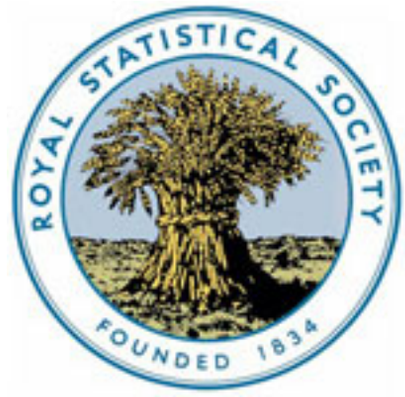


On the Time-Correlation Problem, with Especial Reference to the Variate- Difference Correlation Method
Author(s): G. Udny Yule
Source: *Journal of the Royal Statistical Society*, Vol. 84, No. 4 (Jul., 1921), pp. 497-537
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2341101
Accessed: 24/06/2014 23:39

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Wiley and *Royal Statistical Society* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society*.

http://www.jstor.org

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

Vol. LXXXIV.] [Part IV

JOURNAL

OF THE ROYAL STATISTICAL SOCIETY.

JULY, 1921.

On the Time-correlation Problem, with Especial Reference to the Variate-difference Correlation Method.

By G. Udny Yule, M.A., F.R.S.

[Read before the Royal Statistical Society, May 24, 1921, the President, Sir R. Henry Rew, K.C.B., in the Chair.]

I. *Introduction: the nature of the problem.*

I have entitled this paper "On the time-correlation problem," by which I mean the problem of elucidating (usually in these days, but not necessarily, with the aid of the coefficient of correlation) the relations subsisting between two quantities varying with the time. The problem has, of recent years, received a good deal of attention and, having regard especially to the latest developments, it seems desirable to survey the ground.

The first question to which I wish to devote some consideration is this: what is in fact the real nature of the problem? To answer this I propose briefly to review some of the work that has been done on the subject.

Problems of the kind have, of course, interested statisticians for many years, from days long before the coefficient of correlation was invented. Not to go too far back I take as the first illustration a paper by Professor Poynting in the *Journal* for 1884 (ref. 1).* Poynting's problem was the nature of the relations between movements of wheat-prices in England, France and Bengal, and of cotton and silk imports into Great Britain. Taking the first case for example, he points out that the prices as they stand are not suitable for his purpose: the figures are very irregular and sometimes the average price rises so considerably "that what may be a very low

* *Cf.* list of references at end of Paper.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

"price at a particular time may be a high price compared with the "average of other times." "In order therefore to determine the "fluctuations, we require to know not only the actual price, but "whether that price is above or below the average *for that time.* "It becomes necessary then to average the prices in some way so "as to obtain a standard for each year," and he takes as that standard the average of the ten years of which the given year is the fifth. Further "there are so many irregularities of short duration, "say two or three years, that it is more convenient to take, instead "of the price for each year, the average for a short period," and accordingly Poynting replaces the price of the year by the average price of the four years of which the given year is the second. Finally the curve is drawn showing the four-year average as a percentage of the corresponding "standard" price, and these curves are the basis of the discussion. Poynting's process consists then in an endeavour to isolate for discussion oscillations of about ten years' duration, eliminating so far as possible both more gradual and more rapid movements.

Hooker, in his Paper on the relation between the marriage-rate and trade (ref. 2) uses a method which is essentially the same but is rendered much more precise by the use of the correlation-coefficient. He points out that, while it is often said that the marriage rate and the trade per head of the population are closely connected, it is readily seen on drawing a diagram that it is only the oscillations which correspond: the oscillations rise and fall together, although over the period considered the trade curve has risen and the marriage rate has fallen. To eliminate this general movement and "correlate the "oscillations of the two curves" he suggests that "all deviations "should be reckoned, not from the average of the whole period, "but from the instantaneous average at the moment," the curve or line representing the successive instantaneous averages being termed the *trend.* In the given case he regards it as being sufficient for practical purposes to take an average of nine years round a given year as the "instantaneous average," and calculates the correlations between the oscillations accordingly. The method, it will be remembered, led to very interesting results, and the calculation of correlations between deviations of the marriage rate and deviations in the trade or other curves for earlier and later years rendered it possible approximately to determine the time-lag between the two.*

* Some curves obtained by this method and exhibiting graphically the relation between the oscillations in the marriage rate, in prices, in unemployment, &c., will be found in my Paper on the marriage and birth-rates, *Journal,* vol. lxix, 1906, p. 88.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

Hooker did not consider it necessary to eliminate the irregularities of short duration "say two or three years," before correlating, in the same way as Poynting did.

Further, these movements of shorter duration were precisely the point of interest in the case with which he was dealing immediately afterwards (ref. 3). In that Paper, on the effect of the suspension of the Berlin Produce Exchange, Hooker suggested that "the "problem—to what extent do the fluctuations at one market follow "those at another ?—requires to be attacked by the use of some "formula which should correlate the differences between the prices "on consecutive days, instead of the differences from the average "price." This is the first suggestion, so far as I am aware, of a "difference-method." The work was not carried out, however, till four years later, in 1905 (ref. 5), in a Paper which specifically discussed the purpose of the method and gave other illustrations of its use. The introductory paragraph to this Paper is illuminating as showing Hooker's view of the whole problem. If, he points out, a coefficient of correlation is formed in the ordinary way between the simultaneous values of two time-variables, "we shall obtain (*a*) a value that is "very high if the 'secular' changes are similar (the value being "almost entirely independent of the similarity or otherwise of the "more rapid changes), (*b*) a value approximating to 0 if the 'secular' "changes are of quite dissimilar character, even although the "similarity of the smaller rapid changes may be extremely marked." He then refers to the moving-average method, just described, "a "method which is often useful in the particular case of two variables "subject to oscillations of a more or less periodic character." "I "now desire to direct attention," he continues, "to a method of a "very simple character, applicable whether the smaller rapid "changes under investigation are of a quasi-periodic character or "not." This statement is interesting as relating the first or moving-average method to the second or difference-method, and the point is brought out again in the concluding paragraph of the Paper. "Perhaps it may be suggested that, speaking generally, in examining "the relationship between two series of observations extending over "a considerable period of time, correlation of absolute values "(deviations from the arithmetic mean) is the most suitable test of "'secular' interdependence, and may also be the best guide when "the observations tend to deviate from an average that may be "regarded as constant. Correlation of the deviations from an "instantaneous average (or *trend*) may be adopted to test the "similarity of more or less marked periodic influences. Correlation "of the differences between successive values will probably prove

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

"most useful in cases where the similarity of the shorter rapid "changes (with no apparent periodicity) are the subject of investiga- "tion, or where the normal level of one or both series of observations "does not remain constant." I would only be inclined to modify the general view expressed in these citations by suggesting that the "shorter rapid changes" which give the appearance of irregularity to a statistical curve may not be wholly non-periodic in character but may be oscillations of two years' duration or thereabouts. Poynting's wording (irregularities of short duration, say two or three years) also suggests this view. The birth-rate curve, for example, looks notably irregular and I am almost certain contains an oscillation of about two years' duration, possibly due to the average interval between births, which accounts for this appearance.*

Miss F. E. Cave, in her Paper of 1904 on barometer correlations (ref. 4) which preceded the Paper last quoted, assigns no reason for the choice, in one section of her work, of the daily rise or fall as the variable to be correlated. She only remarks: "Another point "which seemed to deserve investigation was the correlation between "the daily rise or fall of the barometer at Halifax or Wilmington" (p. 407). In a later Paper by Miss B. M. Cave and Professor Pearson (ref. 9) it is stated, however, that she "endeavoured to get rid of "seasonal change by correlating first differences of daily readings "at two stations." This then assigns the same reason as is given by Hooker for the use of the difference-method: the use of first differences represents an attempt to get at the correlation of short-period, as against long-period, changes.

The Paper by Monsieur March, which closes this first group of Papers (ref. 6), reviews the whole problem, evidently independently. He deals first with the difference-method, and then proceeds to the method of the moving average which was determined in his case by graphical interpolation. I should like to make lengthy quotations, but space will hardly permit. Generally speaking, his ideas seem very similar to those of Hooker. He points out that one must distinguish "des changements annuels, des changements poly- "annuels (décennaux par exemple), des changements séculaires, "sans parler des périodes plus courtes qu'une année" (Section iv). Most statistics are too recent in date to give us much information about the secular correlations. But we are concerned with the "changements annuels" (*i.e.*, changes from one year to the next and therefore, if periodic, of a periodicity of two years or so), and the difference-method deals with these. Changes of a slower kind, it is suggested, might be dealt with by applying the difference-

* *Cf. Journal*, vol. 69, pp. 125–6.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

method, for example, to decennial averages, but the moving-average method gives greater precision. Monsieur March emphasises, as does Hooker, the fact that all these movements of greater or less rapidity, of which the total movement may be regarded as composed, are due—largely or wholly—to different groups of causes, and the correlations found, say, between the seasonal movements, between the oscillations with a period of about two years, and between the secular movements in two variables, may therefore differ from each other not only in magnitude but in sign, and they do in fact so differ.

I hope that I have not been unduly labouring the obvious, but wish to emphasize that it is not my view alone but the view of most writers on the subject up to 1914, that the essential difficulty of the time-correlation problem is the difficulty of isolating for study different components in the total movement of each variable : the slow secular movement, probably non-periodic in character or, if periodic, with a very long period ; the oscillations of some ten years' duration, more or less, corresponding to the wave in trade ; the rapid movements from year to year which give an appearance of irregularity to the curve in a statistical chart and which may in fact be irregular or may possess a quasi-periodicity of some two years' duration ; the seasonal movement within the year, and so on. It is unfortunate that the word " periodic " implies rather too much as to the character of such more rapid movements ; few of us, I suppose, now believe that they are strictly periodic in the proper sense of the term, and hence the occurrence in writings on the subject of such terms as " quasi-periodic " and " pseudo-periodic." They are wave-like movements, movements which can be readily represented with a fair degree of accuracy over a moderate number of years by a series of harmonic terms but which cannot be represented in the same way, for example, by a polynomial ; movements in which the length of time from crest to crest of successive waves is not constant, and in which, it may be added, the amplitude is not constant either, but would probably, if we could continue our observations over a sufficient number of waves, exhibit a frequency distribution with a fairly definite mode ; to avoid the suggestion of strict periodicity and the use of the term *period* I propose to speak of them as *oscillations* of a given *duration*, the word *duration* to imply, not a fixed and constant duration, but an average only. In these terms, the problem of time-correlation may be said to be the isolation, for separate study, of oscillations of differing durations. Most writers up to 1914—indeed all writers so far as I am aware—seem to be agreed on this.

