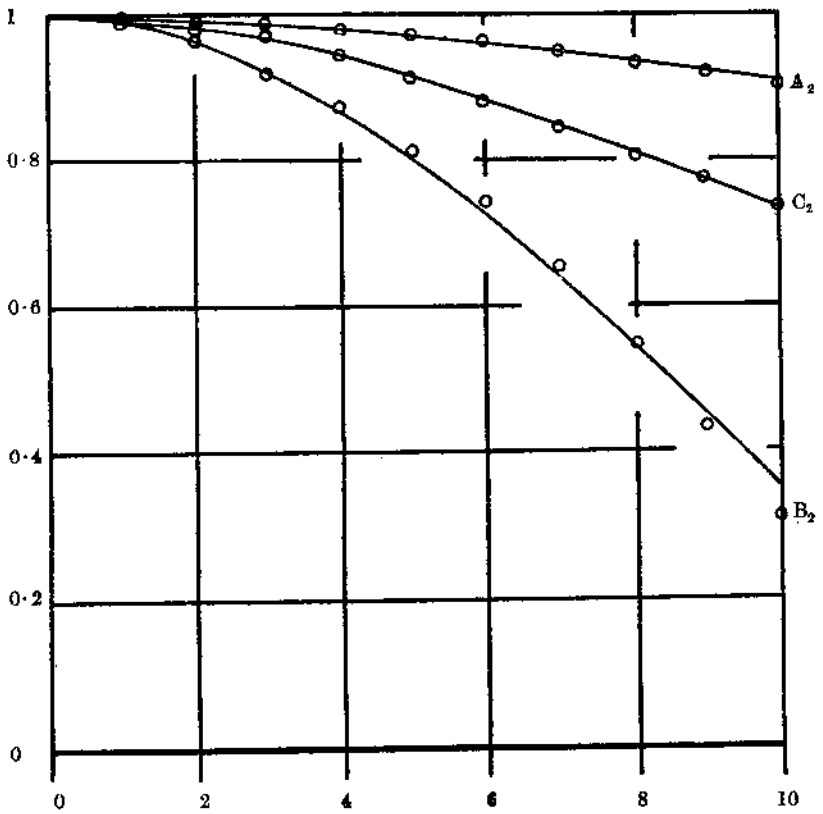

FIG. 12.—Serial correlations up to r_{10} for three experimental series (of 100 terms) with positively correlated (conjunct) differences.

adjacent first differences, is — 0·5 and all the remaining correlations are zero, so the differences of a random series form a very simple type of disjunct series.

Oscillatory series.—Series for which the serial correlations change sign, being alternately positive and negative. These are very important in many forms of statistics (quite possibly they are the most frequent form), but I am not able to consider them in the

present Address, though I take one series with oscillatory differences as an illustration for analysis (Section V). The truly periodic series is a special case; an oscillatory series is not necessarily periodic. If, for example, we take a random series and form a derivative series by calculating the difference of u_s from the mean of the terms u_{s-r} to u_{s+r}, the derived series is oscillatory, but it is not periodic.

These are simple types; but clearly in the endless variety presented by facts we may expect to meet with compound series of any type, *e.g.*, conjunct series with an oscillatory series superposed (*cf.* Section V). It is also imaginable, obviously, that we might for such purposes of classification desire to go further and consider the serial correlations for second, third or nth differences.

In the immediately following work we are concerned only with *random series*, to which the ordinary theory of sampling applies, and two sub-types of *conjunct series*—

(*a*) *conjunct series the differences of which are random.*

(*b*) *conjunct series the differences of which are themselves conjunct series.*

We have concluded that if we take random samples from two conjunct series and work out the correlations between them, series of type (*a*) will tend to give a distribution of correlations certainly divergent from the distribution given by random samples from random series, more scattered, and *possibly* bimodal: series of type (*b*) will tend to give an entirely divergent and probably U-shaped frequency-distribution of the correlations. In the next section an experimental investigation is described to test these tentative conclusions.

As the distinctions seem to me of possible importance for much statistical work, I give in Figs. 13–15 illustrations of the three types—random series, conjunct series with random differences, and conjunct series with conjunct differences. Fig. 13 shows two random series; there is no secular trend, and the whole movement is highly irregular. The graphs are not, to the eye at least, very unlike graphs of some annual averages in meteorological data. Fig. 14 gives graphs of two series with random differences. We now get a marked "secular movement," with irregular oscillations superposed on it. Finally, Fig. 15 gives two graphs of series with conjunct differences. The curves are smoothed out, the secular movements or long waves are conspicuous, but there are no evident oscillations of short duration. The graphs of both Fig. 14 and Fig. 15 could, I think, be matched from statistical data, but it is quite possible that what looked a good match to the eye would not seem at all a good match when subjected to strict analysis.

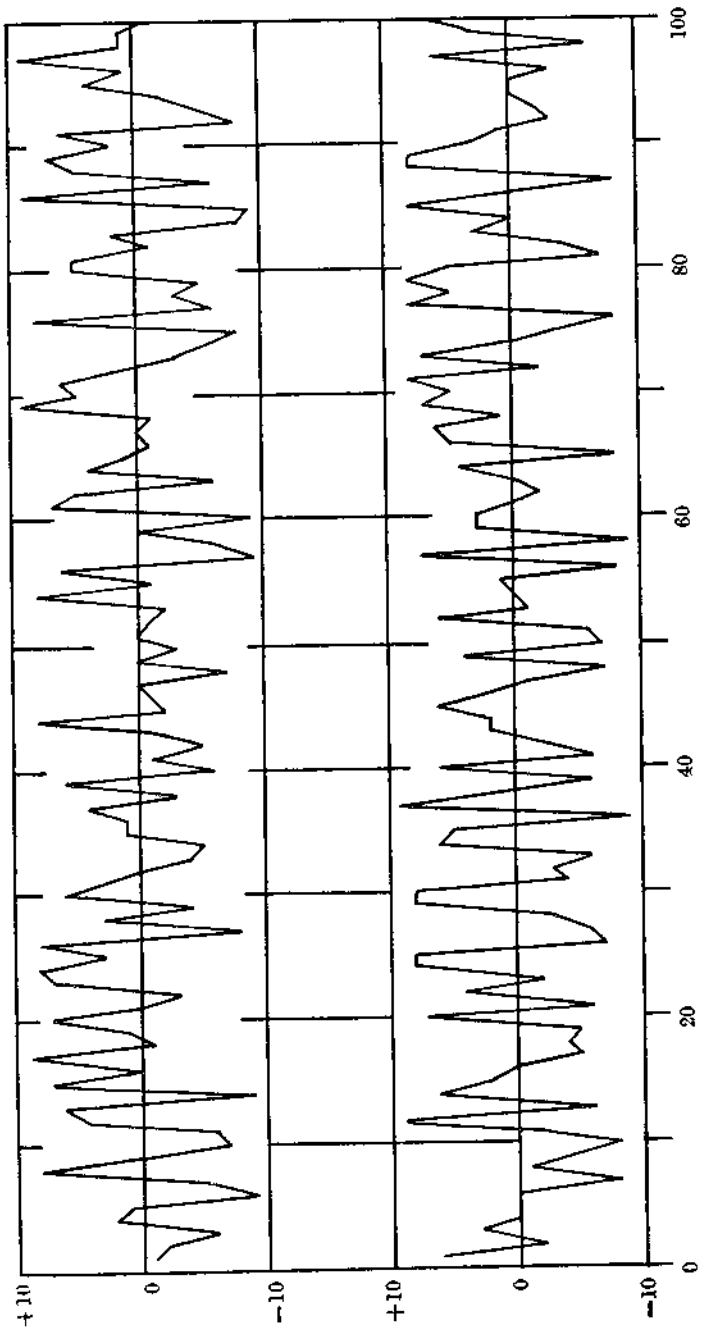

FIG. 13.—Two random series.

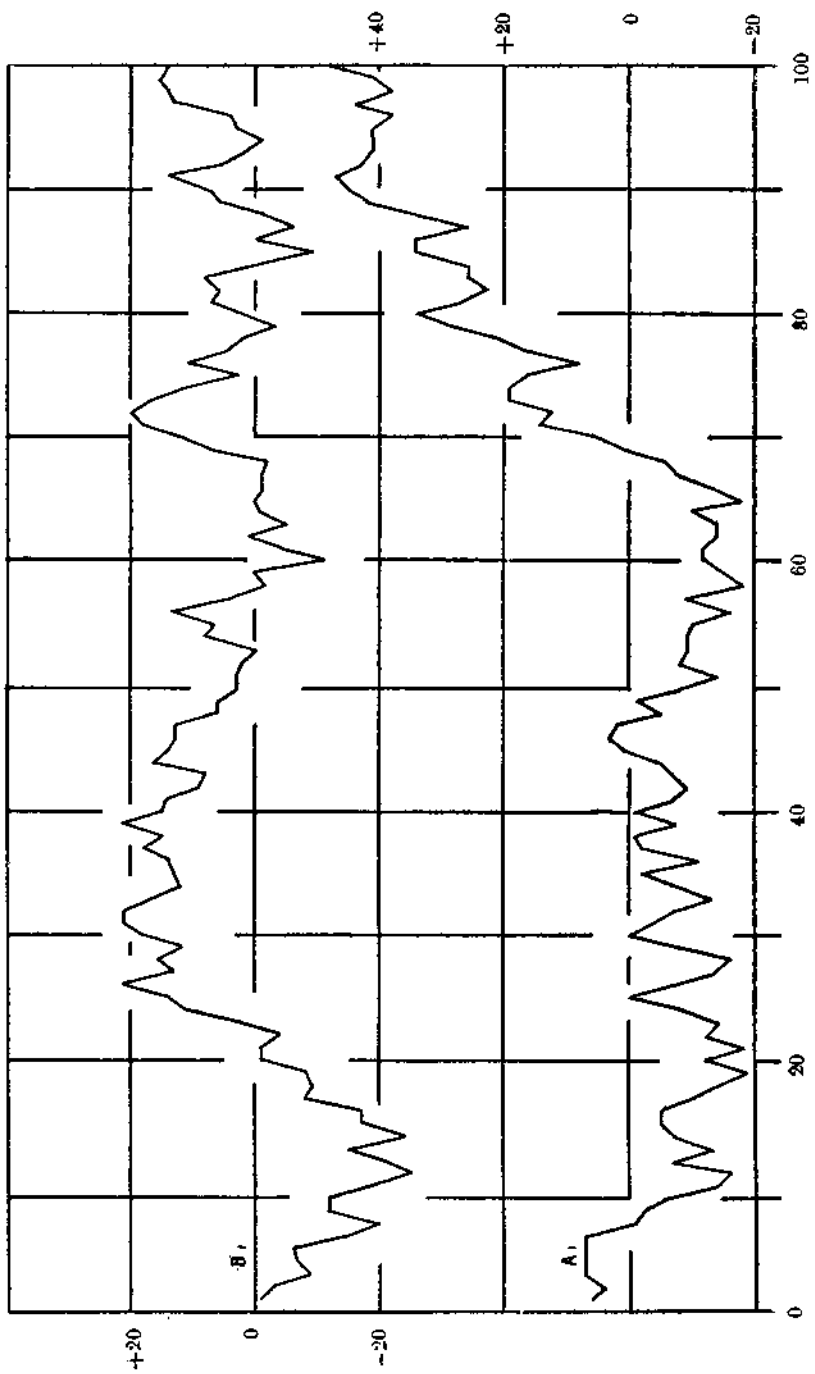

FIG. 14.—Two series with random differences (conjunct series with random differences).