It is accordingly with some surprise that we find a totally

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

different view taken in the Paper by " Student " (ref. 7) in which the extension of the difference-method is proposed. In his introductory paragraph " Student " says : " In the *Journal of the Royal* " *Statistical Society* for 1905, p. 696, appeared a paper by Mr. R. H. " Hooker giving a method of determining the correlation of variations " from the 'instantaneous mean' by correlating corresponding " differences between successive values. This method was invented " to deal with the many statistics which give the successive annual " values of vital or commercial variables ; these values are generally " subject to large secular variations, sometimes periodic, some- " times uniform, sometimes accelerated, which would lead to alto- " gether misleading values were the correlation to be taken between " the figures as they stand." The first sentence of this paragraph confounds the method of the instantaneous mean with the difference-method. The second sentence may express " Student's " views but does not express Hooker's as developed in his Papers. Hooker, and the other writers whom I have cited, regard the secular and the various oscillatory components of the total movement as liable, almost necessarily, to lead to *different*—not " misleading "—values of the correlation. And if " Student " desires to remove from his figures secular movements, periodic movements, uniform movements, and accelerated movements—well, the reader is left wondering with what sort of movements he *does* desire to deal. This appears in what follows. After stating that Professor Pearson had pointed out to him that the first difference-method was " only valid when the " connection between the variables and time is linear," " Student " proceeds to show that if a series of values of x are in random order, so that the correlation between x_r and x_{r+s} is zero for all values of r and s, and if the same holds good for the values of another variable y, then the correlation between the nth differences of x and y is the same as that between x and y themselves. If x and y are not random in their order with respect to time, suppose that—

$$x = X + b.t + c.t^2 + d.t^3 + \ldots$$
$$y = Y + b'.t + c'.t^2 + d'.t^3 + \ldots$$

where X and Y are random in their order. Then, assuming of course that the series in t stops at some finite power of t, we have only to proceed to a sufficiently high order of differences in x and y to eliminate the function of t altogether, and when we have done this the correlation between the nth differences of x and y will be the same as the correlation between the nth differences of X and Y, and therefore the same as the correlation between X and Y themselves. " Hence," continues " Student," " if we wish to eliminate variability " due to position in time or space and to determine whether there is

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

" any correlation between the residual variations, all that has to " be done is to correlate the 1st, 2nd, 3rd, . . . nth differences " between successive values of our variable with the 1st, 2nd, 3rd " . . . nth differences between successive values of the other " variable. When the correlation between the two nth differences " is equal to that between the two $(n + 1)$th differences, this value " gives the correlation required."

" Student " therefore introduces quite a new idea that is not found in any of the writers previously cited. He desires to find the correlation between x and y when every component in each of the variables is eliminated which can well be called a function of the time, and nothing is left but residuals such that the residual of a given year is uncorrelated with those that precede or that follow it.

Anderson (ref. 8) adds much to the mathematics of the generalised method but little to the discussion of its purpose. He refers to " das von Cave und Hooker vorgeschlagene Verfahren, den Korre " lationskoeffizienten zweier oscillierender Variablen durch Berech- " nung erster Differenzen . . . vom evolutorischen Element zu " befreien," and this might pass for a description of Mr. Hooker's intention if it were not for the fact that in a passage at the end of the Paper, cited later, it seems to be claimed for the method that even periodic terms can be eliminated by differencing. The writer apparently regards periodic terms as " evolutorische Elemente."

Miss B. M. Cave and Professor Pearson (ref. 9), at the commencement of the paper in which they give some arithmetical examples of the generalised method, describe " Student's " work practically without criticism, summing up the purpose of the method as the determination of the correlation between x and y " free from the " spurious time (or it might be position) correlation." " The spurious " correlation arising from x and y being both functions of the time " it is also remarked in a preceding sentence " could be got rid of by " correlating the differences of x and y." Such language shows how distant is the standpoint of this group of writers from that of Poynting, Hooker or March. Oscillatory movements that may be described as functions of the time are the whole subject of the discussions by the latter writers ; the correlation between two such movements may arise either because the one is causally dependent on the other, or because both are " functions of the time," *i.e.*, of some third variable or group of variables on which both are causally dependent, as in the case of the correlation between oscillations in the marriage rate and in the proportion of signatures by mark (*cf.* diagram in *Journal,* vol. 69, p. 111). There is nothing essentially " spurious " or " misleading " about such a correlation.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

But which view of the problem is correct? Do we want to isolate oscillations of different durations, two years, ten years, or whatever it may be, or nothing but these random residuals? Personally I cannot hesitate for a moment as to the answer. The only residuals which it is easy to conceive as being totally uncorrelated with one another in the manner supposed are errors of observation, errors due to the "rounding off" of index-numbers and the like, fluctuations of sampling, and analogous variations. And an error of observation or fluctuation of sampling in x would normally be uncorrelated with an error of observation or fluctuation of sampling in y, so that if the generalised variate-difference method did finally isolate nothing but residuals of the kind supposed I should expect it in general to lead to nothing but correlations that were zero within the limits of sampling. The older writers were, I think, perfectly right; the problem is not to isolate random residuals but oscillations of different durations, and unless the generalised method can be given some meaning in terms of oscillations it is not easy to see what purpose it can serve.

"Student" and Anderson assume that for the purpose of their method each variable can be held to consist only of (1) a polynomial function of the time and (2) random residuals. Is this assumption legitimate? As indicated above (p. 501) it seems to me very difficult to suppose so: it would take a parabola of very high order indeed to represent, say, the course of the marriage rate over thirty or forty years—a harmonic series seems to be called for at once. Can we then ignore the presence of harmonic terms in the functions expressing our variables?

"Student" says nothing on this head, but as in his introductory sentence he includes periodic movements with those which may give rise to "altogether misleading values" of the correlation, presumably he assumed that differencing tends to eliminate any periodic terms. Anderson says definitely (p. 279) "so kommen wir zum Schluss, "dass nicht nur Komponenten, die durch eine Parabel höherer "Ordnung darstellbar sind, sondern auch solche, denen nur "tranzendentale Gleichungen (z.B. Sinus-reihen) genügen, beim "endlichen Differenzieren eliminiert werden." Miss Elderton and Professor Pearson (ref. 10, footnote on p. 503), however, reject this conclusion: "we do not believe that a very short periodicity would "be eliminated by the variate-difference method using any moderate "number of differences. We cannot on this point accept Dr. "Anderson's view." But neither "Student," nor Anderson, nor Miss Elderton and Pearson have given any investigation on the point. Persons, in a paper (ref. 11) which I regret I had not seen while

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

making my own investigations and to which I was referred by the kindness of Professor Edgeworth after the present paper had already been written, shows that *alternations* (up-and-down movements in alternate years) are emphasized, but does not consider the general case of a periodic movement covering n years or intervals. I accordingly devote my next section to the effect of differencing on a series of terms given by a harmonic function.

II. *The differences of a harmonic function.*

Let the successive values of the function to be differenced be—

$$u_0 = A \sin\left(2\pi \frac{t+\tau}{T}\right)$$

$$u_1 = A \sin\left(2\pi \frac{t+\tau+h}{T}\right)$$

$$u_2 = A \sin\left(2\pi \frac{t+\tau+2h}{T}\right)$$

and so on, where T is the period, τ gives the phase, h is the interval, and A the amplitude. Then—

$$\begin{aligned}\Delta_0 &= A\left\{\sin\left(2\pi\frac{t+\tau+h}{T}\right) - \sin\left(2\pi\frac{t+\tau}{T}\right)\right\}\\ &= 2A\sin\left(\pi\frac{h}{T}\right)\cos\left(2\pi\frac{t+\tau+0{\cdot}5\,h}{T}\right)\\ &= 2A\sin\left(\pi\frac{h}{T}\right)\sin\left(2\pi\frac{t+\tau+0{\cdot}5\,h+0{\cdot}25\,T}{T}\right)\end{aligned}$$

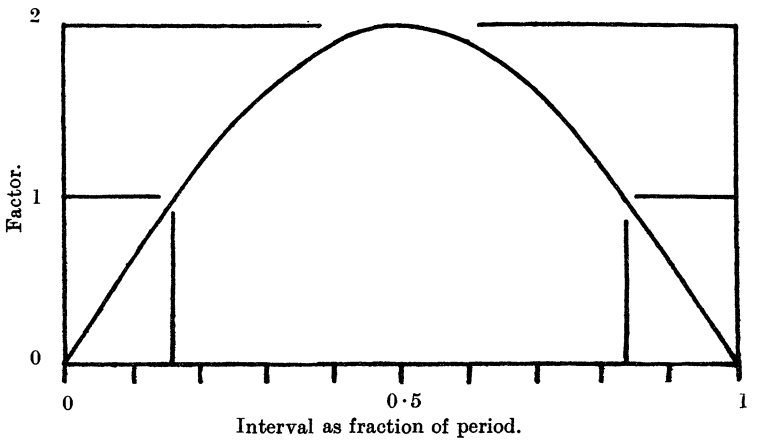
That is to say, the phase is shifted—a point with which we are not at present concerned—and the amplitude is multiplied by $2 \sin (\pi h/T)$, but otherwise the first differences are given by a harmonic term of the same period as the original function. The second differences will, therefore, be derived from the first by multiplying the amplitude again by $2 \sin (\pi h/T)$ and shifting the phase again by the same amount.

Our interest centres on the factor $2 \sin (\pi h/T)$, for according as this is greater or less than unity the successive orders of difference will either continually diverge, or will converge and tend to become smaller and smaller. Fig. 1 shows the course of the function, the half of a sine curve, for all values of h/T from 0 to 1. The value is evidently a maximum and equal to 2 when h is $0{\cdot}5\ T$, and decreases towards zero when h is either greater or less than $0{\cdot}5\ T$. When h is five-sixths of T or one-sixth of T, $2 \sin (\pi h/T)$ is unity. Differencing, then, does not necessarily tend to eliminate periodic terms but tends either to eliminate them or to emphasize them according as h lies outside or within the limits $\frac{1}{6}\ T$ and $\frac{5}{6}\ T$—ignoring the cases where h is greater than T, a case that seems hardly likely to be of practical importance.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

FIG. 1.

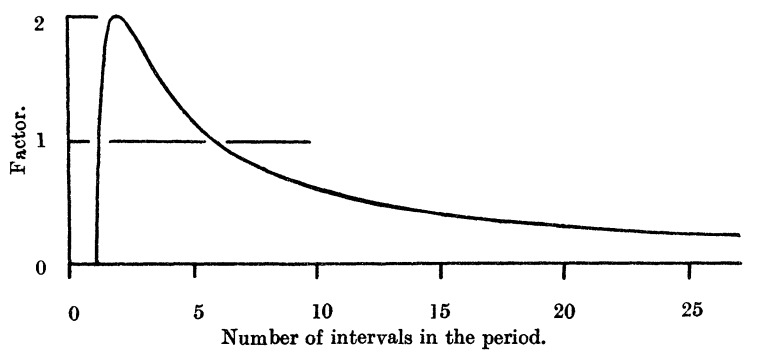
Diagram showing the factor by which the amplitude of a harmonic function is multiplied on differencing, when the interval is a given fraction of the period.



If our unit of time and interval for differencing is the year, differences will diverge for all periods between 1·2 years and 6 years, and will converge only for periods outside those limits. Fig. 2 is constructed to show the values of the multiplying factor more readily for this case. The value rises rapidly as the period is increased from one year to two years, and then very slowly tails off.

FIG. 2.

Diagram showing the factor by which the amplitude of a harmonic function is multiplied on differencing, when there are a given number of intervals to the period.



This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

In Table I are given the values of powers of the factor up to the sixth, to five decimal places (seven were retained for the calculation). These powers show the amplitudes for successive orders of difference, taking the initial amplitude as unity. A period of two years (or intervals) will have its amplitude multiplied sixty-four-fold by taking differences up to the sixth: a term of three years' period, twenty-seven-fold: a term of four years' period, eight-fold, and so on. It is not till we reach the term of six years' period that the amplitude is unaltered by differencing. Thereafter the effect is to decrease the amplitude. A term of seven years' period will have only 43 per cent. of its amplitude left in the sixth differences; a term of eight years' period, only 20 per cent.; a term of eleven years, only 3 per cent.

The effect of differencing is accordingly, not to eliminate periodic terms, but selectively to emphasize those with a period of two years, or generally with a period of two intervals (*cf.* Persons, ref. 11). Miss Elderton and Professor Pearson were perfectly right in rejecting Anderson's sweeping conclusion.

TABLE I.—*Showing, for harmonic terms of unit amplitude and* 2, 3, *etc. to* 15 *intervals period, the amplitudes of the* 2*nd*, 3*rd*, *etc., up to the sixth differences.*

Order of difference.	Period in intervals.						
	2.	3.	4.	5.	6.	7.	8.
1	2	1·73205	1·41421	1·17557	1	·86777	·76537
2	4	3	2	1·38197	1	·75302	·58579
3	8	5·19615	2·82843	1·62460	1	·65345	·44834
4	16	9	4	1·90983	1	·56704	·34315
5	32	15·58846	5·65685	2·24514	1	·49206	·26263
6	64	27	8	3·64745	1	·42699	·20101
	9.	10.	11.	12.	13.	14.	15.
1	·68404	·61803	·56347	·51764	·47863	·44504	·41582
2	·46791	·38197	·31749	·26795	·22909	·19806	·17291
3	·32007	·23607	·17890	·13870	·10965	·08815	·07190
4	·21894	·14590	·10080	·07180	·05248	·03923	·02990
5	·14976	·09017	·05680	·03716	·02512	·01746	·01243
6	·10244	·05573	·03200	·01924	·01202	·00777	·00517

As these results are curious and may seem rather unexpected, it may be as well to give some numerical illustrations. If h be half the period the result is almost obvious; taking the phase as such that

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

the values of u fall alternately at the maxima and minima of the wave, and an amplitude of unity, we have—

u	Δ
+1	−2
−1	+2
+1	−2
−1	+2
+1	

and so on. The amplitude is doubled and the phase is shifted by a quarter period plus half the value of h/T—that is another quarter period, or half a period altogether.

If h be a quarter of the period and zero time be the time corresponding to u_0 we have the following results—

u	Δ	Δ^2
0	+1	−2
+1	−1	0
0	−1	+2
−1	+1	0
0	+1	−2
+1	−1	
0		

Here the meaning of the first differences is not obvious, but the second differences clearly repeat the first differences with a doubling of the amplitude, as required by Table I, and a shift in phase of 270° or three-quarters of a period. By the formula shown the shift of phase on differencing is $\frac{1}{4} + \frac{1}{2}\, h/T$ and as h/T is $\frac{1}{4}$ that is $\frac{3}{8}$ of a period, or 135°, and therefore 270° on differencing twice. As the +1 of the leading first difference corresponds to sin . 135° the amplitude of the curve of first differences is $\sqrt{2}$ or 1·414 . . . again as in Table I.

If h is exactly one-sixth of the period Table I indicates that the amplitude is unaltered by differencing, while the formula gives a shift of phase amounting to one-third of the period We have in fact—

u	Δ
0·	+0·866
+0·866	0
+0·866	−0·866
0·	−0·866
−0·866	0
−0·866	+0·866
0·	+0·866
+0·866	0
+0·866	

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

If h be anything less than one-sixth of the period the amplitude is lessened on differencing. Thus if h be one-eighth of the period or 45°, we have—

u	Δ
0	+ ·707
+ 0·707	+ ·303
+ 1·000	− ·303
+ 0·707	− ·707
0	− ·707
− 0·707	− ·303
− 1·000	+ ·303
− 0·707	+ ·707
0	+ ·707
+ 0·707	

Here the shift in phase is $\frac{5}{16}$ of the period or 112·5° and sin . 112·5° is 0·924, so that the reduction of amplitude is $\frac{707}{924}$ or 0·765 . . . as shown in Table I.

The effect then of differencing the values of a function which is given by a series of harmonic terms is not gradually to extinguish all the terms, but selectively to emphasize the term with a period of 2 intervals; terms with a period between 2 and 6 intervals, or between 2 and 1·2 intervals have their amplitude increased, but not so largely; terms with a period between 1 and 1·2 intervals, or greater than 6 intervals, are reduced in amplitude. Further, every term is altered in phase, by an amount depending on its period. Correlations between high differences will accordingly *tend* to give the correlations between component oscillations of very short period—predominantly of a two-yearly period, in so far as such oscillations exist in the original observations, even though they may not be the most conspicuous or characteristic oscillations.

The data used by Miss Cave and Professor Pearson to give numerical illustrations of the method (ref. 9) were some series of index-numbers for economic progress in Italy. Each index-number is given only to the nearest unit (1 per cent.) and differences are taken up to the sixth. The data for the most part (and with some exceptions, *e.g.* the index for 8. Coal) run very smoothly, the numbers rising with increasing rapidity towards the more recent years and showing no conspicuous oscillations; Table II gives the data for the Savings Banks index, with the differences, as an example. Table III gives in col. 2 the mean-deviation of each set of sixth differences, from its arithmetic mean, as a standard of comparison,

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

TABLE II.—*Index numbers of progress of Savings Banks in Italy, with their differences.*

Year.	Index-number.	Δ^1.	Δ^2.	Δ^3.	Δ^4.	Δ^5.	Δ^6.
1885	47	+ 6	− 4	+ 4	− 5	+ 7	−10
1886	53	+ 2	0	− 1	+ 2	− 3	+ 7
1887 ..	55	+ 2	− 1	+ 1	− 1	+ 4	−16
1888 ...	57	+ 1	0	0	+ 3	−12	+29
1889 ...	58	+ 1	0	+ 3	− 9	+17	−23
1890 ...	59	+ 1	+ 3	− 6	+ 8	− 6	− 4
1891	60	+ 4	− 3	+ 2	+ 2	−10	+24
1892 ...	64	+ 1	− 1	+ 4	− 8	+14	−24
1893 ...	65	0	+ 3	− 4	+ 6	−10	+19
1894 ...	65	+ 3	− 1	+ 2	− 4	+ 9	−19
1895 ...	68	+ 2	+ 1	− 2	+ 5	−10	+17
1896 ..	70	+ 3	− 1	+ 3	− 5	+ 7	− 8
1897 ...	73	+ 2	+ 2	− 2	+ 2	− 1	0
1898	75	+ 4	0	0	+ 1	− 1	− 1
1899	79	+ 4	0	+ 1	0	− 2	+ 4
1900	83	+ 4	+ 1	+ 1	− 2	+ 2	+ 3
1901 ...	87	+ 5	+ 2	− 1	0	+ 5	−13
1902 ...	92	+ 7	+ 1	− 1	+ 5	− 8	− 1
1903 ...	99	+ 8	0	+ 4	− 3	− 9	+40
1904	107	+ 8	+ 4	+ 1	−12	+31	−61
1905	115	+12	+ 5	−11	+19	−30	+40
1906	127	+17	− 6	+ 8	−11	+10	− 2
1907	144	+11	+ 2	− 3	− 1	+ 8	
1908	155	+13	− 1	− 4	+ 7		
1909 ...	168	+12	− 5	+ 3			
1910 ..	180	+ 7	− 2				
1911 ...	187	+ 5					
1912	192						

and in col. 3 the amplitude of the two-year period.* In the case of the Savings Banks index, for example, the mean of the sixth differences for the odd years is +7·5, for the even years −7·4, average 7·4 with the positive sign, the sign used being that for the odd years as the initial year is odd. The total for the 22 sixth differences without regard to their sign is 365 and the correction for reduction to the mean being very small this gives a mean deviation of 16·6. It will be seen that in the first case the amplitude of the two-year term is some 80 per cent. of the mean deviation, and it amounts to some half of the mean deviation in lines 2, 4, 6 and 8. If we estimated the signs of the sixth-difference correlations from the signs of these oscillations alone we should be right in thirty-two out of the forty-five cases. Guessing at random we might expect to be right

* Properly speaking this is not, of course, the amplitude, which is really indeterminate. If zero time is the time of u_0, the sine term has no influence on the data. The figure given is only the apparent or effective amplitude.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

TABLE III.

1. Index-number.	2. In sixth differences. Mean deviation.	3. In sixth differences. Mean amplitude of two-year term.	4. In sixth differences. Corresponding amplitude in data.
1. Railways	22·3	−17·3	−0·27
2. Shipping	54·3	−21·4	−0·33
3. Revenue	11·2	− 3·8	−0·06
4. International commerce	56·4	+23·5	+0·37
5. Posts and telegraphs	18·1	+ 3·3	+0·05
6. Stamp duties	7·8	+ 4·1	+0·06
7. Savings banks	16·6	+ 7·4	+0·12
8. Coal	102·6	−53·7	−0·84
9. Tobacco	11·4	+ 1·0	+0·02
10. Coffee	46·6	+12·1	+0·19

in 22·5 cases, with a standard error of $\frac{1}{2}\sqrt{45}$ or 3·35, and there is an excess of 9·5 right cases, so these terms have a marked influence in the results. But can we regard them as significant? Col. 4, giving the figures of col. 3 divided by 64, shows the corresponding amplitudes in the original data; in no case is the amplitude equal to a unit, and in four cases it is less than one-tenth of a unit, *while the original data were rounded off to the nearest unit.* Our division by 64 may not be strictly correct, as is suggested by the next example, since we may not be dealing with a term of precisely two years' period; but even if the true value of the factor be as low as 40 it will only bring up a single amplitude (8. Coal) above the value of a unit. The process of differencing so magnifies even trifling figures that they become important and may finally dominate.

As another case on which to study the effect of differencing I chose the data given by Miss Elderton and Professor Pearson for infantile mortality in males (ref. 10). The data refer to England and Wales, 1850 to 1908 inclusive, the figures being worked out to three places of decimals so as to ensure reasonable accuracy in the higher differences. The movement of infantile mortality is very different from that of the index-numbers previously studied, it shows large oscillations and at times, as between 1876 and 1887, as will be seen from the chart in Figure 3, a real or apparent tendency to up-and-down movement in alternate years. Further there is comparatively little long-period or secular movement until after 1901. I hoped accordingly that it might be possible in this case to trace the magnification of the alternate oscillation through the whole series of differences.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