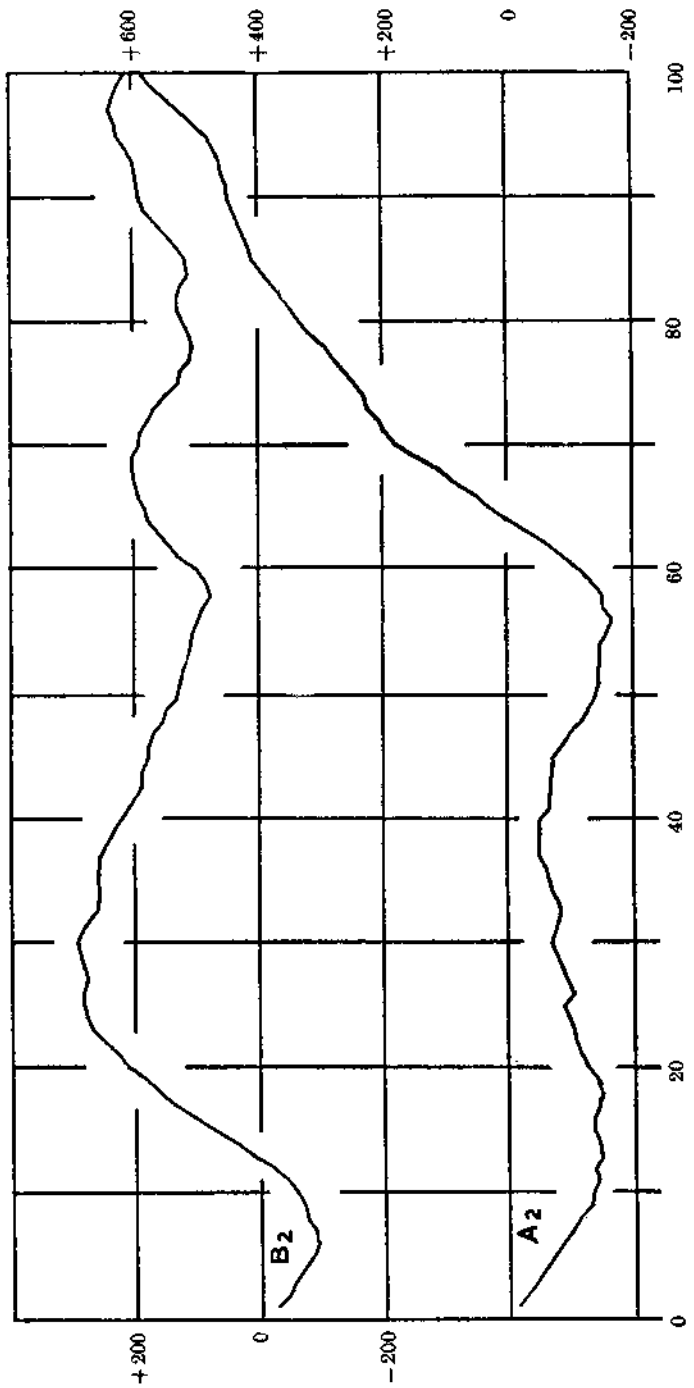

FIG. 15.—Two series with positively correlated differences (conjunct series with conjunct differences).

SECTION IV.—*Experimental investigations.*

When the main ideas developed in Sections II and III had been reached, I decided to carry out experimental tests. The fundamental random series were formed by drawing cards from a pack in the way described in a former paper*—in fact, the record used in that paper was employed as one series. The court cards were removed from two patience packs; black cards were reckoned as positive, red cards as negative and tens as zeros, so that the frequency-distribution in the pack was uniform from -9 to $+9$, with the exception that there were two zeros. The mean of this distribution is zero, and the standard deviation is $\sqrt{28\cdot5}$, or $5\cdot3385$. The pack was shuffled and a card drawn; thoroughly shuffled again and another card drawn, and so on. Every precaution was taken to avoid possible bias and ensure randomness. The use of a double pack helps, I think, towards this, as the complete series is repeated four times. Shuffling was very thorough after every draw; after shuffling, the pack was cut and, say, the fifth card from the cut taken as the card drawn, so as to avoid any possible tendency of the cards to cut at a black rather than a red, or a ten rather than an ace, and so on.

When the random series had been obtained, a series with random differences was calculated from it by adding term by term from the beginning. To obtain a series with correlated differences, the natural procedure would have been, as already suggested in Section III, to go on and obtain the second sum of the random series. But at the time the experiments were begun this did not strike me, and it seemed desirable to work with known correlations between the differences. I therefore added up the random series by successive groups of 11 terms, u_0 to u_{10}, u_1 to u_{11}, u_2 to u_{12}, and so on; this gave the difference series, and adding term by term gave the series with correlated differences, the serial correlations between the differences being 10/11, 9/11, 8/11, . . . 1/11, and thenceforward zero.

But the process used for sampling was very slow, and to shorten both the work of sampling and the arithmetic I adopted a procedure which was certainly very effective to that end, but proved itself by no means desirable in other respects; it tended, in fact, to give lumpy and irregular frequency-distributions. Had I fully realized its disadvantages as well as its advantages, I might rather have chosen to adopt the straightforward method of obtaining completely independent samples for every correlation to be calculated. This

* "On the Time Correlation Problem," *J.S.S.*, vol. lxxxiv, 1921; *cf.* pp. 517–18.

would have necessitated a much longer time for the investigation, but I had, in fact, to make one supplementary series of experiments by the better method. The procedure used was this for each type of series. I formed three series of 100 observations each, A, B, C. I then divided up each series into 10 sets of 10 observations. Finally, for the correlations I combined every set of A with every set of B (100 pairs), every set of A with every set of C (100 pairs), and every set of B with every set of C (100 pairs). I thus obtained 300 correlations each based on 10 observations, but only 30 completely independent sets of 10 observations were used in the whole set. As a control I carried out another set, however, in the same way with three series, D, E, F. To make the experimental test complete and afford some control of the method, I began with the random series where the theory is known and familiar.

(A.) *Random series.*

The distribution of correlations in this case should be symmetrical about zero, and, though it can hardly be normal, should approximate to the normal form with the mode at zero; the standard deviation should be $1/\sqrt{10}$, or 0·3162.* The results given by experiment are shown in Table VIII, which shows separately the distributions for

TABLE VIII.—Frequency-distributions of correlations for samples of 10 observations from random series.

Correlation.	Frequency.		
	Series A_0, B_0, C_0.	Series D_0, E_0, F_0.	Total.
− 0·9 — − 1·0	—	—	—
− 0·8 — − 0·9	1	—	1
− 0·7 — − 0·8	1	2	3
− 0·6 — − 0·7	4	8	12
− 0·5 — − 0·6	9	8	17
− 0·4 — − 0·5	18	13	31
− 0·3 — − 0·4	37	31	68
− 0·2 — − 0·3	30	37	67
− 0·1 — − 0·2	24	20	44
0 — − 0·1	32	33·5	65·5
0 — + 0·1	27	30·5	57·5
+ 0·1 — + 0·2	38	37	75
+ 0·2 — + 0·3	28	25	53
+ 0·3 — + 0·4	26	20	46
+ 0·4 — + 0·5	12	15	27
+ 0·5 — + 0·6	6	9	15
+ 0·6 — + 0·7	3	8	11
+ 0·7 — + 0·8	1	3	4
+ 0·8 — + 0·9	2	—	2
+ 0·9 — + 1·0	1	—	1
Total	300	300	600

* As we are sampling from material that is not merely uncorrelated but completely independent, the expression for the standard error of r reduces to its simplest form.

each set of three series; Fig. 16 gives a graph of the results for the two sets combined. It will be seen that the distributions are at least moderately symmetrical, though by no means as regular as might be wished. The means and standard deviations are as follows :—

A_0, B_0, C_0	...	$M = -0{\cdot}019.$	$\sigma = 0{\cdot}3191.$
D_0, E_0, F_0	...	$M = -0{\cdot}0075.$	$\sigma = 0{\cdot}3263.$
Together	...	$M = -0{\cdot}013.$	$\sigma = 0{\cdot}3227.$

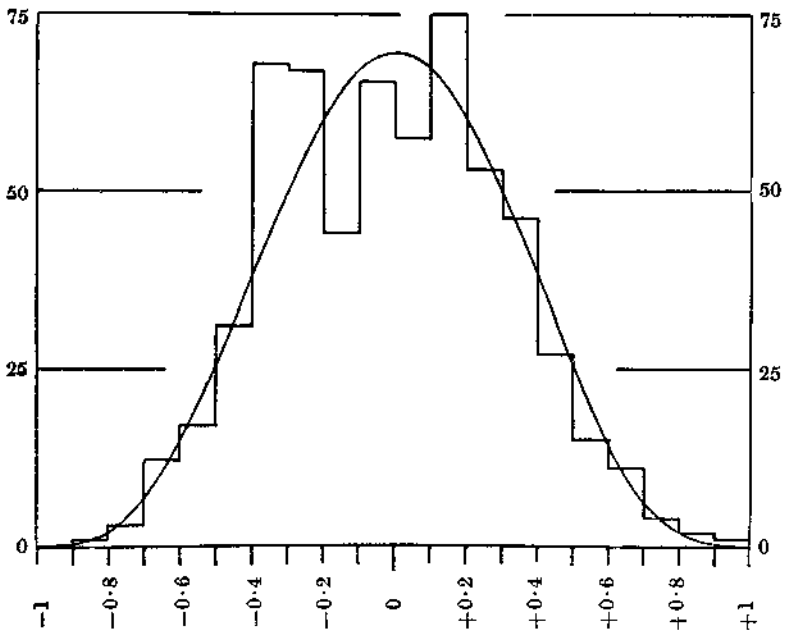

FIG. 16.—Frequency-distribution of 600 correlations between samples of 10 observations from random series (Table VIII).

The standard error of the mean with 300 observations is 0·0183, with 600 observations, 0·0129: the divergence from zero is just greater than the standard error in the first case, and less than half the standard error in the second; for the two sets together it is just equal to the standard error. The standard error of the standard deviation is 0·0129 for 300 observations, 0·0091 for 600 observations; all the divergences are well within the standard error. Mean and standard deviation agree very fairly with theory: it is only the irregularity of the distribution which is not pleasing.

To get some measure of the divergence in this respect, I calculated Professor Pearson's symmetrical limited range-curve with the theoretical value of the standard deviation :—

$$y = 69{\cdot}846\ (1 - x^2)^{3\cdot5}.$$

This is the curve of which the graph is shown in Fig. 16. The principal excess of frequency is in the interval $-0\cdot3$ to $-0\cdot4$, but there are also marked excesses at $-0\cdot2$ to $-0\cdot3$, and $+0\cdot1$ to $+0\cdot2$, compensated by deficiencies over the range $-0\cdot2$ to $+0\cdot1$. Grouping the frequencies below $-0\cdot7$ and above $+0\cdot7$, χ^2 comes to $29\cdot36$ and n' is 16, so that P is $0\cdot015$, a low though not impossible value. The odd thing is that the two separate distributions from A_0, B_0, C_0, and from D_0, E_0, F_0, agree in the sign of the most marked divergences, and this can hardly be anything but an unfortunate fluke. If the two distributions are treated as forming a two-row contingency table, with the same grouping χ^2 comes to $6\cdot40$ only and n' is 16, which gives $P = 0\cdot97$: the two distributions agree much too well with each other even in their irregularities.

The serial correlations for these random series A_0 to F_0 will be found in Table X below, p. 36. The number of observations on which they are based range from 100 down to 90, so that the standard errors range from $0\cdot1$ to $0\cdot105$. Whichever value we take, there are 47 of the correlations less than the standard error, 13 between once and twice the standard error, and none greater. Expectation, assuming normal distribution, would be 41 : 16 : 3.

(B.) *Series with random differences.*

The frequency-distributions of the correlations for samples of 10 observations from these series are shown in Table IX. It is evident that both the distributions, from A_1, B_1, C_1, and D_1, E_1, F_1, respectively, are much more widely dispersed than the correlations from samples of random series, and the set D_1, E_1, F_1, like the total, is clearly bimodal. This is what the argument of Section III led us to expect. But the two contributions from A_1, B_1, C_1, and from D_1, E_1, F_1, differ much too largely from each other to enable us to attach much weight to the pool of the two. To begin with, the second set is more widely dispersed than the first: the respective standard deviations are:—

A_1, B_1, C_1	0·500
D_1, E_1, F_1	0·601
Combined series	0 555

In the second place the set A_1, B_1, C_1 is not clearly bimodal, but merely irregular. At the same time the distribution, when I obtained it, seemed rather puzzling. The sub-contributions were rather suggestive of outlying modes, and it will be noticed that in the total of 300 observations the highest frequency is that for the interval $+0\cdot6$ to $+0\cdot7$. I felt some doubt whether the distribution was

really bimodal or merely flat-topped; it was this doubt, and the desire to clear it up, which originally led me to carry through the experiments with the second series D_1, E_1, F_1. The second series is quite clearly bimodal, with modes *circa* 0·7, and these modes remain marked when the results of the two sets are taken together. But as I have said, not much weight can be attached to this when the two components are so different.