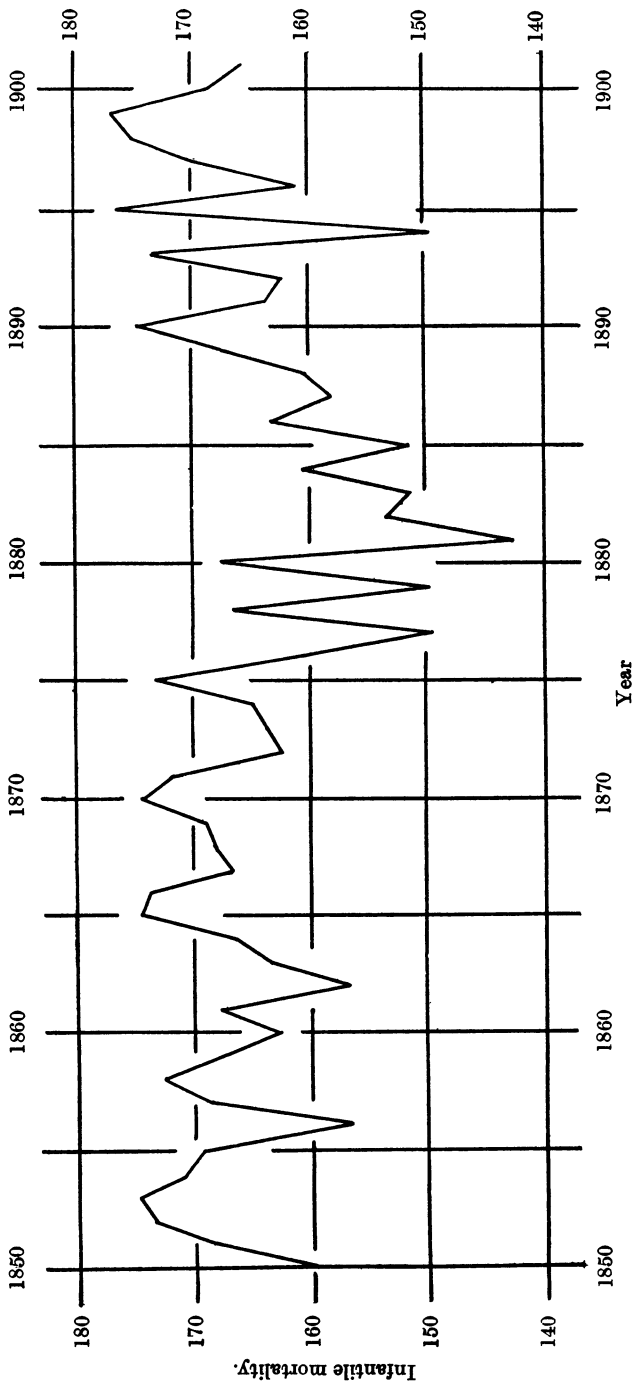


FIG. 3.

Diagram showing the course of infantile mortality (deaths per thousand births in the same year) for males in England and Wales, 1850–1900.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

Using only the years 1850–1901 inclusive, so as to avoid the period of rapid fall, and their differences, I got the following results :—

TABLE IV.

—	Amplitude of two-year term.	Factor.
Data....	− 0·374	....
First differences	+ 0·647	1·73
Second differences	− 1·207	1·87
Third differences	+ 2·273	1·88
Fourth differences ...	− 4·326	1·90
Fifth differences	+ 8·330	1·93
Sixth differences	−16·162	1·94

The first figure in col. 2 is got by taking half the difference between the average of the figures for the even years (164·017) and the average for the odd years (164·765) and the remainder as explained above. Col. 3 shows the ratio of each term to that preceding it ; the factor never reaches 2, but appears to be gradually approaching it. The final amplitude reached is small compared with the mean deviation 191·7, and there is in fact a more conspicuous three-year term.* But these results and examination of the figures suggested to me that there was in fact some period—or oscillation—of nearly but not quite two years' duration of very much greater amplitude. Smoothing the sixth differences in groups of three so as to reduce the amplitude of the three-year oscillation, the figures showed large positive values (signs being reversed from those of the unsmoothed figures) for the *even* years about 1858–60, for the *odd* years about 1874–76, and for the *even* years again about 1890–92. This suggested a reversal of phase in each sixteen years—more or less—and hence that the year was not precisely half a period but either 15/32 or 17/32 of a period—with annual values alone available one cannot tell which. Analysis of this period† gave an amplitude, not of a mere 16 or 20 points, but of roundly 265 points, utilising for the analysis the differences for the years 1850 to 1898 inclusive. Table V shows the actual values of the sixth differences, together with this periodic term and the residuals, rounded off in each case to the nearest unit. The residuals are still large, but the standard-deviation has been reduced from 255 for the differences to 178 for the residuals, the square of the standard-deviation having been reduced to less than half of its original value. The amplitude-factor for six

* *Cf.* Table VI below.

† See note in Appendix at the end of the Paper.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

TABLE V.—*Showing the sixth differences of Infantile Mortality (Males) in England and Wales, together with the periodic term in which the year is* 15/32 *(or* 17/32*) of a period, and the residuals. Figures rounded to the nearest unit: 3 decimal places retained in differencing and fitting.*

Year.	Δ^6	Periodic term.	Residual.	Year.	Δ^6	Periodic term.	Residual.
1850	− 44	+112	−156	1875	−485	−257	−228
1851	+ 98	− 64	+162	1876	+569	+264	+305
1852	−182	+ 12	−194	1877	−578	−261	−317
1853 ...	+210	+ 40	+170	1878	+477	+249	+228
1854	−114	− 90	− 24	1879 ...	−323	−227	− 96
1855	+ 2	+137	−135	1880	+247	+195	+ 52
1856	− 17	−178	+161	1881	−269	−157	−112
1857	+124	+213	− 89	1882	+275	+112	+163
1858	−206	−239	+ 33	1883	−175	− 64	−111
1859	+196	+257	− 61	1884	+ 7	+ 12	− 5
1860	−134	−264	+130	1885	+ 80	+ 40	+ 40
1861	+ 89	+261	−172	1886	− 32	− 90	+ 58
1862 ...	− 43	−249	+206	1887	− 19	+137	−156
1863	− 28	+227	−255	1888 ...	− 57	−178	+121
1864	+ 68	−195	+263	1889	+262	+213	+ 49
1865	− 64	+157	−221	1890	−505	−239	−266
1866	+ 58	−112	+170	1891	+647	+257	+390
1867	− 25	+ 64	− 89	1892	−591	−264	−327
1868	− 51	− 12	− 39	1893	+380	+261	+119
1869	+ 93	− 40	+133	1894	−162	−249	+ 87
1870	− 99	+ 90	−189	1895	+ 64	+227	−163
1871	+138	−137	+275	1896	− 82	−195	+113
1872	−133	+178	−311	1897	+143	+157	− 14
1873	− 41	−213	+172	1898	−173	−112	− 61
1874	+303	+239	+ 64	1899	+ 71	+ 64	+ 7

differencings with this period I make 62·17, so that in the original data we should expect an amplitude of something like 4·26 points. Applying the same process to the data, using the years corresponding to those employed in the differences, viz., 1850 to 1904 inclusive, I find an amplitude 3·80, some 11 per cent. too low. The phase for this term in the data is 54°·5, and the shift of phase is 31/64 of a period at each differencing, or 326°·75 as the effective alteration of phase over six differencings. This would make the phase for the corresponding term in the sixth differences 20°·75. Actually it is 25°·1. There is therefore a rough correspondence in phase and amplitude. I should hesitate to say whether or no this period (or something near it—no attempt has been made to determine the period with maximum amplitude) is in any proper sense of the term a real one—very likely it is not—and for present purposes the question is immaterial; the point is that it exists in the data before

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

us, and in the sixth differences its amplitude is so magnified that it must exercise something like a dominant influence.

I applied periodogram analysis to the data and the sixth differences to determine the amplitudes of the three-year, four-year and five-year oscillations as well, and give the results in the following table, including the amplitudes already stated for completeness.* The amplitudes for the data and for the differences cannot be taken as strictly comparable, as all the amplitudes for the data were determined on the years 1850 to 1901, to avoid the period of fall, while forty-eight years of differences were used for the three and four-year periods, and fifty for the five-year period. The results are sufficiently comparable, however, for my present purpose.

TABLE VI.

Period in years.	Amplitude: data.	Amplitude Δ^6.
32/15 or 32/17	3·80	265
2	·37	16·2
3	1·51	48·9
4	1·45	18·0
5	1·48	5·0

In the data the special period shows an amplitude some two-and-a-half times that of the periods of three, four and five years, the apparent amplitudes of which are all about equal. In the differences the amplitude of the special short period is magnified up to overwhelming dimensions; the amplitude of the three-year term is only about one-sixth as large, but in its turn has been rendered nearly ten times as large as the five-year term and nearly three times as large as the four-year term.

The "variate-difference correlation method" so far as appears from the present investigation tends accordingly to stress, not merely the fluctuations that are random with respect to time as stated by the original authors, but the oscillations of two years' duration.† Further, I think it may be emphasized (as the point is obscured by the preceding analysis) that there is no need for oscillations which are so nearly periodic as to show an appreciable amplitude when the data are analysed as a whole. Suppose that a very long period of time is available, that the movement is usually

* Elimination of the special period before analysis would, however, seriously affect the apparent two-year period but not, I think, the others.

† The conclusion would seem to vitiate the reasoning of ref. 10, which has been already challenged by Dr. Brownlee and Mr. Trachtenberg (*Journal*, Vol. 80).

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

steady, but that from time to time some disturbance occurs which starts a two-year oscillation, expanding for a few years and then again subsiding, thus giving rise to a patch of years (such as the infantile mortality exhibits in 1876–87) with a well-marked alternation. Suppose such patches of alternation to occur at random; then over a long period of time the average amplitude of the two-year oscillation will appear to be zero. But all the conditions are present for rendering such two-year oscillations the dominant factor in correlations obtained by the difference method. Regularity is not necessary. How dominating the influence of the alternations in infantile mortality round about 1880, and again round about 1895 must be, is clear from a glance at Table V; and if another century of data were available, the random occurrence of such alternations from time to time—without respect to their relative phases—would continue to render alternation the dominating factor. Some work in the next section (p. 521) throws a little light on the possibility of measuring such a generalized tendency to alternation in a given series.

In view of these conclusions it is necessary to give some further study to the influence of differencing on the *random* series. In particular we have to answer the question whether differencing tends to lay *more* stress on random fluctuations or on alternations.

III. *The differences of a random series.*

If any series such as the series $a\ b\ c\ d\ \ldots$ below is differenced, the coefficients of the terms in the differences of order k are given by the binomial coefficients in the expansion of $(1-1)^k$:—

a	$b-a$	$c-2b+a$	$d-3c+3b-a$
b	$c-b$	$d-2c+b$	$e-3d+3c-b$
c	$d-c$	$e-2d+c$	
d	$e-d$		
e			

Now suppose the series to be a *random series*,* in the sense that the rth term u_r is uncorrelated with the $(r+s)$th term u_{r+s} for all values of s, and further suppose that the series is so long as to render the end terms negligible compared with the central portion of the series. This assumption is necessary since in the first differences the first and last values of the function occur only once, while all other values occur twice; in the second differences the first and last values occur again only once, the second and pen-

* I use this term in preference to Anderson's phrase "oscillatory series" (oscillatorische Reihe), since the word oscillatory suggests some sort of quasi-periodicity and not randomness at all.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

ultimate twice, and all others thrice—and so on. On these assumptions if s_o be the standard-deviation of the function, s_k the standard deviation of the kth differences—

$$s^2_k = F(k)\, s_o^2$$

where F(k) is the sum of the squares of the binomial coefficients of order k, and by a known theorem* is equal to $(2k!)/(k!)(k!)$. This is one of the fundamental theorems in the work of Anderson (ref. 8). The values of F(k) increase very rapidly with k as shown in Table VII below :—

TABLE VII.

Order of the differences k.	F (k) Ratio of the square of s.d. of kth differences to square of s.d. of the data.	$\sqrt{F(k)}$ Ratio of the s.d. of kth differences to the s.d. of the data.
1	2	1·41
2	6	2·45
3	20	4·47
4	70	8·37
5	252	15·87
6	924	30·40

For fourth differences F(k) attains the value 70 ; that is to say, the square of the standard-deviation of fourth differences of a random series is (for a long series) some seventy times the square of the standard-deviation of the original series. The ratio of the standard deviations themselves is therefore given by the square root of this number, or 8·37 as shown in the last column of Table VII. When we proceed as far as sixth differences F(k) is brought up to 924 and its square root to 30·40—that is to say, the standard-deviation of the sixth differences of a random series is over thirty times the standard-deviation of the original series, given that the series is sufficiently long to enable us to disregard the effect of the end terms compared with the central portion.