TABLE IX.—Frequency-distributions of correlations for samples of 10 observations from series with random differences (conjunct series with random differences).

Correlation.	Frequency.			Series X_1.
	Series A_1, B_1, C_1.	Series D_1, E_1, F_1.	Total A_1 to F_1.	
− 0·9 — − 1·0	2	7	9	8
− 0·8 — − 0·9	11	17	28	21
− 0·7 — − 0·8	14	29	43	24
− 0·6 — − 0·7	18	19	37	34
− 0·5 — − 0·6	21	22	43	27
− 0·4 — − 0·5	17	13	30	38
− 0·3 — − 0·4	20	11	31	42
− 0·2 — − 0·3	12	14	26	41
− 0·1 — − 0·2	22	13	35	33
0 — − 0·1	21	8	29	31
0 — + 0·1	10	7	17	43
+ 0·1 — + 0·2	18	11	29	34
+ 0·2 — + 0·3	13	10	23	28
+ 0·3 — + 0·4	18	12	30	33
+ 0·4 — + 0·5	20	18	38	34
+ 0·5 — + 0·6	18	18	36	26
+ 0·6 — + 0·7	24	20	44	31
+ 0·7 — + 0·8	13	28	41	30
+ 0·8 — + 0·9	7	16	23	34
+ 0·9 — + 1·0	1	7	8	8
Total	300	300	600	600

I decided therefore that I must carry through for this case another series of experiments in which all the sets of observations should be taken independently. To keep the same frequency-distribution as before for the fundamental random series, counters (cardboard wads for No. 12 cartridges) were taken and a set of 20 was inscribed with the numbers from − 9 to 0 and 0 to + 9. Fifteen such sets, or 300 counters in all, were prepared and put in a bag: a counter was drawn at random, noted, put back, stirred up with the others, another drawn, and so on. Ten such drawings having been made, the addition of the numbers, step by step, gave the sum series for the correlation: another set of ten drawings gave its fellow-set, and the

correlation between them could then be worked out. Six hundred correlations were worked out in this way, and the frequency-distribution is shown in the last column of Table IX under the heading "Series X_1." A graph is given in Fig. 17. The mean and standard deviation are:—

$$M = +0{\cdot}0093 \qquad \sigma = 0{\cdot}513.$$

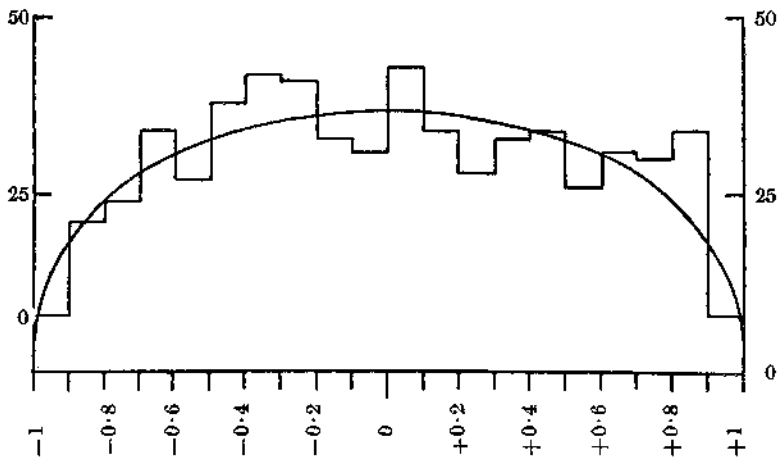

FIG. 17.—Frequency-distribution of 600 correlations between samples of 10 observations from conjunct series with random differences (Series X_1, Table IX).

The standard deviation lies accordingly between those for A_1, B_1, C_1, and D_1, E_1, F_1, but nearer the former. There is no indication, in the total, of bimodality. The graph shows Pearson's symmetrical curve of limited range with the same range and standard deviation,

$$y = 36{\cdot}717\ (1 - x^2)^{0{\cdot}40152},$$

which in this case gives a curve somewhat resembling a semi-ellipse. This gives too high a frequency at both ends, so I have some doubt whether it truly represents the actual form; but, even so, on testing the distribution by the χ^2 method, I find $\chi^2 = 21{\cdot}69$, $n' = 20$, which gives $P = 0{\cdot}30$, indicating quite a passable fit.

I think we must accept Series X_1 as giving the best evidence at present available as to the form of the distribution, and take it as unimodal. But I remain exceedingly puzzled. I booked up the correlations of Series X_1 by separate hundreds; the last hundred was the most widely dispersed (s.d. 0·553), and as clearly bimodal as D_1, E_1, F_1. The distribution seems to be in some way highly unstable and liable to break up into a distribution with relatively low central

frequencies and much higher frequencies round about 0·6 to 0·8. To endeavour to throw light on the reasons for this instability, I worked out the serial correlations for the fundamental random series A_0 to F_0 inclusive and give them in Table X. If, it occurred to me, owing to imperfect fulfilment of the conditions of simple sampling or otherwise, series D_0, E_0, and F_0 proved to be on the whole slightly conjunct series, we might have quite enough to account for the difference between the results given by D_1, E_1, F_1 and by A_1, B_1, C_1, having regard to the results of the next section. As we are dealing with samples of 10 observations only, we are not really concerned with r_{10}, the last correlation given in Table X; omitting this and looking at the others, it will be seen that there is a certain preponderance of negative correlations in A_0, B_0, C_0, and of positive correlations in D_0, E_0, F_0: there are, in fact, 16 negatives out of the 27 correlations in A_0, B_0, and C_0, 16 positive out of the 27 in D_0, E_0, and F_0. But the differences look hardly adequate to account for the divergence between the second and third columns of Table IX.*

TABLE X.—Serial correlations for the random series A_0 to F_0.

r.	A_0.	B_0.	C_0.	D_0.	E_0.	F_0.
1	− 0·130	− 0·089	+ 0·080	+ 0·007	+ 0·014	+ 0·071
2	− 0·075	− 0·005	+ 0·010	+ 0·133	+ 0·014	− 0·191
3	− 0·009	− 0·068	− 0·001	+ 0·094	+ 0·085	+ 0·010
4	− 0·167	− 0·147	− 0·059	− 0·098	− 0·028	+ 0·071
5	+ 0·116	+ 0·087	− 0·083	+ 0·035	+ 0·037	− 0·020
6	+ 0·047	− 0·141	− 0·043	− 0·027	− 0·127	+ 0·006
7	− 0·090	+ 0·141	+ 0·056	− 0·005	+ 0·040	− 0·016
8	+ 0·024	+ 0·184	+ 0·093	+ 0·073	+ 0·061	− 0·055
9	+ 0·037	− 0·015	− 0·035	+ 0·006	− 0·170	− 0·044
10	+ 0·128	− 0·020	− 0·026	− 0·059	− 0·047	− 0·117

(C.) *Series with correlated differences.*

The results of the experiments with these series are given in Table XI, and a graph of the frequency-distribution for the 600 observations from the two sets combined is shown in Fig. 18. In complete accordance with expectation, the distribution is U-shaped; a little over one-third of the correlations from the samples exceeding $\pm 0·9$ and about 58 per cent. exceeding $\pm 0·8$. The results from the first set, A_2, B_2, C_2, and the second set, D_2, E_2, F_2, are in good

* Treating these as a two-row contingency table, I make $\chi^2 = 38·57$, $n' = 20$, $P = 0·01$ roughly.

accordance with each other, but the second set shows slightly greater dispersion. The form of distribution in this case is indeed so marked that it is brought out quite clearly by a very short series of trials.

TABLE XI.—Frequency-distributions of correlations for samples of 10 observations from series with correlated differences (conjunct series with conjunct differences).

Correlation.	Frequency.		
	Series A_2, B_2, C_2.	Series D_2, E_2, F_2.	Total.
− 0·9 — − 1·0	51	61	112
− 0·8 — − 0·9	30	36	66
− 0·7 — − 0·8	20	17	37
− 0·6 — − 0·7	11	12	23
− 0·5 — − 0·6	11	9	20
− 0·4 — − 0·5	10	8	18
− 0·3 — − 0·4	7	4	11
− 0·2 — − 0·3	5	6	11
− 0·1 — − 0·2	4	1	5
0 — − 0·1	6	1	7
0 — + 0·1	8	2	10
+ 0·1 — + 0·2	7	3	10
+ 0·2 — + 0·3	4	1	5
+ 0·3 — + 0·4	2	10	12
+ 0·4 — + 0·5	5	4	9
+ 0·5 — + 0·6	6	6	12
+ 0·6 — + 0·7	12	14	26
+ 0·7 — + 0·8	20	15	35
+ 0·8 — + 0·9	32	31	63
+ 0·9 — + 1·0	49	59	108
Total	300	300	600

It is an interesting question, though of more theoretical than practical importance, whether the distribution is strictly U-shaped, with the frequency increasing indefinitely towards unity at either end of the range, or whether there is a true mode in the neighbourhood of unity. Table XII gives a detailed analysis of the distribution of correlations exceeding 0·9 at either end of the range. The figures are naturally irregular, but taking those for both positive and negative correlations together, a mode is suggested between 0·98 and 0·99.* The bimodality met with in some of the sub-series for series with random differences suggests that as the correlation between differences is gradually increased from zero, the distribution

* But even beyond 0·99 there is no rapid falling-off in frequency. Of the 24 coefficients numerically exceeding 0·99, 11 numerically exceed 0·995.

becomes bimodal, and the modes shift out to the extremities of the range; but this is rather a speculative deduction from the facts observed.

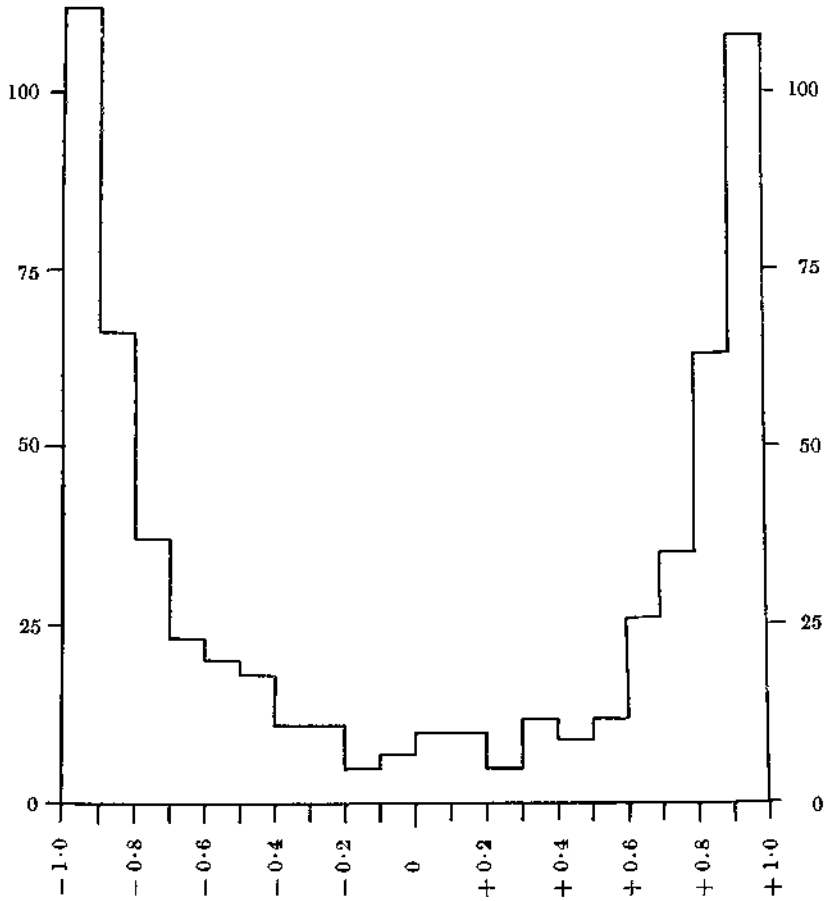

FIG. 18.—Frequency-distribution of 600 correlations between samples of 10 observations from conjunct series with conjunct differences (Table XI).

The experimental work has thus completely borne out the tentative conclusions reached in Section III, and the general result is clear. Considering only the two simple types of conjunct series, those with random differences and those with conjunct differences respectively, correlations between samples of the first type are subject to a much higher standard error than that given by the usual formula, but do not tend definitely to mislead; correlations between samples of the second type tend definitely to be "nonsense-

correlations"—correlations approaching plus or minus unity in value. The tentative answer to the problem of my title is therefore this: that some time-series are conjunct series with conjunct differences, and that when we take samples from two such series the distribution of correlations between them is U-shaped—we tend to get high positive or high negative correlations between the samples, without any regard to the true value of the correlation between the series that would be given by long experience over an indefinitely extended time.

TABLE XII.—Detailed analysis of the distribution of correlations exceeding $\pm$ 0·9 in Table XI.

Correlation.	Frequencies of correlations.		
	Positive.	Negative.	Together.
0·99 — 1·00	10	14	24
0·98 — 0·99	16	19	35
0·97 — 0·98	17	7	24
0·96 — 0·97	14	13	27
0·95 — 0·96	4	17	21
0·94 — 0·95	7	7	14
0·93 — 0·94	13	11	24
0·92 — 0·93	12	8	20
0·91 — 0·92	7	6	13
0·90 — 0·91	8	10	18
Total	108	112	220

Suppose we form a random series, for which the mean of the terms is zero, and regard each term as representing an observation during one unit of time, *e.g.*, an annual observation. Obviously this series is not correlated with the time. Form the first sum of the random series. This series will swing about above and below the zero base-line, but will not tend as the length of the series is increased to be correlated with the time. Now form the second sum of the random series, thus obtaining a conjunct series with conjunct differences. The swings above and below the base-line will now be smoother, longer and of greater amplitude, but still as the length of the series is increased there will be no tendency for it to be correlated with the time. Now I mentioned early in this Address the view that to interpret such "nonsense-correlations" as are here considered as implying causation is to "ignore the common influence of the time-factor," or as it has otherwise been put, the fact that both variables are correlated with the time. And I added that, while I could not accept the phrase, there was a special and

definite sense in which it might be said to cover the explanation. We see, in fact, that conjunct series with conjunct differences are *not* necessarily correlated with the time, so the phrase criticized is at least inexact. But, successive differences being correlated with each other, there is a tendency for the curve to rise or fall consistently over more or less prolonged periods ; there is a greater or less degree of *continuity* with time, and hence a tendency for the variable to be correlated with the time *over short samples*. This is, I think, the only sense in which the "common influence of the time factor" can be held to be responsible.

I give my answer to the problem as a tentative answer only, for I quite recognize that the discussion is inadequate and incomplete. The full discussion of the mathematical problem—given two series, each with specified serial correlations, required to determine the frequency-distribution of correlations between samples of n consecutive observations—I must leave to more competent hands. It is quite beyond my abilities, but I hope that some mathematician will take it up. The results that he may obtain may seem to be of mere theoretical importance, for in general we only have the sample itself, which may be quite inadequate for obtaining the serial correlations. But to take such a view would, I think, be short-sighted. The work may not lead, it is unlikely to lead, to any succinct standard error, or even frequency-distribution applicable to the particular case. But only such direct attack can, it seems to me, clear up the general problem ; show us what cases are particularly liable to lead to fallacious conclusions, and in what cases we must expect a dispersion of the sample-correlations greater than the normal. I have only considered two cases, and there is more variety in fact than this—compound curves of every sort* may occur. If my view is correct, that the serial correlations of the difference series are the really important factor, even the special solution for the special problem may not be so hopeless as at first sight it may seem ; for the sample may be a more adequate basis for the approximate determination of the difference correlations than for the determination of the serial correlations of the series itself.

* The mortality curve of Fig. 1 does not suggest a conjunct series with conjunct differences, but rather a segment of a series that might be regarded as compound—a conjunct series with an oscillatory series superposed like the Beveridge series of the next section. It may be noted that when we separate out the oscillations in such a series by taking the difference of u_s from the mean of the terms u_{s-r} to u_{s+r}, we are in fact splitting up the series into (1) an oscillatory series, (2) a conjunct series.

In a mathematical series any term u_s is some definite mathematical function of s, and has precise and definite mathematical relations to the terms that precede and the terms that follow. In a statistical series u_s is no longer a definite mathematical function of s, and no longer has precise and definite relations to the terms that precede and follow it. I have suggested replacing, as we usually have to do in statistics, the conception of mathematical functionality by the conception of correlation, and thus specifying the characteristics of the series by its serial correlations. Apart from its application to the theory of sampling in time-series, such a specification is of interest in itself as a method of analysis. I give an illustration or two in the next section.

SECTION V.—*Serial correlations for Sir William Beveridge's index-numbers of wheat prices in Western Europe; and for rainfall at Greenwich.*

The great majority of statistical series that we possess seem to me to be far too short to afford any adequate basis for determining the serial correlations; few of them extend even for as long as a century. And brevity of the sample has more than one disadvantage. That it may not be adequately representative is the primary fault. But, further, it must be remembered that in determining r_1 from a series of n terms we use u_1 to u_{n-1} for the one series, u_2 to u_n for the other; in determining r_2 we use u_1 to u_{n-2} for the one series, u_3 to u_n for the other, and so on. Each successive correlation in the series is determined from different observations, and if k is not small compared with n, the number of terms in the given data, r_k may be seriously inconsistent with r_1. Moreover, the equation that we use for determining the difference correlations from the serial correlations (the ρ's from the r's) assumes that the "end-effects" are negligible. Bearing these considerations in mind, it seemed to me that Sir William Beveridge's index-numbers for wheat prices in Western Europe,* a series extending over more than 300 years, was about the only one worth detailed study. Following his practice in the periodogram analysis,† I have used only the 300 years 1545 to 1844 inclusive, but it must be understood that I have worked on the index-numbers themselves, not the derived figures obtained by taking the ratio of each index-number to the average of the 31 of which it forms the centre, which were used for periodogram analysis.

* *Economic Journal*, vol. xxxi, p. 429, December, 1921.

† *J.S.S.*, vol. lxxxv, p. 412, 1922.

The work was executed as follows, without any grouping of observations. The squares of all the index-numbers were first added on a Burroughs Adding Machine, and the slip of squares retained. The machine was then "split," and the numbers themselves entered in duplicate on the right and left halves of the machine and added. The resulting slip was then cut longitudinally down the centre; by putting these two half-slips against each other so that observation s of the first is opposite observation $s + k$ of the second, corresponding observations could be added in one's head and the squares entered direct on the machine. Let us call an observation on the first slip X and on the second slip Y. Then the slip thus obtained gives $S(X + Y)^2$. $S(X^2)$ will be obtained from the slip of squares by deducting k squares from the bottom, and $S(Y^2)$ by deducting k squares from the top, and

$$S(X + Y)^2 - S(X^2) - S(Y^2) = 2S(XY).$$

The means M_x and M_y are obtained from the addition-slip for the observations themselves by deducting k observations from the bottom and the top respectively, and the reductions of the mean product and the standard deviations to the mean are effected in the usual way. The slips giving $S(X + Y)^2$ were, of course, read over and checked, and I hope the results are accurate. A serious blunder can hardly escape the mere graphic check of plotting the results—one error was so found. Another partial check is given by the fact that, since the index-numbers are whole numbers, $S(XY)$ must be a whole number, or $2S(XY)$ must be even. This check led to the discovery of four minor errors that had escaped detection in the first reading over; one of these only affected the correlation coefficient in the sixth place of decimals, one in the fifth place and two in the fourth place. I had originally intended only carrying the calculations up to r_{30}, one-tenth of the whole number of observations, thinking this might be as far as it was safe to go, but some curiosity as to whether there would be any apparent effect of the Brückner cycle of 35 years led me to continue up to r_{40}. The correlations are given in Table XIII, and a graph is shown in Fig. 19. The correlations are all positive, as they evidently must be in a series that sweeps up from values round about 20 or 30 in the earlier years to 100, 200 and over in the later years. They fall away at first with some rapidity to a minimum of 0·71 at r_8; there is then a large broad hummock in the curve followed by some minor oscillations, and finally, from about r_{25} onwards, the curve tails away comparatively smoothly to 0·53 at r_{40}. There is no trace of any special maximum suggesting the Brückner cycle.

TABLE XIII.—Serial correlations for Sir William Beveridge's index-numbers for wheat prices in Western Europe, 1545–1844. All correlations are positive.

k.	r_k.	k.	r_k.
1	0·92240	21	0·63432
2	0·83353	22	0·62901
3	0·79639	23	0·61136
4	0·79560	24	0·59658
5	0·79146	25	0·59193
6	0·76013	26	0·60030
7	0·72850	27	0·61241
8	0·71063	28	0·60680
9	0·72170	29	0·60770
10	0·75356	30	0·60789
11	0·78013	31	0·60877
12	0·77661	32	0·59589
13	0·74508	33	0·58851
14	0·73330	34	0·58553
15	0·73625	35	0·57505
16	0·73609	36	0·56441
17	0·70015	37	0·55683
18	0·65054	38	0·55342
19	0·62692	39	0·54495
20	0·62319	40	0·53479

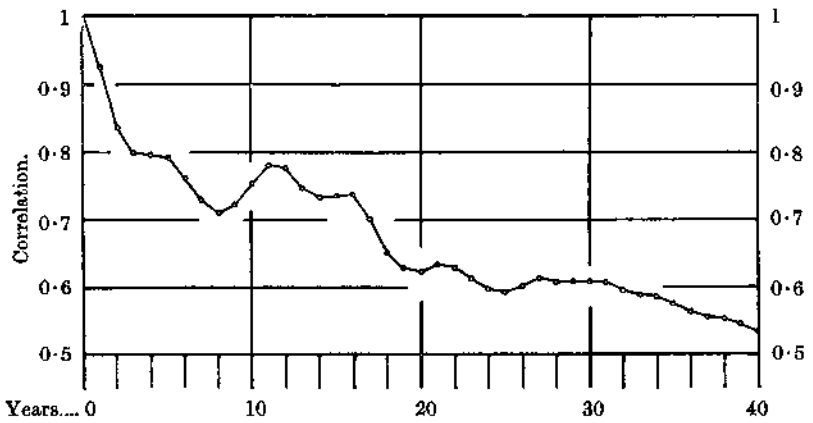

FIG. 19.—Serial correlations up to r_{40} for Sir William Beveridge's index-numbers of wheat prices in Western Europe, 1545–1844 (Table XIII).