As an illustration I built up a random series in the following way. Two packs of cards from which the court cards had been removed were shuffled together, a card drawn and noted. The pack was then shuffled again, another card drawn and noted, and so on, each card being returned to the pack after being noted. Tens were entered as zeros, and red cards as negative. The drawing was continued till 120 cards had been noted. In this way a random series was obtained, the expected frequency distribution ranging between the limits ± 9, and the expected frequency of every digit

* *Cf.* Chrystal's *Algebra*, Part II, p. 18.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

being the same with the exception of that of the zeros, which were twice as frequent as each of the other digits. Differences were taken up to the fourth and the standard-deviations worked out. The results were as follows :—

TABLE VIII.—*Standard-deviations of an experimental random series.*

Order of differences.	Standard-deviation.		
	Observed.	Calculated from s_0 observed	Calculated directly.
0	5·14	—	5·30
1	7·70	7·25	7·47
2	13·40	12·59	12·99
3	24·48	22·98	23·69
4	45·72	43·02	44·36

The standard-deviation of the data is 5·14 as against a calculated value of 5·30, and the remaining standard-deviations are slightly high as compared with the theoretical values, whether the latter are calculated from the observed value of s_0 (5·14) or directly from the theoretical frequency distribution. But the deviations from expectation are well within the limits of fluctuations of sampling and the agreement as a whole very satisfactory, considering the moderate number of observations.

As another illustration of the general formula, take the effect of errors due to rounding-off the figures of the data to be differenced to, say, the nearest unit as in the case of the index-numbers for economic progress in Italy cited from ref. 9. Regarding such errors as uniformly distributed over the range $\pm 0{\cdot}5$, the square of their standard-deviation is 1/12, the square of the standard-deviation of sixth differences, if the errors may be regarded as random with respect to time, is therefore 924/12 or 77 and the standard-deviation itself 8·77. The smallest sixth-difference standard-deviation in the index-numbers in question (I cite from ref. 9) is that for the index-number of Stamp Duties which is 9·92. Within the limits of fluctuations of sampling this may be no greater than the s.d. of rounding off; taking the figures as they stand, however, the significant s.d. for sixth differences of the Stamp Duties index would be only (the square root of the difference of the squares or) 4·63. Any sixth-difference correlations with the Stamp Duties index would therefore be reduced at least in the ratio 4·63/9·92 or 0·467 ; and one would not be surprised therefore if none of them were significant—and in fact none of them are significant. In the other cases the resulting reductions are not nearly so large. But

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

the assumption that the errors of rounding-off are random with respect to time may have led to an under-estimation of their influence. In a slowly rising series rounding-off may obviously lead to small alternations ; *cf.* the successive differences for the years 1894–98 in Table II.

With these illustrations of the formula we can now turn to the problem that was raised at the end of the previous section : on which does the process of differencing tend to lay the more stress—random fluctuations, or alternating fluctuations, *i.e.*, fluctuations with a period of two intervals ?

The answer seems to be given at once by comparing the figures of the last column of Table VII with those in the first column of Table I. The figures are brought together below.

TABLE IX.

1. Order of difference.	2. Ratio of amplitude of two-year period to amplitude in data.	3. Ratio of s.d. of differences to s.d of data for random series.	4. Ratio of column 2 to column 3.
1	2	1·41	1·41
2	4	2·45	1·63
3	8	4·47	1·79
4	16	8·37	1·91
5	32	15·87	2·02
6	64	30·40	2·11

The amplitude of the two-year period is doubled at each differencing. The standard-deviation of the differences of the random series is multiplied by a factor which only approaches 2 in the limit : for the ratio of the square of the s.d. of the kth differences to that of the $(k-1)$th is (*cf.* ref. 9) :—

$$\frac{(2k\,!)}{(k\,!)\,(k\,!)} \times \frac{(k-1)\,!\,(k-1)\,!}{(2k-2)\,!} = 4 - \frac{2}{k}$$

and the square root of this only becomes nearly equal to 2 as k increases. Hence continued differencing tends more and more to emphasize the alternations as against the random fluctuations, as shown by the ratio of amplitude to s.d. in Col. 4 of Table IX. This ratio tends to increase indefinitely, approaching in the limit, using Stirling's approximation for the factorials in the expression for the standard-deviation, to the value $\sqrt[4]{\pi k}$; but the rate of increase after the first few differences is slow, even tenth differences only raising the ratio to 2·367. Reference to Table I shows that up to the fifth and sixth differences the amplitude of even a three-year period is raised practically with the same rapidity as the standard

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

deviation of the differences of the random series. The variate-difference correlation method *tends*, we must conclude—slowly and subject to more or less dilution by longer periodicities and by the effects of random fluctuations—to give correlations due to two-year oscillations. If, in the data, the amplitudes of longer periods greatly exceed that of the two-year period, the longer periods may, of course, still dominate even in the sixth differences, as in the case of some at least of the Italian index-numbers (*cf.* below, Table XIII).

But now consider the correlations between $\Delta^k{}_m$ and $\Delta^k{}_{m+n}$ in an indefinitely long random series. These correlations are readily worked out step by step. Thus for first differences we have, using an accent to denote a deviation from the mean—

$$S(u'_m - u'_{m-1})(u'_{m-1} - u'_{m-2}) = -(N-1)s_0^2$$

and therefore—

$$r_{m(m+1)} = \frac{-s_0^2}{2s_0^2} = -\frac{1}{2}$$

If, however, we take not adjacent first differences but those next but one to each other, *e.g.*, $u_m - u_{m-1}$, $u_{m-2} - u_{m-3}$, these differences are uncorrelated, or—

$$r_{m(m+2)} = 0.$$

The correlations between second differences can be worked out similarly, and so on, and the results up to sixth differences are summarised in Table X, where the true values of the correlations are given in the upper line of each row as fractions and the approximate decimal equivalents underneath. I have not obtained a proof

TABLE X.—*Correlations between differences of a random series.*

Order of difference.	Correlation between differences.						
	$m(m+1)$.	$m(m+2)$.	$m(m+3)$.	$m(m+4)$.	$m(m+5)$.	$m(m+6)$.	$m(m+7)$.
1	−1/2 −0·5000	0 0	0 0	0 0	0 0	0 0	0 0
2	−2/3 −0·6667	+1/6 +0·1667	0 0	0 0	0 0	0 0	0 0
3	−3/4 −0·7500	+3/10 +0·3000	−1/20 −0·0500	0 0	0 0	0 0	0 0
4	−4/5 −0·8000	+2/5 +0·4000	−4/35 −0·1143	+1/70 +0·0143	0 0	0 0	0 0
5	−5/6 −0·8333	+10/21 +0·4762	−5/28 −0·1786	+5/126 +0·0397	−1/252 −0·0040	0 0	0 0
6	−6/7 −0·8571	+15/28 +0·5357	−5/21 −0·2381	+1/14 +0·0714	−1/77 −0·0130	+1/924 +0·0011	0 0

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

of the general result, but it will be seen that the correlations for differences of order k are given by the series—

$$-\frac{k}{k+1}+\frac{k(k-1)}{(k+1)(k+2)}-\frac{k(k-1)(k-2)}{(k+1)(k+2)(k+3)}+\ldots$$

The result is interesting. The correlation starts with a high negative value between adjacent terms, and the values slowly die away with alternating signs. Differencing a random series tends therefore to produce a series in which the successive terms are alternately positive and negative. The result will be familiar to anyone who has differenced a tabular function while retaining only a few figures; after the first few differences the figures become irregular and finally the signs become erratic with a clear tendency to alternation: the errors of rounding-off here form the initial random series. In the figures given by the experiment described above, the phenomenon was very striking; even in the fourth differences there were series of alternations of sign extending in one or two instances to twelve or thirteen terms. The figures of Table II will serve as one example, though rather a short one, of the production of such alternations in a practical case. Persons (ref. 11) has also emphasized the presence of such alternations in the higher orders of difference, and shown how, in consequence, the correlations between simultaneous differences of two variables and differences 1, 2, 3 . . . steps apart tend to alternate in sign.

At the end of Section II it was stated, without special proof as the result seemed obvious, that if the variable considered were subject, not to a regular two-year oscillation, but to " patches " of such alternation, such patches would still provide the dominant factor in the variate-difference method. The conclusion is confirmed by the present result. For if the variable-series be supposed to contain as one component such an erratic tendency to alternation as exists in the kth differences of a random series—and this is very much the sort of tendency that I had in mind—the amplitude of that component, as measured by its standard-deviation, will be multiplied on a first differencing by—

$$\sqrt{4-\frac{2}{k+1}}$$

which will approach 2 the more nearly the larger k, *i.e.*, the more marked the tendency to alternation in question. But the standard-deviation of any random component will only be multiplied on the first differencing by $\sqrt{2}$ or 1·414 . . . It is the alternations, not the random fluctuations that are most emphasized.

A comparison of the correlations between the mth and $(m+n)$th sixth differences of the infantile mortality series, with the corre-

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

sponding correlations for the sixth differences of a random series is illuminating as regards the tendency to alternation existing in the original data. It will be seen from Table XI that the correlations markedly exceed the values that would be shown by the sixth differences of a random series, and more nearly approach the values that would be given by, say, the 11th differences of such a series.

TABLE XI.—*Correlations between the mth and (m + n)th sixth-differences of the data for infantile mortality: values determined from the rounded figures given in Table V.**

n.	Correlation.	Theoretical values for a random series in order of differences.			
		6.	10.	11.	12.
1	−·906	−·857	−·909	−·916	−·923
2	+·678	+·536	+·682	+·705	+·725
3	−·469	−·238	−·420	−·453	−·484
4	+·288	+·071	+·210	+·242	+·272
5	−·207	−·013	−·084	−·106	−·128
6	+·137	+·001	+·026	+·037	+·050
7	−·027	0	−·006	−·010	−·016

* I give no probable errors in this or the following table. In view of the work of Section II I do not at the moment even see how to approach the determination of probable errors of difference-correlations.

The infantile mortality *data* exhibit, I think we may say, a component resembling—roughly—in its tendency to alternation the fifth *differences* of a random series.

The number of years available in the case of the Italian index-numbers that afforded the other example are too few (twenty-eight years of data, twenty-two years of sixth differences) to render any extended comparison worth the labour, but the following values for the correlations between adjacent sixth differences suggest—as indeed does a glance at graphs of the data—that we are dealing with series of a different type :—

TABLE XII.—*Correlations between adjacent sixth differences of index-numbers for economic progress of Italy.*

Index-number.	Correlation.
1. Railways	−·870
2. Shipping	−·689
3. Revenue	−·637
4. International commerce	−·696
5. Posts and telegraphs	−·599
6. Stamp duties	−·424
7. Savings banks	−·738
8. Coal	−·843
9. Tobacco	−·669
10. Coffee	−·895

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

Here only two of the correlations exceed the theoretical value for a random series (0·857), all the rest being lower, and in some cases considerably lower. Time has not allowed me to make a complete investigation of the problem, but I have little doubt that the reason is to be sought in the much greater importance, for these data, of the longer periods. For a simple alternation or period of two years the correlation between adjacent terms is −1: for a three-year period −0·5: for a four-year period zero: for a five-year period +0·309 (provided that the series in each case may be regarded as indefinitely long) and the values for longer periods continuously increase. Hence the presence of longer periodicities will always tend to reduce the correlation between adjacent terms. The following table contrasts the amplitudes of the periods up to five years for the two cases above in which the correlation is highest, with the two cases in which it is lowest.