Clearly the series of index-numbers is a conjunct series with oscillatory differences, that is probably marked oscillatory components. The next step is to obtain from the correlations of Table XIII the serial correlations for the differences. These are shown, for a

few selected values of the interval h, in Table XIV. Consider first the column for $h=1$, giving the limit values (equation (4), Appendix II) for the serial correlations of the differences between consecutive years, and compare with the graph at the top of Fig. 20. It will be seen from the graph that up to about ρ_{26} the correlations are fairly regularly oscillatory, but after this they become more irregular and the

TABLE XIV.—Index-numbers of wheat prices in Western Europe: limit values of the serial difference correlations, derived from the correlations of Table XIII by equation (5), Appendix II, for various values of the interval h.

k.	$h = 1$.	$h = 5$.	$h = 6$.	$h = 11$.	$h = 15$.
	$_1\rho_k$.	$_5\rho_k$.	$_6\rho_k$.	$_{11}\rho_k$.	$_{15}\rho_k$.
1	+ 0·073	+ 0·693	+ 0·677	+ 0·715	+ 0·712
2	− 0·333	+ 0·341	+ 0·335	+ 0·455	+ 0·421
3	− 0·234	+ 0·117	+ 0·156	+ 0·338	+ 0·314
4	+ 0·022	− 0·127	+ 0·009	+ 0·288	+ 0·349
5	+ 0·175	− 0·409	− 0·249	+ 0·197	+ 0·391
6	+ 0·002	− 0·437	− 0·534	+ 0·065	+ 0·311
7	− 0·089	− 0·367	− 0·439	+ 0·025	+ 0·222
8	− 0·186	− 0·288	− 0·303	− 0·005	+ 0·154
9	− 0·134	− 0·205	− 0·186	− 0·030	+ 0·164
10	+ 0·034	− 0·049	− 0·051	− 0·113	+ 0·235
11	+ 0·194	+ 0·154	+ 0·143	− 0·156	+ 0·312
12	+ 0·180	+ 0·299	+ 0·297	+ 0·004	+ 0·274
13	− 0·127	+ 0·309	+ 0·281	+ 0·137	+ 0·094
14	− 0·095	+ 0·283	+ 0·277	+ 0·178	− 0·120
15	+ 0·020	+ 0·230	+ 0·243	+ 0·193	− 0·257
16	+ 0·231	+ 0·138	+ 0·187	+ 0·155	− 0·112
17	+ 0·088	− 0·013	+ 0·018	+ 0·076	− 0·055
18	− 0·167	− 0·133	− 0·150	− 0·080	− 0·159
19	− 0·128	− 0·182	− 0·173	− 0·147	− 0·241
20	− 0·096	− 0·196	− 0·182	− 0·191	− 0·228
21	+ 0·106	− 0·162	− 0·167	− 0·184	− 0·106
22	+ 0·080	− 0·131	− 0·177	− 0·252	− 0·052
23	− 0·018	− 0·083	− 0·177	− 0·317	− 0·078
24	− 0·065	− 0·099	− 0·136	− 0·289	− 0·139
25	− 0·084	− 0·113	− 0·108	− 0·259	− 0·198
26	− 0·024	− 0·102	− 0·039	− 0·210	—
27	+ 0·114	− 0·000	+ 0·004	− 0·147	—
28	− 0·042	+ 0·033	− 0·002	− 0·072	—
29	+ 0·005	+ 0·080	+ 0·060	+ 0·068	—
30	− 0·004	+ 0·117	+ 0·092	—	—
31	+ 0·089	+ 0·127	+ 0·143	—	—
32	− 0·035	+ 0·054	+ 0·079	—	—
33	− 0·028	+ 0·040	+ 0·041	—	—
34	+ 0·048	+ 0·044	+ 0·061	—	—
35	+ 0·001	+ 0·018	—	—	—
36	− 0·020	—	—	—	—
37	− 0·027	—	—	—	—
38	+ 0·033	—	—	—	—
39	+ 0·011	—	—	—	—

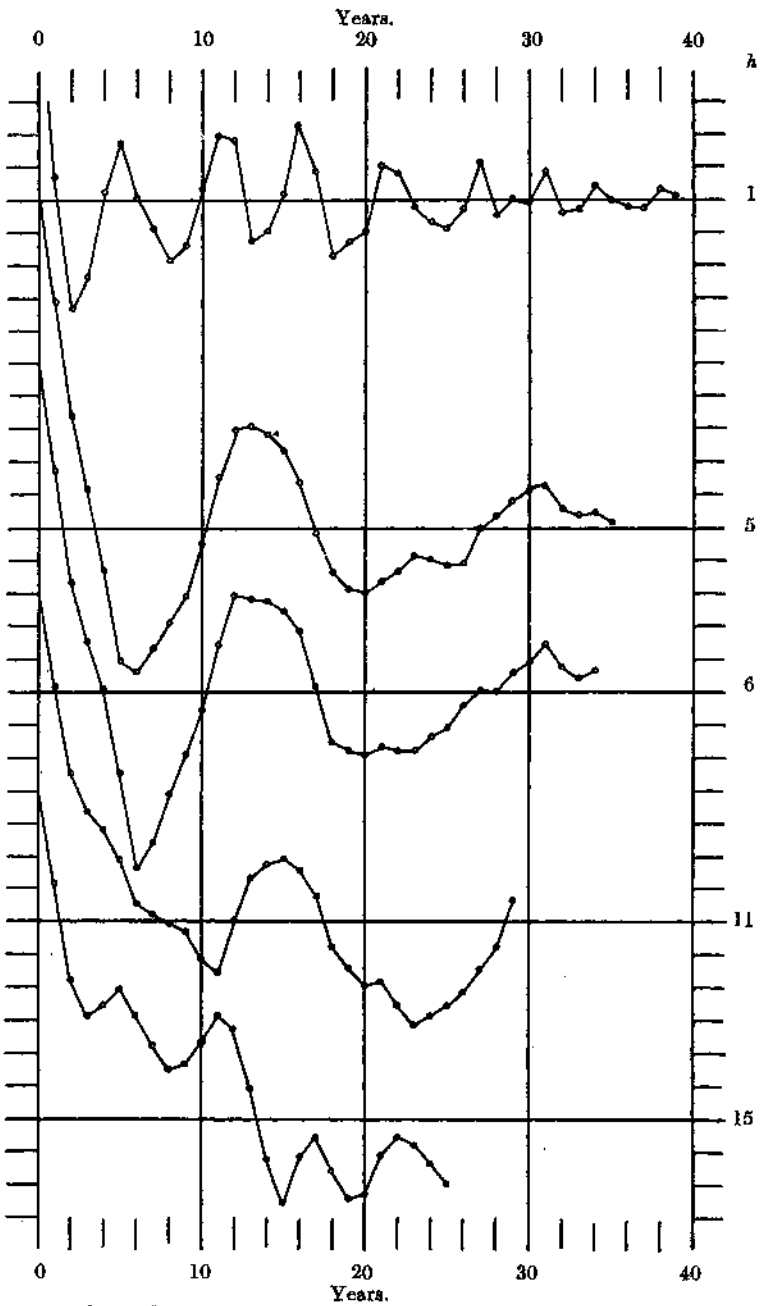

FIG. 20.—Serial difference correlations for the index-numbers of wheat prices in Western Europe: intervals of differencing h = 1, 5, 6, 11 and 15 years respectively (Table XIV).

oscillations diminish in amplitude. At a rough judgment by eye, I should put the successive maxima and minima at about 2·2, 5, 8, 11·5, 13·7, 16·2, 18·7, 21·5 and 22·3; the intervals for the half-oscillations would be on this reckoning 2·8, 3, 3·5, 2·2, 2·5, 2·5, 2·8, and 3 years, giving an average duration for the oscillation of about 5·6 years. Beveridge finds four marked periods* between 5·100 and 5·960 years inclusive, of intensities (amplitudes squared) in round numbers 42, 23, 33 and 23. Of shorter periods he accepts only three as definite—one of period 2·735 and intensity roundly 8, one of length 3·415 and intensity roundly 16, and the third of length 4·415 and intensity roundly 16. Differencing tends, of course, to emphasize the shorter periods,† but it is a little surprising to find the effect of the terms of about 5 to 6 years' duration standing out, apparently, almost by themselves. After the first minimum only one correlation exceeds 0·2, viz., ${}_1\rho_{16}$, and the regularity of the oscillations is greater than one might have expected.

To bring out the oscillations of longer duration, if there are any, it is only necessary to work out the difference-correlations for an interval longer than one year. Judging from the first curve, the predominant oscillations of shorter duration were of 5 to 6 years' duration. These oscillations would be practically eliminated by taking 5 years or 6 years for the interval of differencing h, and so the serial correlations for these values of h were worked out next.‡ The figures are given in Table XIV, and the graphs are the second and third in Fig. 20. The two curves are very like each other; both give a good clean sweep and the correlations are considerably higher than in the last case. The similarity between the two extends to points of detail. The drop to the first minimum is abrupt, the respective minima being $-0·437$ and $-0·534$, both at 6 years. Thence there is a sharp rise to a maximum in the neighbourhood of year 13, with correlations of about 0·3, this maximum being flat and the slope up to the maximum steeper than the slope away from it. The third half-oscillation, below the base-line, is double-humped, clearly in the first case, less markedly in the second. We seem to have here as the predominant factors Sir William Beveridge's periods.

* I refer specifically to his table of "Apparent Periods" on pp. 444–45.

† As regards the effect of differencing on the amplitudes of harmonic terms, *cf.* my paper on the "Time Correlation Problem," *J.S.S.*, vol. lxxxiv, 1921, especially Table I, p. 507.

‡ ${}_h\rho_k$ is the correlation between $u_{s+h} - u_s$ and $u_{s+h+k} - u_{s+k}$; hence ${}_h\rho_h$, for example, is the correlation between consecutive differences taken with interval h.

Length.					Intensity.
12·840	...	...	...	...	46·00 +
15·225	...	...	...	...	76·16 +
17·400	...	...	...	...	54·12 +

Scaling off on my original chart the lengths from the zero-point to the points at which the curves cut the base-line for the first, second and third times, the durations suggested for the oscillations by the first quarter-period and the following half-periods would be 13·6, 13·6 and 13·4 years from the first curve, and 16, 12·6, 13·4 years from the second curve. The first minimum in the following half-oscillation is at 20 years, nearly, in both curves, suggesting a duration for the oscillation of about 14·3 years. The second minimum I should put at about 25 years in the first curve and 23 years in the second, suggesting durations of about 16·7 and 15·3.

For my next case, again with the intention of eliminating as far as possible the oscillations shown in the curve for $h = 1$, I took $h = 11$. The form of the curve now obtained, with so long an interval of differencing, suggests that we have, superposed on the effect of the oscillations, some effect of either very long oscillations or secular movement. Judging therefore rather by the maxima and minima, the first maximum suggests oscillations of a duration near 15 years: the second minimum, placing it at 23 years, would suggest a duration of 15·3.

For my final case I took $h = 15$, about as far as it seemed worth while to go with only 40 serial correlations from which to construct the difference-correlations. The oscillations which are predominant when $h = 1$ are again, rather unexpectedly, conspicuous in this curve, the lowest on Fig. 20. On the other hand, the oscillations predominant in the second and third curves are more or less eliminated. The short oscillations are now too troublesome and the whole extent of our curve is too short to judge durations with any precision. If we may take the time to the point at which the curve first cuts the base as a half-duration—and it falls just about half-way between a maximum and a minimum of the minor oscillations, so that its position would not be greatly disturbed by them—this is roughly 13·5 years, suggesting an oscillation of duration 54 years, which is one of the periods noted by Sir William Beveridge, with an intensity 26.

The work may suffice to suggest the interesting way in which the serial correlations can be used to bring out, at least by a first rough analysis, the predominant characteristics of a given series. In the series in question there can be no doubt about the differences being oscillatory. I had some hopes that by making h sufficiently large

one would practically eliminate the effect of oscillations, but even with $h = 15$ the correlations are still conspicuously oscillatory. As emphasized, the mere fact that a series is oscillatory, as defined, is no evidence that it is periodic: but if it is periodic it must be oscillatory. In so far, then, the results are in accordance with Sir William Beveridge's periodogram analysis, and its indication that a considerable part of the price movement is periodic. Is there anything, on the other hand, which suggests that the movement is oscillatory rather than truly periodic? At first, inspection of the curves of Fig. 20 made me suspicious. The oscillations in the values of the correlations tend notably to decrease in amplitude as k is increased. This comes out clearly in the curves for $h = 1$, 5 and 6, and it is exactly the sort of effect that may be obtained with a series which is oscillatory but not periodic. Further consideration showed me, however, that exactly the same effect will be given by the interference of different incommensurable periodicities. It is shown in Appendix I that if we have a function of the time expanded in a Fourier Series, so that

$$y = \mathrm{S}\left\{ \mathrm{A}_m \sin 2\pi \frac{t + \phi_m}{m\mathrm{T}} \right\},$$

where $m = 1, \frac{1}{2}, \frac{1}{3} \ldots$, the correlation between two ordinates of such a series at a time τ apart is given by

$$r = \frac{1}{\mathrm{S}(\mathrm{A}_m{}^2)} \mathrm{S}\left\{ \mathrm{A}_m{}^2 \cos 2\pi \frac{\tau}{m\mathrm{T}} \right\}.$$

In the present instance we have not got a Fourier Series with its simple periodicities in the proportions $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4} \ldots$, but a whole collection of incommensurable periodicities. The serial correlations can now take no simple form when we sum over a finite time. But I think it will be true that if we sum over a very long time the inconvenient product-terms which occur in the summation will tend to become very small and that as a first approximation we may take a similar expression for r:—

$$r = \frac{1}{\mathrm{S}(\mathrm{A}_m{}^2)} \mathrm{S}\left\{ \mathrm{A}_m{}^2 \cos 2\pi \frac{\tau}{\mathrm{T}_m} \right\}:$$

It seemed to me that it might be illuminating to calculate such a curve for r and compare it with the figures obtained from the data. I chose the case $h = 5$ of Table XIV. I took out the periods (Table XV) which seemed likely to have any appreciable effect, from Sir William Beveridge's table on pp. 444–45 of his paper, together with their intensities. But here occurred the first difficulty —the intensities given are not, in the majority of cases, the true intensities, but those of neighbouring trial periods. There seemed

nothing for it but to take them as the true intensities. The next step was to calculate the correcting factors for these intensities, to allow for the fact that we have taken differences with an interval of 5 years; these are given in column 3 of Table XV. The intensities are then multiplied by these factors (column 4) and finally the products are divided by the sum of the intensities, so as to make the total sum unity (column 5). It will be noted that the effect of differencing on the original intensities is to make the predominant periodicities 15·225, 12·840, 17·400, 11·000, and 9·750, in the order given, the last three having almost equal intensities and the periodicities shorter than 12·84 having more importance than I had estimated from Fig. 20.

TABLE XV.—Calculation of the curve of Fig. 21 from certain of Sir William Beveridge's periods for wheat prices.

	(1) Period years.	(2) Intensity.	(3) Factor.	(4) Factor × Intensity.	(5) Divided by sum of intensities.
1	7·417	21·72	2·919	63·40	0·057
2	8·050	23·23 +	3·448	80·10	0·072
3	9·750	33·89	3·879	131·46	0·119
4	11·000	33·84	3·919	132·62	0·120
5	12·050	23·30 +	3·721	86·70	0·078
6	12·840	46·00 +	3·536	162·66	0·147
7	15·225	76·16 +	2·946	224·37	0·203
8	17·400	54·12 +	2·465	133·41	0·120
9	19·900	37·88 +	2·016	76·37	0·069
10	35·5	23·29 +	0·735	17·12	0·015
Sum	—	—	—	1108·21	1·000

The compound cosine-curve with these intensities as amplitudes was now calculated, and the results are shown in Table XVI against the observed difference correlations; the graph is shown in Fig. 21. It will be seen that there is a broad, though only a broad, agreement with the data. There are only three discrepancies in sign, and the dying away of the oscillations is just as conspicuous in the calculated curve as in the data. The second "dip" is markedly double-humped, but the second minimum is markedly later than in the data and is deeper than the first. The agreement is, perhaps, as good as we have any right to expect, having used an approximate expression in the first place and approximate intensities in the second, and having ignored not only many other periodicities actually found in the data within the given range, but also all others outside it and all non-periodic components of the series.

TABLE XVI.—The ordinates of the compound harmonic curve derived from Table XV, compared with the observed coefficients in the column for $h = 5$ of Table XIV (*cf.* Fig. 21).

	Observed coefficient.	Calculated curve.		Observed coefficient.	Calculated curve.
0	+ 1·000	+ 1·000	18	− 0·133	− 0·007
1	+ 0·693	+ 0·862	19	− 0·182	− 0·063
2	+ 0·341	+ 0·496	20	− 0·196	− 0·062
3	+ 0·117	+ 0·033	21	− 0·162	− 0·028
4	− 0·127	− 0·360	22	− 0·131	+ 0·002
5	− 0·409	− 0·608	23	− 0·083	− 0·008
6	− 0·437	− 0·648	24	− 0·099	− 0·068
7	− 0·367	− 0·539	25	− 0·113	− 0·143
8	− 0·288	− 0·357	26	− 0·102	− 0·192
9	− 0·205	− 0·180	27	− 0·000	− 0·179
10	− 0·049	− 0·037	28	+ 0·033	− 0·102
11	+ 0·154	+ 0·066	29	+ 0·080	+ 0·004
12	+ 0·299	+ 0·153	30	+ 0·117	+ 0·095
13	+ 0·309	+ 0·222	31	+ 0·127	+ 0·132
14	+ 0·283	+ 0·261	32	+ 0·054	+ 0·113
15	+ 0·230	+ 0·252	33	+ 0·040	+ 0·064
16	+ 0·138	+ 0·190	34	+ 0·044	+ 0·023
17	− 0·013	+ 0·089	35	+ 0·018	+ 0·023

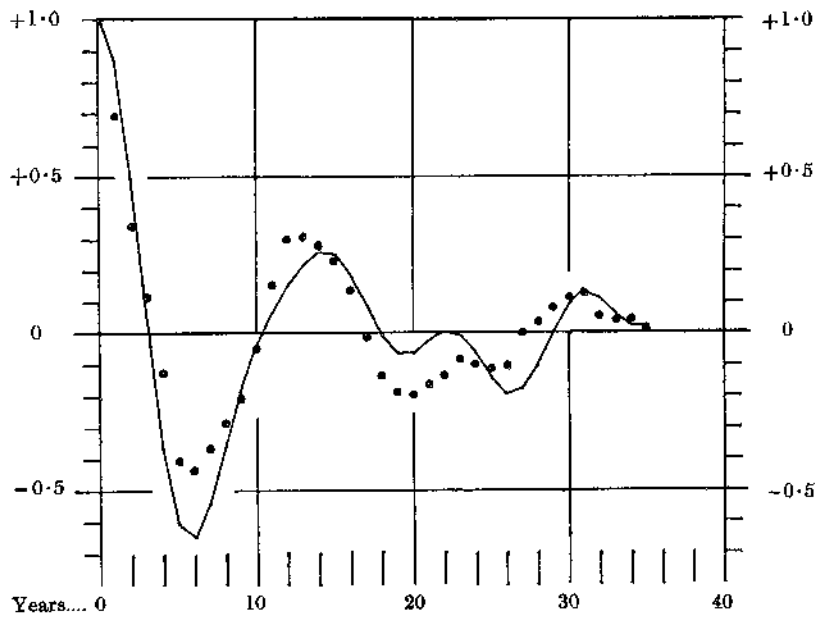

FIG. 21.—Serial difference correlations for $h = 5$ (Table XIV) (dots) and a curve constructed from certain of the periodicities given by Sir William Beveridge (line) (Table XVI).

Greenwich Rainfall.—The only other series which I have submitted to detailed analysis by this method is the rainfall at Greenwich over the 110 years 1815–1924. Records for a Continental station would have been better for comparison with the Beveridge Series, but this was the longest unbroken series that I could obtain.

Mere inspection of a graph suggests that the series is totally different in character from the last, and this impression is confirmed by the serial correlations given in Table XVII. All the correlations lie practically within the limits $\pm 0{\cdot}2$, only one (r_{18}) just exceeding this value. Since the standard errors are all of the order $0{\cdot}1$, this would suggest that none of the correlations are significant, and that the series is practically random. But looking at the graph, Fig. 22, there are some slight suggestions of order The correlations rise continuously over three years to the conspicuous maximum at r_7. At first I had only calculated the serial correlations up to r_{10}, but this led me to continue the work up to r_{20} to see if there was a corresponding maximum at r_{14}. There is. And having got this, I was enticed to continue up to r_{24} to see if there was a maximum again at r_{21}; it is a poor thing, but still a

TABLE XVII.—Serial correlations for Greenwich Rainfall, 1815–1924, and difference correlations for $h = 3$ and $h = 9$.

k.	r_k.	$_3\rho_k$.	$_9\rho_k$.
1	− 0·0036	+ 0·093	+ 0·061
2	− 0·0594	− 0·011	− 0·093
3	+ 0·0459	− *0·466*	+ 0·025
4	− 0·1248	− 0·099	− 0·114
5	− 0·0944	− 0·031	− 0·078
6	− 0·0182	− 0·014	− 0·033
7	+ 0·1858	+ 0·295	+ 0·280
8	− 0·0706	+ 0·032	− 0·072
9	− 0·0556	− 0·078	− *0·431*
10	− 0·0658	− 0·211	− 0·036
11	− 0·1086	− 0·130	− 0·073
12	+ 0·0562	+ 0·095	+ 0·031
13	+ 0·0857	+ 0·208	+ 0·193
14	+ 0·1010	+ 0·155	+ 0·176
15	− 0·0133	+ 0·062	− 0·073
16	− 0·1597	− 0·185	—
17	+ 0·0149	− 0·035	—
18	− 0·2008	− 0·204	—
19	− 0·0521	+ 0·087	—
20	− 0·0036	+ 0·028	—
21	+ 0·0002	+ 0·029	—
22	− 0·1103	—	—
23	− 0·0756	—	—
24	+ 0·1462	—	—

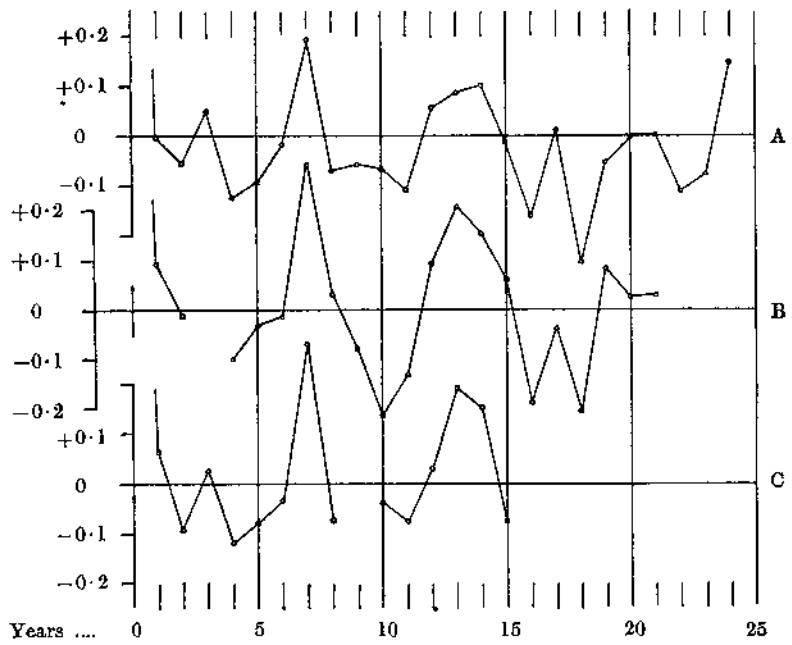

FIG. 22.—Serial correlations up to r_{24} for rainfall at Greenwich (Curve A): and difference correlations for $h = 3$ (Curve B), and for $h = 9$ (Curve C) (Table XVII).

maximum. This suggests a period—or an oscillation—of a duration close to 7 years. But there seems to be some interference by a shorter period, judging by the first maximum of about 3 years' duration. Taking difference correlations with $h = 3$ to eliminate this, the first thing that will be noted in the column for ${}_3\rho_k$ in Table XVII is the high negative correlation, $-0{\cdot}466$, for $k = 3$. This does not indicate any oscillation; it merely shows that the original series is nearly random, for if that series were purely random the correlation between consecutive first differences would be $-0{\cdot}5$.* I have therefore omitted the corresponding point from the graph, the second curve in Fig. 22. There is now a very marked maximum at 7, the second and third maxima lying at or just beyond 13 and 19. Taking the difference correlations for $h = 9$, and omitting for the same reason as before the point corresponding to ρ_9, the graph is very similar, with the first conspicuous maximum at 7 and the

* In the difference correlations of Table XIV, ${}_6\rho_6$, ${}_{11}\rho_{11}$ and ${}_{15}\rho_{15}$ are minima, but the values seem to run well with the others, and do not suggest any large random component.

second about 13·5. There is little to suggest oscillations of long duration, which should have been brought out by such an interval of differencing. But a correlation of nearly 0·3 based on over 100 observations would seem to be probably significant, and I think there must be an oscillation present of about 7 years' duration, as the most conspicuous component, with possible oscillations of shorter duration in the neighbourhood of 3 years. But it remains true that something like 90 per cent. of the entire variance of the series appears to be random.