TABLE XIII.

Sixth differences in:	Amplitude of period: years.			
	2.	3.	4.	5.
Coffee	12·1	11·6	8·9	5·5
Railways	17·3	13·4	6·9	2·9
Posts and telegraphs	3·3	20·6	5·7	5·3
Stamp duties	4·1	4·3	1·4	5·5

With the warning given by the case of infantile mortality before us, it is necessary to remember that there may be a period in the neighbourhood of two years with a much greater amplitude than is suggested by the analysis for integral numbers of years, but taking the figures as they stand the amplitudes in the case of the first two index-numbers continuously converge. The figures for the last two numbers are very different. Both show an amplitude for the five-year period greater than that for the two-year period, and in the case of the last greater than that of both the two and the three-year periods. Only four of the sixth-difference amplitudes in Table XIII, it may be noted, correspond to amplitudes exceeding a unit in the data. These are the four-year amplitudes for coffee, and the five-year amplitudes for coffee, posts and stamp duties, and in these cases the amplitudes in the data (using the factors of Table I) would be only 1·1, 1·5, 1·5, 1·5. If any oscillations in the data are significant they are those of four or five years' duration rather than the two-year oscillations emphasized by differencing.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

IV. *Conclusion.*

The problem of time-correlation is then, in my view, the problem of isolating, for the purpose of discussing the relations between them, oscillations of different durations—such oscillations being, in all probability, not strictly periodic but up-and-down movements of greater or less rapidity. "Le plupart des statistiques," as Monsieur March remarks, "sont de date trop récente pour que l'on ait à "s'occuper des changements séculaires."

The problem is not that of isolating uncorrelated residuals.

The variate-difference method does not tend to isolate nor to lay most stress on such uncorrelated residuals. It tends to stress preponderantly oscillations with a duration of two years, the actual weight of oscillations of two, three, four, five . . . year durations in the *k*th differences naturally depending, however, on their relative amplitudes in the data.

In so far as the problem consists in finding the relations between such shorter oscillations, eliminating or reducing the effects of others so far as may be possible, the variate-difference method may possibly be of service on appropriate data in which the short oscillations are significant. The work already done by the method requires re-interpretation, however, in the light of the present discussion.

Speaking generally, and tentatively, I am inclined to think that Mr. Hooker's first method—the isolation of oscillations from the "trend"—is the better method and leads to the more readily interpretable results. Subsequent writers have varied the method of determining the "trend," and on this point, no doubt, improvement may be possible. Persons (ref. 11) discusses the point, with illustrations. The method, as it seems to me, isolates or may isolate (subject to the use of suitable processes for determining the trend) the oscillations with relatively little distortion and—no mean advantage—they can be exhibited graphically, so that investigator and reader can see what are actually the movements considered The variate-difference method distorts the actual oscillations, altering the various harmonic components in amplitude and phase.

In no case, I would suggest, is it of service to deal with movements which are not sufficiently conspicuous to call for remark in a graph of the data. Had this test been applied to the Italian index-numbers, for example, I cannot help thinking some would have been rejected as material unsuitable for illustrating the correlations between oscillations of any kind, as such oscillations are hardly visible. Two at least of the curves (savings, tobacco) are so smooth that they can be fitted with a three-difference series showing errors of less than a unit in sixteen of the twenty-eight years.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

APPENDIX.

In the work on male infantile mortality it was required to fit a harmonic series when the number of years to the period was not integral, and the number of years used for fitting covered a non-integral number of periods. When the period is an integral number of years the fitting is done by what is virtually the method of least squares, and the same method was used in the general case; the only consequence is that the work is a little more complex. Let $u_0, u_1, u_2 \ldots$ be the series to be fitted; $f_0, f_1, f_2 \ldots$ the sines; $g_0, g_1, g_2 \ldots$ the cosines. Required, to determine a, b and c so that—

$$\mathrm{S}\,(u - \overline{a + bf + cg})^2$$

shall be a minimum. The equations are—

$$\mathrm{S}\,(u) = N\,a + b\,\mathrm{S}\,(f) + c\,\mathrm{S}\,(g)$$
$$\mathrm{S}\,(uf) = a\,\mathrm{S}\,(f) + b\,\mathrm{S}\,(f^2) + c\,\mathrm{S}\,(fg)$$
$$\mathrm{S}\,(ug) = a\,\mathrm{S}\,(g) + b\,\mathrm{S}\,(fg) + c\,\mathrm{S}\,(g^2)$$

For fitting the 15/32 term in the sixth differences 49 years were used, and S(g) and S(fg) vanish, so that the equations simplify But with the fifty-five years used for fitting to the data the general equations must be employed.

I should like, on the general question of fitting a harmonic series or determining trial periods, to enter a warning against a suggested simplification of the periodogram method given in the very useful *Edinburgh Mathematical Tract*, No. 4 (A course in Fourier's Analysis and Periodogram Analysis). On p. 34 it is stated that, instead of finding the true value of the amplitude c of a given trial period (not the c of the above, but the quantity corresponding to $\sqrt{b^2 + c^2}$) in the periodogram method, we can save much of the labour involved "by taking instead of c the difference between the maximum and "the minimum in the series of numbers obtained" by averaging the columns. "This oscillation is, of course, twice the value of c "if the series of numbers represents a purely harmonic variation, "while if this is not the case it is a sufficiently accurate measure of "the importance of the trial period in question." The process may, in my very limited experience, lead to quite misleading results. Not only does the four-year grouping contain the two-year as well as the four-year amplitude, the six-year grouping contain both the two-year and the three-year as well as the six-year amplitude, and so on, but there may also be casual fluctuations. In one of my cases, for example, the difference between the maximum and minimum in the column-averages for determining a six-year period was 124·8, which would correspond to an amplitude of 62·4 on the above

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

approximation. The true amplitude was only 3·14. In another case the difference between the maximum and the minimum in the column-averages for determining a five-year period was 222·8, which would correspond to an amplitude of 111·4 on the above approximation. The true amplitude was only 4·67.

REFERENCES.

This list only contains the principal references bearing on method ; other references are given in footnotes.

1. Poynting, J. H. A comparison of the fluctuations in the price of wheat and in the cotton and silk imports into Great Britain. *J.S.S.*, Vol. 47, 1884, p. 34.
2. Hooker, R. H. On the correlation of the marriage-rate with trade. *J.S.S.*, Vol. 64, 1901, p. 485.
3. Hooker, R. H. The suspension of the Berlin Produce Exchange and its effect upon corn-prices. *J.S.S.*, Vol. 64, 1901, p. 574.
4. Cave-Browne-Cave, F. E. On the influence of the time factor in the correlation between the barometric heights at stations more than 1,000 miles apart. *Proc. Roy. Soc.*, A, Vol. 74, 1904, p. 403.
5. Hooker, R. H. On the correlation of successive observations illustrated by corn-prices. *J.S.S.*, Vol. 68, 1905, p. 696.
6. March, L. Comparaison numérique de courbes statistiques. *Jl. de la Société de Statistique de Paris*, 1905, pp. 255 and 306.
7. "Student." The elimination of spurious correlation due to position in time or space. *Biometrika*, Vol. X, 1914, p. 179.
8. Anderson, O. Nochmals über "The elimination of spurious correlation due to position in time or space." *Biometrika*, Vol. X, 1914, p. 269.
9. Cave, Beatrice M., and Pearson, Karl. Numerical illustrations of the variate-difference correlation method. *Biometrika*, Vol. X, 1914, p. 340.
10. Elderton, Ethel, and Pearson, Karl. On further evidence of natural selection in man. *Biometrika*, Vol. X, 1915, p. 488. (A paper using the variate difference method.)
11. Persons, Warren M. On the variate difference correlation method and curve-fitting. *Publications of the American Statistical Association*, Vol. XV, 1917, p. 602.

DISCUSSION ON MR. YULE'S PAPER.

PROFESSOR EDGEWORTH proposed a vote of thanks for a Paper which, he said, would be of great use to statisticians in the performance of their principal task, the treatment of averages. The case of correlated averages which had been presented reminded him of a scene which combined in one view the sea and a fresh water lake. The two surfaces were similarly affected by a rising or falling wind. There was a correlation between the wavelets of the sea and the

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

ripples of the lake. But there was no correlation between the tides of the sea and the level of the lake. In such a case to abstract the tidal oscillation so as to exhibit the correspondence between the minor fluctuations required great skill. Mr. Yule had ably reviewed the different methods which had been proposed for that purpose. He (Professor Edgeworth) hesitated to prefer Professor Poynting's method before that of Mr. Hooker. With regard to the comparison of series consisting of terms that varied with the time he felt, like Mr. Yule, unable to see what there was to compare when all the effects of time had been removed by continued differencing. But he would not venture to pass judgment on the method of differencing until he knew what its eminent advocates had to say in its defence. What the method of differencing might do had been well shown in the Paper. Its use was not to eliminate the influence of time, but to emphasize the terms relating to a period of two years (or other intervals). He liked Mr. Yule's use of "oscillations" instead of "periods," with reference to movements which were only "quasi-periodic." Adapting his methods to the irregularity of the phenomena, Mr. Yule was not open to the sarcasm which a distinguished statistician had directed against his mathematical colleagues, that they shot at sparrows with a cannon.

Dr. BROWNLEE: I have very much pleasure in seconding the vote of thanks to Mr. Yule. The method of variate differences I have already discussed before this society a number of years ago, and I felt at that time, and still feel, a very profound suspicion of any deductions based on that method. With regard to a large number of methods by which correlation is established there is room for grave criticism. Many correlation coefficients cannot be interpreted except in the light of the mechanism by which correlations arise, and if the mechanism is known the correlation coefficient does not seem to add anything for our information. Take for instance the example of the application of the four-fold method of establishing the correlation to discover the hereditary relationship of coat colour in horses. The result obtained is that the correlation is ·53. Now in this case the method of inheritance is known. A bay or brown horse mated with a chestnut horse produce only dark offspring, while the second generation of this offspring are in a proportion of three dark horses to one light. The correlation of ·53 thus necessarily arises using the four-fold method, but this calculation of correlation is unnecessary as it does not give us even as much information as we already possess. The same difficulty besets the variate difference method which tries to eliminate time changes which are either secular or periodic. Here we are face to face with a very difficult problem. It not only turns up in time phenomena but arises in connection with trade diseases. The more unhealthy a trade is the more of every disease tends to be present. The problem here is to determine if one disease tends to take the place of another in any special trade. Whether for instance if there is more bron-

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

chitis there is less pneumonia, is a problem which presents difficulties of solution which have not yet been overcome. It is the mechanism here that we wish to get at, and if we can get at the mechanism, nothing else is necessary. Mr. Yule has taken, I think, the only way of attacking this problem, and that is to assume a mechanism and to determine what will result on that mechanism. The results he obtains are most interesting and, I think, important. It is further interesting as showing how a period can be discovered by a change in the value of the correlation coefficients. His discovery in the death-rate of young children of a period of about fifteen-sixteenths of two years is a remarkable instance of how a phenomenon, not immediately evident, may be brought out by the application of this method. The period Mr. Yule has found is the period of the epidemicity of measles in London and many of the large towns of England. This period of epidemicity in London was demonstrated first by myself in much the same way as it is found by Mr. Yule, namely, by noticing that there was a sixteen years swing backwards and forwards in the deaths from measles, the years of heavy mortality swinging from the odd years to even years backwards and forwards at intervals of about sixteen years. In thanking Mr. Yule for his Paper, I would like to suggest that he carries his investigations into this method somewhat further, as it seems capable of some development not so much in the capacity of eliminating a time element as in revealing factors which are not immediately apparent. I have to thank Mr. Yule very much for his Paper and have great pleasure in seconding the vote of thanks.