This concludes my work. Starting from a question that may have seemed to some silly and unnecessary, we were led to investigate the correlations between samples of two simple mathematical functions of the time. It appeared that small samples (in time) of such functions tended to give us correlations departing as far as possible from the truth, the correlations tending to approach ± 1 if the time for which we had experience was very small compared with the time necessary to give the true correlation. Asking ourselves, then, what types of statistical series might be expected to give results analogous to those given by the mathematical function considered, we were led to a classification of series by their serial correlations $r_1, r_2, r_3, \ldots r_k, r_k$ being the correlation between terms s and $s + k$. The important matter in classification was the *form* of function relating r_k to k, which indicated the nature of the serial correlations between *differences* of the time-series. If this function is linear, the time-series has random differences; if it gives a graph concave downwards the difference correlations are positive. We concluded that it was series of the latter type (positively correlated series with positively correlated differences, or conjunct series with conjunct differences to use my suggested term) that formed the dangerous class of series, correlations between short samples tending towards unity. Experimental investigation completely confirmed this suggestion. Samples from conjunct series with random differences gave a widely dispersed distribution of correlations; samples from conjunct series with conjunct differences gave a completely U-shaped distribution, with over one-third of the correlations exceeding $\pm 0{\cdot}9$. In the last section the method of analysis by serial correlations was illustrated by a couple of examples.

APPENDIX I.—*The correlation between segments of two sine-curves of the same period, etc.*

Two variables, y_1 and y_2, are harmonic functions of the time t of the same period; say

$$\left.\begin{aligned} y_1 &= \sin\ 2\pi \frac{t}{\mathrm{T}} \\ y_2 &= \sin 2\pi \frac{t+\tau}{\mathrm{T}} \end{aligned}\right\} \tag{1}$$

where T is the period and τ the difference of phase; the amplitude is taken as unity since it does not affect the present question. It is required to find the correlation between simultaneous values of y_1 and y_2 over an interval $\pm h$ round the time u, treating the observed values as continuous. Taking the second variable, as giving the general result from which that for y_1 may be deduced by putting $\tau = 0$, we have

$$\int_{u-h}^{u+h} \sin 2\pi \frac{t+\tau}{\mathrm{T}}\, dt = \frac{\mathrm{T}}{2\pi}\left\{\cos\ 2\pi \frac{u+\tau-h}{\mathrm{T}} - \cos 2\pi \frac{u+\tau+h}{\mathrm{T}}\right\}$$
$$= \frac{\mathrm{T}}{\pi}\ \sin\ 2\pi \frac{u+\tau}{\mathrm{T}} \sin 2\pi \frac{h}{\mathrm{T}},$$

or, dividing by $2h$, we have for the mean of y_2 over the interval,

$$\mathrm{M}_2 = \frac{\mathrm{T}}{2\pi h} \sin 2\pi \frac{u+\tau}{\mathrm{T}} \sin 2\pi \frac{h}{\mathrm{T}}. \tag{2}$$

Further,

$$\int_{u-h}^{u+h} \sin^2 2\pi \frac{t+\tau}{\mathrm{T}}\, dt = h - \frac{\mathrm{T}}{4\pi} \cos\ 4\pi \frac{u+\tau}{\mathrm{T}} \sin\ 4\pi \frac{h}{\mathrm{T}},$$

or, dividing by $2h$,

$$\Sigma_2^2 = \sigma_2^2 + \mathrm{M}_2^2 = \tfrac{1}{2} - \frac{\mathrm{T}}{8\pi h} \cos\ 4\pi \frac{u+\tau}{\mathrm{T}} \sin\ 4\pi \frac{h}{\mathrm{T}} \tag{3}$$

where σ_2 is the standard deviation of y_2 over the interval.

To find the mean product p' of y_1 and y_2 over the interval, we require

$$\int_{u-h}^{u+h} \sin\ 2\pi \frac{t}{\mathrm{T}} \sin\ 2\pi \frac{t+\tau}{\mathrm{T}}\, dt$$
$$= h \cos\ 2\pi \frac{\tau}{\mathrm{T}} - \frac{\mathrm{T}}{4\pi} \cos\ 2\pi \frac{2u+\tau}{\mathrm{T}} \sin\ 4\pi \frac{h}{\mathrm{T}},$$

or

$$p' = \tfrac{1}{2} \cos\ 2\pi \frac{\tau}{\mathrm{T}} - \frac{\mathrm{T}}{8\pi h} \cos\ 2\pi \frac{2u+\tau}{\mathrm{T}} \sin\ 4\pi \frac{h}{\mathrm{T}}. \tag{4}$$

Equations (2), (3) and (4) suffice for the arithmetical solution of the problem. It does not seem possible to simplify the equations sufficiently to obtain any manageable expression for the correlation-coefficient r as a function of u, h and τ. But from (2) we can calculate, in any assigned case, the means M_1 and M_2; (3) will then give the standard deviations, and (4) will give the mean product of deviations p by subtracting the product M_1M_2.

The most interesting case, dealt with at length in Section II of the paper, is given by taking $\tau/T = \frac{1}{4}$, when the correlation between y_1 and y_2 over a whole period is obviously zero. For this case

$$\left.\begin{aligned} M_1 &= \frac{T}{2\pi h} \sin 2\pi \frac{u}{T} \sin 2\pi \frac{h}{T} \\ M_2 &= \frac{T}{2\pi h} \cos 2\pi \frac{u}{T} \sin 2\pi \frac{h}{T} \\ \Sigma_1^2 &= \tfrac{1}{2} - \frac{T}{8\pi h} \cos 4\pi \frac{u}{T} \sin 4\pi \frac{h}{T} \\ \Sigma_2^2 &= \tfrac{1}{2} + \frac{T}{8\pi h} \cos 4\pi \frac{u}{T} \sin 4\pi \frac{h}{T} \\ p' &= \frac{T}{8\pi h} \sin 4\pi \frac{u}{T} \sin 4\pi \frac{h}{T} \end{aligned}\right\} \qquad (5)$$

From these equations the curves showing r as a function of u were drawn (Fig. 4) for values of $2h/T$ equal to 0·1, 0·3, 0·5, 0·7 and 0·9. Since the period of the r-curve is half that of the y-curve, it is not necessary to carry the calculations beyond $u/T = \frac{1}{8}$, *i.e.*, 45°. The values of r were usually calculated at every 5°, with supplementary values at 1° or 2·5° over the range from 0° to 15°.

Finally, to obtain from these calculations the frequency-distribution of r, on the assumption that u is equally likely to fall at any point of the range between 0 and T, the values of u/T for which $r = 0{\cdot}1, 0{\cdot}2, \ldots 0{\cdot}9, 0{\cdot}91, 0{\cdot}92 \ldots$ up to the maximum, were found by the use of the second difference interpolation-formula equation (4) on p. xiv of *Tables for Statisticians and Biometricians*. If k_1, k_2 are the values of u/T corresponding to the values r_1, r_2 of r, $k_2 - k_1$ measures the frequency of values of r between these limits. The interpolation-formula referred to did not give results of any great precision, for the intervals chosen, at some parts of the range, but no more accuracy was desired than sufficed to draw the rough charts (Figs. 5—9).

Taken over a whole period, $2h = T$ and

$$\begin{aligned} M_1 &= M_2 = 0, \\ \sigma_1^2 &= \sigma_2^2 = 0{\cdot}5, \\ p' &= p = \tfrac{1}{2} \cos 2\pi \frac{\tau}{T}. \end{aligned}$$

Hence

$$r = \cos 2\pi \frac{\tau}{\mathrm{T}}, \tag{6}$$

a formula which holds good also, it may be noted, if we do not treat variation as continuous, but are given only ordinates at equal intervals throughout the period. If τ/T is $\frac{1}{6}$, or $2\pi\tau/\mathrm{T}$ is 60°, $r = 0{\cdot}5$, and this is the second case taken for illustration. The interval $2h/\mathrm{T}$ chosen was $0{\cdot}2$. For this case the equations for means, etc., become, converting the angle into degrees and writing for brevity,

$$\theta = 360u/\mathrm{T};$$

and taking this angle in degrees :—

$$\left.\begin{aligned} \mathrm{M}_1 &= 0{\cdot}935\ 4893 \sin\theta \\ \mathrm{M}_2 &= 0{\cdot}935\ 4893 \sin(\theta + 60) \\ \Sigma_1^2 &= 0{\cdot}5 - 0{\cdot}378\ 41335 \cos 2\theta \\ \Sigma_2^2 &= 0{\cdot}5 - 0{\cdot}378\ 41335 \cos 2(\theta + 60) \\ p' &= 0{\cdot}25 - 0{\cdot}378\ 41335 \cos(2\theta + 60) \end{aligned}\right\} \tag{7}$$

The curve for r as a function of u is not shown, but the frequency-distribution is given in Fig. 10.

Referring back to equation (6), we may obtain a general expression that is utilized in Section V. Suppose we have some function expanded in a Fourier Series, the time T being the fundamental period, so that—omitting the constant term which will not affect any correlations—

$$y_1 = \mathrm{S}\left\{\mathrm{A}_m \sin 2\pi \frac{t + \phi_m}{m\mathrm{T}}\right\} \tag{8}$$

where $m = 1, \frac{1}{2}, \frac{1}{3} \ldots$ Then integrating over time T,

$$\sigma_1^2 = \tfrac{1}{2}\mathrm{S}(\mathrm{A}_m^2), \tag{9}$$

the products of terms of unlike period vanishing. If we take the same function shifted in phase by the amount τ so that

$$y_2 = \mathrm{S}\left\{\mathrm{A}_m \sin 2\pi \frac{t + \tau + \phi_m}{m\mathrm{T}}\right\}, \tag{10}$$

the mean product is

$$p = \tfrac{1}{2}\mathrm{S}\left\{\mathrm{A}_m^2 \cos 2\pi \frac{\tau}{m\mathrm{T}}\right\}. \tag{11}$$

The standard deviation is, of course, the same as before, and hence

$$r = \frac{1}{S(A_m^2)} S\left\{A_m^2 \cos 2\pi \frac{\tau}{mT}\right\}. \quad (12)$$

Plotted to τ as base, the curve for r is compounded of cosine curves of the original periods, all shifted into phase at $\tau = 0$, with *intensities* substituted for their amplitudes.

APPENDIX II.—*The relations between the serial correlations of a sum series and of its difference series, when the series may be regarded as indefinitely long.*

THE DIRECT PROBLEM.—Let $u_0, u_1, u_2, u_3 \ldots u_s \ldots u_m$ be a series for which the serial correlations are $r_1, r_2, r_3 \ldots r_k, \ldots r_k$ being the correlation between u_s and u_{s+k}. Let σ_u be the standard deviation of the u's. Then

$$\Sigma (u_{s+1} - u_s)^2 = \Sigma (u_{s+1})^2 + \Sigma (u_s^2) - 2\Sigma (u_{s+1}\, u_s).$$

The sum on the left is extended over all first differences. Hence on the right the first sum only covers u_1 to u_m, where u_m is the last in the series, and the second only u_0 to u_{m-1}: we shall suppose the series to be so long that means and standard deviations are not sensibly affected by this dropping of initial and terminal observations. On this assumption, reading the u's as deviations, we have

$$\sigma_s^2 = 2\sigma_u^2 (1 - r_1). \quad (1)$$

Next, to determine the correlation between adjacent first differences, we have

$$\Sigma (u_{s+2} - u_{s+1})(u_{s+1} - u_s) = \Sigma (u_{s+2}\, u_{s+1}) + \Sigma (u_{s+1}\, u_s) - \Sigma (u_{s+2}\, u_s) - \Sigma (u_{s+1}^2).$$

On the same assumption as before, both the first and the second terms on the right may be written $Nr_1\sigma_u^2$, the third $Nr_2\sigma_u^2$, and the last $N\sigma_u^2$. Hence by (1), using ρ's for the serial correlations of the difference series,

$$\rho_1 = \frac{2r_1 - r_2 - 1}{2(1 - r_1)}. \quad (2)$$

Proceeding in precisely the same way, we have generally

$$\rho_k = \frac{2r_k - r_{k+1} - r_{k-1}}{2\,(1 - r_1)}, \tag{3}$$

which checks with (2), noting that $r_0 = 1$. But this may evidently be written

$$\rho_k = -\frac{1}{2(1 - r_1)}\,\Delta^2(r_{k-1}). \tag{4}$$

This gives the most convenient method of working out the limiting difference correlations when the serial correlations for the u's are given: the second differences of the series 1, r_1, r_2, r_3 . . . are formed, and multiplied though by $1/2\ (1 - r_1)$, reversing signs. Note also that if the ρ's are positive, the graph of the r's must be concave downwards, as in Fig. 12 of the paper; if the ρ's are negative, the graph of the r's must be concave upwards; and, finally, if the ρ's are zero, the graph of the r's must be a straight line, as in Fig. 11 of the paper.