Dr. GREENWOOD: To say that this Paper is an important contribution to the theory of statistical methods would be too much in the manner of "Mr. F.'s Aunt," to be illuminating. All Fellows of the Society know the value of Mr. Yule's researches and we are all proud of the great scientific distinction by which they have recently been rewarded. The object of a discussion is not to pay compliments, however well deserved such compliments may be, but to bring out and, if possible, to elucidate obscurities or objections. I have two difficulties, one respecting the end to be reached, the other as to how it may be reached. I think that some sentences in Mr. Yule's Paper under-rate the importance of what "Student" and those who have developed his method seek to do. Let me suppose, purely for the sake of argument, that the death-rate in any country is substantially modified by two factors and that these factors operate to a greater or smaller degree upon all the constituent items of the death-rate. Let the first be a more or less regularly progressive improvement of the general conditions of life, the second, quasi-periodic oscillations, perhaps consequences of the greater or lesser activities of public health services under the pressure of recurring waves of opinion, political or social. By hypothesis, there will be a general downward trend of the death-rate and more or less considerable quasi-periodic fluctuations imposed upon this

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

trend. Both these phenomena are intrinsically important, but if we seek to learn whether the fatalities of two diseases, A and B, are more intimately related one to another than those of C and D or of A or B with C or D, *admitting that all four are influenced by the two changes mentioned,* we *do* desire to eliminate both secular and oscillatory effects, and, if those effects can be eliminated, what remains is surely of the greatest importance. Indeed, so far as the statistical study of ætiology is concerned—the study which is, to me, of the greatest interest—I should be disposed to leave the "not" out of Mr. Yule's dictum and to say that the "problem *is* "to isolate random residuals." "Student's" sentence does not at all leave me "wondering with what sort of movements he *does* desire "to deal," for I think I perceive that he desires to deal with precisely the sort of movements which are of ætiological importance. For these reasons I rate "Student's" attempt higher than does Mr. Yule. Passing now from the criticism of the objective to the criticism of the manner of attaining that objective, I have the following points. Professor Pearson and Miss Elderton had already recognized that the existence of a harmonic term of short period would prevent the simple application of the variate difference method from leading to the desired end; Mr. Trachtenberg has provided us with important examples. What is, to me at least, novel is Mr. Yule's evidence that a short period of amplitude too small to be easily recognized in the original data may ultimately dominate the results. Upon that, I have this question to ask. The approximate values of the standard deviations of the correlations of successive orders of differences increase, as Dr. Anderson showed, as the square roots of natural numbers; if there be a single short-period quasi-harmonic oscillation too small to catch the eye in the plotting of the original data, is the ultimate correlation of differences dependent upon it likely to mislead the worker who keeps account of the errors of sampling? In ordinary practice is not the method likely to be used as a basis of inferences only when difference correlations of the third and higher orders are sensibly equal? The next point is whether the admitted difficulties of the variate difference line of advance can be escaped by following another line of approach to the "random residuals." Herr Frederik Esscher of Lund in a recent memoir* argues that the way of escape is to correlate the differences between the observed values and the ordinates of a parabola—a plan previously adopted by Professor Pearson and Miss Elderton. Esscher's evidence is that the residual correlations are steadier than those of successive orders of differences and he illustrates his case with the following examples. The data were the probabilities of dying deduced from the vital Statistics of Sweden 1886–1914. For the correlation between the values for males and females the following results were obtained.

* *Ueber die Sterblichkeit in Schweden* 1886–1914, No. 23 of Serie II. *Meddelanden från Lunds Astronomiska Observatorium,* Lund, 1920.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

Age 0–1.		Age 4–5.	
·991	·991	·974	·974
·952	·945	·808	·693
·938	·957	·808	·504
·937	·963	·822	·406
·928	·967	·824	·324
	·972		·268
	·975		·332

The first column in each table gives (excepting the first lines, which contain the correlations of the original data) the correlation of residuals after subtraction of the ordinate of the parabola (fitted by Tchebycheff's method), the second column the successive difference-correlations. Esscher writes :—"The reason for this is to be "sought in the fact that the time free residuals are not, as the method "assumes, independent one of another. Admittedly the method of "least squares also assumes mutual independence of the time free "residuals, but this condition needs only to be fulfilled upon the "average and, even were it not, the method of least squares always "provides the curve which makes the sum of the squares of deviations "from it a minimum. It appears on the other hand very difficult "to give to the method of variate differences an intelligible basis "under such circumstances." I do not think this argument entirely convincing. The assumption is that because the correlation between the residuals after application of an nth order Tchebycheff parabola to each variate is sensibly equal to that obtained from an $n-n$th order parabola, that *therefore* the correlation obtained is the true time-free effect. It might be objected that the basis of the constancy is simply that, over the given range, there may be hardly any difference between the form of the two curves (for n small) so that the calculations are made upon sensibly the same arithmetical numbers. For instance, the death rates from all causes of males aged 45–55 in each successive quinquennium from 1841–45 to 1901–05 (inclusive) England and Wales are per 1,000 :—17·2, 19·2, 18·6, 17·5, 18·9, 19·6, 20·3, 19·8, 19·3, 19·4, 19·5, 18·4, 17·0. When these figures are smoothed by Tchebycheff parabolas of second and third orders, *i.e.*, an ordinary parabola and a cubical parabola, only 4 of the corresponding ordinates differ, viz. : corresponding to 17·2 in the data, the two parabolas give respectively 17·4, 17·5 ; corresponding to 17·5, 19·1 and 19·0 ; corresponding to 19·4, 19·3 and 19·4 ; corresponding to 17·0, 17·7 and 17·6. In fact the argument might be reversed, it might be urged that the inconstancy of the successive difference correlations is giving a warning which only the application of a very high order parabola and correlation of residuals would give, to the effect that the elimination of the trend and of quasi-periodic oscillations is not being successfully achieved, that the correlations are not "organically" trustworthy. It must be remembered that

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

the representation of secular trend by a parabola of any order is only a matter of interpolatory convenience. To suppose that the fitted curve obtained by least squares represents a secular trend in any wider sense is an illusion which extrapolation will at once dispel. This method of assuming that one particular parabolic function, that fitted by least squares to the range, is the right function, is perhaps no less dangerous an assumption than those stigmatised by Esscher in the variate difference method. In conclusion I would repeat that the aim of " Student's " investigation is, as I think, a very important one, the criticisms of the method which we owe to Professor Pearson, to Dr. Brownlee, to Dr. Esscher and now to Mr. Yule are weighty, and if anyone supposed that the rule-of-thumb application of the variate-difference method would suffice to solve intricate problems, he knows better now. But I am still not convinced that, applied with discretion, " Student's " method cannot answer some of the questions which he wished to answer.

Sir JOSIAH STAMP said he remembered in 1917 when he was working on a Paper which was given in 1918 on the relation between trade fluctuations, that is, the fluctuations in the prices and volume, and the fluctuations in profits, he was going along with four separate methods ; the ordinary moving average which he was using in a nine-year period, and—what had just been condemned by Dr. Greenwood —the secular trend, which he called the linear trend, obtained by the use of the method of the least squares. He thought it was the first time it was used in this country for that class of statistics, and therefore he was diffident about the results. He was also using the first and second variate-differences when Professor Pearson's Paper of June, 1917, fell like a bombshell into his quiet studies. He was gravely disconcerted to find these curious alternations of sign in the high differences, and felt at first that the whole thing was a shifting sand, and one could not trust the method at all. It was just as a pupil feels when he finds his teacher out in some glaring error. He felt that he had been decoyed into believing there was something valuable in the variate-difference method ; but on reflection he went on with it because it seemed to him it had certain uses, and he was not quite certain whether Mr. Yule had stressed sufficiently in his Paper the possibility that it is of value along the line that occurred to him at the time. [Sir Josiah quoted the note at the beginning of his own Paper on the variate-methods showing what the variate-difference method involves.] The reason why he went on with it—it might be an absolute mare's nest—was this : he felt he had to employ every method he could to get rid of this terrible difficulty of the time element when he was trying to correlate annual differences, because there were so many " faults " in the statistics. He rather fancied Mr. Yule had looked upon the statistics as being in all cases a pretty perfect series. When one came to deal with statistics of prices or

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

railway tonnage, bank clearing house figures, and all kinds of figures of that sort, one could never make that assumption. He had to be prepared for two kinds of critics when he took out his linear trend. He had to be prepared for the critic who said "But your "bank statistics are no good at all; we have had a constant process "of amalgamation, and all your clearing house figures are useless." He felt that every kind of what one might call systematic or natural alteration in the figures that had a reasonable chance of being regular, might be regarded as falling into the same kind or order of changes as the time-change itself, and that the linear trend method should be a sufficient answer in order to remove any systematic change in the trend or character of the figures, as well as the common cause of growth of population and wealth. But there was the possibility that some other person would come along and say "It "is all very well; you have used a lot of railway tonnage statistics, "but I am in the office where they made them up, and in the course "of such a year we had a complete change." What he had to meet there was the possibility, in every one of these series of statistics, that there was some unknown break in one year in the method in which they were made up, and that that would ruin the application of the secular or linear trend; for whereas at the point of bringing it out from a break there should be two parallel trends, and any single trend would be quite artificial, he felt that the variate-differences, the first and second, would enable him to say that out of say forty observations there was only *one* sinner, and he did not turn the whole lot into an absolutely vicious series. It was true his error would to a small extent affect the coefficient, but *only* to a small extent; at any rate thirty-nine of the differences were right, and only the fortieth was wrong. But he hoped by the use of the two methods he might counter the two criticisms. Whether he was wrong or right in that, Mr. Yule might give some idea. It seemed also that if the variate-difference method had this tendency to show the two-year oscillation as over-emphasized, that might have some virtue in this class of statistics. He was afraid, from the observation he made about those Italian figures, the Mortara statistics, Mr. Yule would condemn practically every kind of mathematical gymnastic done with trade and similar classes of figures as not being sufficiently perfect in their form for the use of these particular exercises. But he had wondered, supposing the particular yearly figures of profits being made up from assets running to particular dates were what one might call rather "fluffy" and not clear cut, and the averages of two years round any given point were probably truer than any single year by itself, whether this fact of itself might not give the variate-difference method some virtue.

Mr. H. C. TRACHTENBERG said he agreed with Dr. Greenwood that it was too much to say that "Student" was misleading in his definition of what he set out to do. He thought what he meant was that if one used the coefficient as it stood, one would not know

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

how to interpret the result. Therefore he set out to separate the components, and in this he was successful. With regard to the point Dr. Greenwood had brought up as to the difficulty of countering chance fluctuations, he had thought of that difficulty himself, and had hit, he thought, on the solution. Mr. Yule had paid so much attention to the single series and its differences in isolating certain qualities, that he only reminded them from time to time that the real problem he was dealing with was correlation. For instance, if they were correlating two quantities, and they found there was a two-year period in one of them, it was very unlikely on the lines of chance that there would be a two-year period in the other as well. When they correlated two periodic series, unless the two periods were identical, the correlation would not be unity but zero, and therefore the correlation would not be touched by that fact. It would simply dismiss the chance periods, except in the very exceptional case when chance gave a coincidence of the two-year period, say, in each service; and there was always the Schuster test to fall back upon, which would show in each case whether the period was a real one or not. With regard to the type of things which the variate-difference method safeguarded them against, Mr. Yule had called attention to three types, but he wished to mention some others. The types Mr. Yule had mentioned were, firstly, the secular trend—and he did not think Dr. Greenwood's plaint was to be taken so seriously as he took it, because he thought when one referred to the secular trend, one meant such a cumulative process as continued throughout the series one was examining, without regard to the effect of extrapolation. The variate-difference method isolated that factor. Another factor it isolated was the oscillation type, and, finally, Mr. Yule had mentioned the random series type. He (the speaker) desired to point out that it isolated oscillations of different periods as well, because it produced a zero correlation. But the particularly interesting point that had not been mentioned in detail was the case where they had, not a random series in one case, and another independent random series in the other, which gave a zero correlation, but when, as in the case he had treated, they had a random series arising in one case, and in the other case they got that random series repeated, *e.g.*, in the case of the same disease striking different ages. In that case it could lead to fallacious results if not interpreted correctly. Professor Pearson had drawn a conclusion as to natural selection out of a figure which simply arose out of the pure mathematics of the calculation. In the same way he was puzzled by Mr. Yule's reference to infantile mortality, in which he said that the sixth order of difference was of a character inherent in the original statistics. But surely that difference was simply due to the mathematics, coming down from ·906 and ·678 to minus ·027. Where they got a difference which was outside the scope of the sixth differences they got an unimportant negative correlation. Mr. Yule said, taking the correlation between the adjacent sixth differences, it was minus ·906, whereas with the

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

sixth differences of the random series it would be only minus ·07. They had to go on to the tenth or eleventh differences to get a correlation as big as the others ; that was five orders of difference higher. So that they started off with the high orders of co-ordination. Mr. Trachtenberg said he was sorry he had looked at the wrong column, but he would like to emphasize his point that the difficulty of the chance period was one which did not affect the correlation.

Mr. R. A. FISHER said there was one difficulty which he had met with in his practical experience with the variate-difference correlation method which Mr. Yule had not mentioned, and which might perhaps throw some light on the remarks which Dr. Greenwood had made, apparently in its defence. That was that the variate-difference correlation method assumed that if they had a series with terms x_1, x_2, etc., and a second series y_1,y_2, etc., the only correlation between those series was the correlation between corresponding terms, that was to say, between x_1 and y_1, and so on ; there were no lagging correlations. Suppose they were birth-rates and marriage-rates, for example, that was an instance where they would expect a lagging correlation, and not only an immediate or contemporary correlation of the kind which was assumed. It was rather difficult to explain in a few words exactly what the variate-difference method led to in such cases, although mathematically it was very simple. Considering the ultimate residuals, which he agreed with Dr. Greenwood were the really valuable things that they wanted to get at, and supposing the correlation between the corresponding and contemporary residuals was r, and if shifted one year was r_1, shifted two years was r_2, and so on, you have a series of correlations running backwards and forwards to negative and positive suffixes ; if there was such a series of correlations—as you would get in the case of marriage-rates and birth-rates—what the variate-difference method gave them was a high order difference of that r series. If they correlated first differences, they would get $r_o - \frac{1}{2}(r_1 + r_1)$, that is to say one half the second difference of the r series. If they correlated sixth differences, they would get $\frac{1}{924}$ of the 12th difference of the r series. Those expressions written out in full reduced to r_o quite satisfactorily, if all the other r's were zero. If lagging correlations were absolutely evanescent they got the true value of the contemporary correlation ; that consideration was overlooked in one of the most important applications of the variate-difference correlation method, and one which was being discussed that evening, that was the question of the selective effect of infantile and child death-rates in successive years. In that case the death-rate in one year was correlated by Professor Pearson with the death-rate of the same group of children in the next year, and he found a series of correlations in every case steadily rising up to a value — ·7 with extraordinary uniformity. As Dr. Greenwood had said, he thought that did tell them something which they ought to know, and perhaps Professor Pearson should have suspected, that there

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

was something very wrong in the method which brought out negative correlation of the same value—that was to say relative effects, as he interpreted them—between the death-rates in the 1st and 2nd years of life, as between the death-rates of the 4th and 5th year; one would have expected that the selective effect of the mortality of the first year, which was very high, would be higher than any possible selective effect of the much lower mortality of the 4th year. If they put that alongside the criticism he originally made, they would see that if they took their r_0 as the death-rates of the same group of children in adjacent years, then their r_1 would be the correlation between different groups of children in the same year, and that was what was being neglected. They had the same environment and the same meteorological and epidemiological conditions acting on two groups of children in their fourth and fifth years respectively, and any correlation which existed between those death-rates was ignored entirely by the method as applied in this particular example. He suggested that that was a very high correlation. Without going into the statistics, he could not say what value should be assigned to the correlation, but looking at them cursorily he found every indication of the correlation being very high, and one would expect them to be very high. He suggested $+\cdot 8$ as a reasonable figure, or, to be moderate, $+\cdot 7$. The figures themselves indicated, without going into a more elaborate method which he hoped to mention later, that the value was more than $+\cdot 8$. If they considered the twelfth difference of a correlation coefficient divided by 924, they had r_0 minus $\frac{6}{7}$ths of r_1 if all other lagging correlations were ignored; and if they took $r_{1\,2} + \cdot 7$ as the correlation between the death-rates of different groups of children exposed side by side to the same environment, they would have a spurious error of $-\cdot 6$ introduced into the correlation obtained by the variate-difference method, in fact the same order as total effect, $-\cdot 695$, $-\cdot 72$, and so on, which were actually obtained by that method. He suggested that there was every reason to suspect a spurious factor in those correlations, of the same order as the very high correlations themselves, and that the data as treated at present did not even show whether the correlation which they were seeking was a positive or a negative one. They did not answer the question as to whether there was or was not a selective action of the kind that was sought for. He suggested that the method mentioned, and to some extent criticised by Dr. Greenwood, would meet that difficulty, and meet it in a manner quite above suspicion. What the variate-difference correlation method attempted to do was to eliminate any time function which would be represented by a polynomial of the sixth degree. In his opinion it did that successfully. It also unfortunately entangled up the adjacent lagging correlations into expressions from which it was impossible to extricate them, but it did successfully eliminate any secular change which could be represented by the terms up to the sixth degree. If they fitted polynomials to the data and took the residuals, they also get a series of

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

values from which any time factor to the sixth degree was eliminated, and in which the residues are not entangled together, or, rather, correlations between them were not entangled together; in fact they would find actual values of the correlations between the same group of children in successive years, or between different groups of children in the same year. He suggested they would at any rate find the value of the latter as a high positive value, and that that was the correlation which had governed the actual values found by the variate-difference method.

The vote of thanks to the author of the Paper was carried unanimously.

Mr. YULE, in reply, said he thanked the Society for the way in which they had received the Paper, and he would ask them to forgive him if he did not at that moment endeavour to cover all the remarks that had been made. Much of the discussion had been very interesting, especially the remarks made by Dr. Brownlee, Dr. Greenwood, Sir Josiah Stamp, and Mr. Fisher. But they were not remarks that one could deal with readily at short notice, and it was very difficult to follow the points when merely listening to a speaker. On the general question of "random residuals"—referred to both by Dr. Greenwood and by other speakers—he must confess to a certain prejudice. He found it very difficult indeed to conceive that any annual residuals due to real, definite, assignable causes—as distinct from chance residuals of the nature of errors of observation or fluctuations of sampling—could be of the random character supposed, the residual for year n uncorrelated with the residual for year $n + m$. At any rate if such residuals did exist he would like to see them isolated and brought out for inspection. As mentioned in a line or two at the end of the Paper, he thought it of considerable importance in every case to endeavour to isolate the particular oscillations with which it was desired to deal, so that the student could see then graphically before him. He agreed with the speaker (Mr. Trachtenberg) who remarked that he did not think that the failure of any curve to extrapolate was a good test, as the sole problem of determining the "trend" was one of interpolation. There was nothing metaphysical or obscure in his own conception of the "trend"; it was simply the line given when the particular oscillations with which it was desired to deal were removed from the original data. The difficulty in determining it lay solely in the precise definition, and then in the elimination, of these oscillations. It should be noted that, on this conception, the "trend" was not necessarily purely of the type that could be described by a polynomial. If, for example, it was desired to discuss the three-year oscillations, and these were removed by a process of averaging over groups of three years, the "trend" would still contain the marked oscillations of about ten years' duration. Esscher's method, it appeared to him, wanted a good deal of consideration. The

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions

fall in the difference-correlations, as he conceived it, might simply be due to the fact that the oscillations existing in the data were compounded of oscillations of different durations due to different causes, and to the tendency of the variate-difference method to lay stress on one particular oscillation of short duration only. Again, he would suggest that any special case of the sort was difficult to discuss without isolating the oscillations so that you could see with what you were dealing. Sir Josiah Stamp had put a very interesting problem, as to the best method to be employed when the data were subject to discontinuous changes due to alterations of method. He did not feel at all competent to advise without a good deal of consideration, but would still be inclined to suggest Hooker's method. Even the difference-method would, step by step, spread the effect of such a discontinuity over a number of years. He did not agree with Mr. Trachtenberg's view, if he rightly understood it, that a period found in the one variable would be unlikely to be found in the other. As variables selected for discussion were, in any practical case, either similar or in all probability related, it appeared to him most likely that a period conspicuous in the one would be conspicuous in the other. In conclusion, he would like to point out that he had not written down the variate-difference method as of no service whatever. He conceived it to be quite possible that there might be cases where the short oscillations were the important thing, and the tendency of the method to stress such oscillations would then be of value. But it was important that it should be realized what the method did, and it was very important that they should specify more clearly than was often done what was the *precise* problem and what were the precise oscillations with which they desired to deal. He would also like to suggest, from his own experience, that it might often be of service to use differences for periodogram analysis, thus eliminating, before analysis, any part of the time-movement which could be expressed by a polynomial function. The fact that part of the movement could be expressed by a polynomial might in some cases lead to appreciably false values for the amplitudes when the periodogram method was applied to the data.

The following Candidates were elected Fellows of the Society :—

John Balch Blood.
Sir Charles Bright, F.R.S.E.
Herbert Douglas Buchanan.
Edgar Leyshon Chappell.
Lawrence W. Duffett.
Lewis Edwards.
Lewis Jones.
Godfrey Edward Mappin.
Arthur Pedoe, B.Sc.
Frederick Steadman.
F. H. G. P. Thomson.
Thomas Wharton.

This content downloaded from 185.44.79.160 on Tue, 24 Jun 2014 23:39:22 PM
All use subject to JSTOR Terms and Conditions