Suppose that the first differences are formed with the interval h instead of the interval unity, *i.e.*, the differences are taken as

$$\begin{aligned} &u_h - u_0 \\ &u_{h+1} - u_1 \\ &u_{h+2} - u_2 \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ &u_{h+k} - u_k. \end{aligned}$$

Then by similar reasoning we have

$$_h\rho_k = \frac{2r_k - r_{k+h} - r_{k-h}}{2\,(1 - r_h)}. \tag{5}$$

Putting $h = 1$, this becomes identical with (3). Where $k < h$, remember that $r_{k-h} = r_{h-k}$.

THE INVERSE PROBLEM.—Now consider the inverse problem: given the ρ's for the difference series, required to find out what we can about the r's for the sum series. We will consider only certain special cases.

A.—*The differences are random*, so that all ρ's are zero. We then have

$$\Delta^2 r_{k-1} = 0 \tag{6}$$

for all values of k. It is obvious that the r's must form an arithmetic

series, but the series is not determinate unless one term, say r_1, is given. The series is then 1, r_1, $2r_1 - 1$, $3r_1 - 2$, etc., or generally

$$r_k = k\,r_1 - (k - 1). \tag{7}$$

In any actual case the r_k's will not, of course, form a strictly arithmetical series, owing partly to the inevitable chances of sampling and partly to the "end-effects," and consequently a "best fitting" series will have to be determined by assigning some special value, say r_1', to r_1. The readiest, and on the whole the best, method to determine r_1' seems to be to make the sum of the calculated correlations equal to the sum of those observed, so that the mean error is zero. This gives

$$\tfrac{1}{2}k(k+1)\,r_1' = \tfrac{1}{2}k\,(k-1) + \Sigma\,(r_k), \tag{8}$$

or for the special case when k is 10,

$$11\,r_1' = 9 + 0{\cdot}2\Sigma\,(r_k). \tag{9}$$

Fitting by least squares offers no difficulty, but does not make the mean error zero and does not seem, in the cases tried, at all markedly to reduce the errors.

I worked out r_1 to r_{10} for my first three series with random differences (A_1, B_1, and C_1) each of 100 terms. The original correlations were taken to five figures, and r_1' was calculated from these. Table A shows the observed correlations against the fitted series to three digits. For A_1 and C_1 the fit seems very satisfactory; for B_1 it is poor. Fig. 11 of the paper (p. 24) shows the results. It is odd that all the series give positive errors (r in excess of the calculated values) in the later terms. Is this due in some way to the end-effect, or is it merely chance ? The next case does not show the same thing.

TABLE A.—Comparison of serial correlations for three series with random differences, with fitted arithmetical progressions.

	Series A_1.		Series B_1.		Series C_1.	
	Observed correlation.	Calculated series.	Observed correlation.	Calculated series.	Observed correlation.	Calculated series.
1	0·963	0·976	0·882	0·921	0·954	0·950
2	0·934	0·951	0·792	0·843	0·900	0·900
3	0·911	0·927	0·705	0·764	0·842	0·850
4	0·889	0·903	0·626	0·686	0·780	0·800
5	0·879	0·878	0·587	0·607	0·729	0·750
6	0·859	0·854	0·525	0·528	0·689	0·700
7	0·836	0·829	0·513	0·450	0·654	0·650
8	0·817	0·805	0·455	0·371	0·613	0·600
9	0·797	0·781	0·350	0·292	0·571	0·550
10	0·776	0·756	0·242	0·214	0·519	0·500

But these figures and the chart of Fig. 11 will perhaps, and legitimately, raise a difficulty in the mind of the reader. If the lines are continued downwards, they will lead first to negative and then to impossible values of the correlation. Any line with a finite slope must give the same trouble if continued sufficiently far. But the point is that we can only obtain such series as those of Table A if the serial correlations are determined from a *finite* series, and for a finite series (6) will be only approximately true for moderate values of k and will cease to be valid for large values. As the u-series is extended indefinitely, σ_u tends to increase indefinitely: but ${}_h\sigma_\delta$—the standard deviation of first differences with an interval h—remains finite for all finite values of h. Hence, since

$$ {}_h\sigma_\delta{}^2 = 2\sigma_u{}^2 (1 - r_h), $$

r_h must tend to unity for all finite values of h. For an indefinitely long series of the type considered all the serial correlations tend to unity.

B.—*The differences are correlated,* ρ_k being a linear function of k. Since ρ_0 must be unity, we may take

$$ \rho_k = 1 - \alpha k. \tag{10} $$

Hence

$$ \Delta^2 (r_{k-1}) = -2 (1 - r_1) (1 - \alpha k) \tag{11} $$

Since the second differences are a linear function of k, the series $r_0, r_1, r_2 \ldots r_k$ must evidently be a polynomial in k involving powers up to the third, say,

$$ r_k = 1 + bk + ck^2 + dk^3. \tag{12} $$

Here

$$ \Delta^2 (r_{k-1}) = 2 (c + 3dk), \tag{13} $$

and hence, equating coefficients

$$ \left.\begin{aligned} c &= -(1 - r_1) \\ d &= \tfrac{1}{3}\alpha (1 - r_1) \end{aligned}\right\} \tag{14} $$

Inserting these values in (12) and putting $k = 1$, we have

$$ b = -\tfrac{1}{3}\alpha (1 - r_1) = -d. \tag{15} $$

Hence finally, writing for brevity

$$ 1 - r_1 = m, \tag{16} $$

we have

$$ r_k = 1 - mk^2 + \tfrac{1}{3}\alpha mk (k^2 - 1). \tag{17} $$

For the special case of correlated differences in the experiments of Section IV, $\alpha = \frac{1}{11}$. If we determine the " best " value of m, say, m', by making the sum of the observed values of r_s from $s = 1$ to $s = k$ equal to the sum of the values calculated from (17), we have as the general equation for determining m',

$$\sum_0^k (r_s) = k - m' \{\tfrac{1}{6}k\,(k+1)\,(2k+1) + \tfrac{1}{6}\alpha k\,(k+1) - \tfrac{1}{12}\alpha k^2\,(k+1)^2\} \quad (18)$$

which, for $k = 10$ and $\alpha = \frac{1}{11}$, reduces to

$$295m' = 10 - \Sigma\,(r_s). \quad (19)$$

The first ten serial correlations for the experimental series A_2, B_2, C_2 were calculated in the original work to five figures. Table B shows these observed values to three figures against the series (17) fitted by equation (19). For series A_2 and C_2 the fit is excellent : B_2, like B_I, is rather more irregular. Graphs are shown in Fig. 12, p. 25.

TABLE B.—Comparison of serial correlations for three series with correlated differences, with fitted cubic series.

	Series A_2.		Series B_2.		Series C_2.	
	Observed correlation.	Calculated series.	Observed correlation.	Calculated series.	Observed correlation.	Calculated series.
1	0·999	0·999	0·991	0·991	0·996	0·996
2	0·995	0·995	0·966	0·965	0·986	0·985
3	0·990	0·989	0·927	0·923	0·968	0·968
4	0·982	0·982	0·876	0·868	0·945	0·945
5	0·973	0·972	0·814	0·802	0·916	0·918
6	0·963	0·962	0·741	0·725	0·883	0·886
7	0·951	0·950	0·654	0·640	0·848	0·851
8	0·937	0·937	0·552	0·548	0·811	0·813
9	0·922	0·924	0·436	0·451	0·774	0·773
10	0·906	0·910	0·307	0·351	0·738	0·731

We have only worked out the calculated series up to r_{10}. Since $\Delta^2\,(r_{10})$ is zero, the series beyond this point becomes linear.

For this type of series, as for the last, serial correlations such as are shown in Table B are only possible for a finite series. For an infinite series, all serial correlations would tend to unity.

C.—*The second differences of the given series are random*, i.e., the given series is the second sum of a random series.

In this case the first differences of the given series are the sum

of a random series, and therefore the serial correlations of the differences are given by equation (7), or, writing this in the form of (10),

$$\rho_k = 1 - k(1 - \rho_1), \tag{20}$$

so that the α of equation (10) is $1 - \rho_1$. The r-series is consequently given by (17).

So far as samples of no more than 10 observations are concerned, the special mode of forming the series used for the experiments on series with correlated differences leads therefore to precisely the same results as regards the frequency-distribution of correlations, as would the second summation of a random series. This conclusion was utilized in Section III of the paper.

The actual mode of formation used was to sum successive batches of 11 terms of the random series and then use these as differences. If $a_1, a_2, a_3 \ldots a_3$ is the random series, $u_1, u_2, u_3 \ldots u_3$ the final series,

$$\begin{aligned}
\Delta^1(u_1) &= a_1 + a_2 + \ldots + a_{11} \\
\Delta^1(u_2) &= a_2 + a_3 + \ldots + a_{12} \\
&\cdots\cdots\cdots \\
\Delta^1(u_{10}) &= a_{10} + a_{11} + \ldots + a_{20},
\end{aligned}$$

and therefore

$$\begin{aligned}
\Delta^2(u_1) &= a_{12} - a_1 \\
\Delta^2(u_2) &= a_{13} - a_2 \\
&\cdots\cdots\cdots \\
\Delta^2(u_{10}) &= a_{21} - a_{10}.
\end{aligned}$$

Within the sample of 10 terms only, second differences *are* uncorrelated. Not until we reach $\Delta^2(u_{12})$ would there be a negative correlation with $\Delta^2(u_1)$.

D.—*A special case of causation.*—Let us now, instead of assuming a special form for the serial correlations, assume a special mechanism of causation and ask to what serial correlations it leads.

It is familiar that if we take a set of dice of which n_1 are red, n_2 white and n_3 green, the correlation between the number of successes in the red and white together and the number of successes in the white and green together is $n_2/(n_1 + n_2)$ or the proportion of dice common to the two sets. Now suppose that the magnitude of our variable in any year is determined by a number of independent, unitary, elementary causes (analogous to the dice), and that n_1 of these causes come into existence in every successive year, of which pn_1 survive to the next year only, p^2n_1 for two years, and so

on. The total number of causes operating in any one year will then be

$$n = n_1 (1 + p + p^2 + p^3 + \ldots)$$
$$= n_1/(1-p), \tag{21}$$

and a proportion p of these will be common to years s and $s+1$, p^2 to years s and $s+2$, and so on. The serial correlations will therefore be 1, p, p^2, p^3. . . . As the graph of this geometric series is concave upwards, we have the rather unexpected result that for this type of continuity of causation the serial correlations for the differences must be *negative*. We have, in fact,

$$\rho_k = -\tfrac{1}{2}(1-p)p^{k-1}. \tag{22}$$

The difference correlations, from ρ_1 onwards, are a geometric series of negative sign.

It is of interest to ask now a further question. Supposing that such a system of causation as we have assumed determines, not the *values* of the variable, but its *changes* from year to year, *i.e.*, the first differences, what will be the serial correlations for the sum series ?

We have now

$$\Delta^2 (r_{k-1}) = -2(1-r_1)\rho^k, \tag{23}$$

and the general solution is of the form

$$r_k = A - Bk + Ce^{-bk}. \tag{24}$$

Hence

$$\Delta^2 (r_{k-1}) = Ce^{-b}(e^b - 1)^2 e^{-bk} = -2(1-r_1)\rho^k.$$

We must therefore have

$$e^{-b} = \rho, \tag{25}$$

and thence, writing for brevity, $1 - r_1 = m$ as before,

$$C = -\frac{2m\rho}{(1-\rho)^2}. \tag{26}$$

Further, for $k = 0$, $r_k = 1$, and therefore

$$A = 1 - C. \tag{27}$$

Whence

$$r_1 = 1 + \frac{2m\rho}{(1-\rho)^2} - B - \frac{2m\rho^2}{(1-\rho)^2},$$

or

$$B = m\frac{1+\rho}{1-\rho}. \tag{28}$$

Therefore, finally, (24) becomes

$$r_k = 1 - m\frac{1+\rho}{1-\rho}k + \frac{2m\rho}{(1-\rho)^2}(1-\rho^k). \tag{29}$$