IV. The Elimination of Spurious Correlation due to position in Time or Space.

By "STUDENT."

In the *Journal of the Royal Statistical Society* for 1905*, p. 696, appeared a paper by 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, sometimes uniform, sometimes accelerated, which would lead to altogether misleading values were the correlation to be taken between the figures as they stand.

Since Mr Hooker published his paper, the method has been in constant use among those who have to deal statistically with economic or social problems, and helps to show whether, for example, there really *is* a close connection between the female cancer death rate and the quantity of imported apples consumed per head!

Prof. Pearson, however, has pointed out to me that the method is only valid when the connection between the variables and time is linear, and the following note is an effort to extend Mr Hooker's method so as to make it applicable in a rather more general way.

If x_1, x_2, x_3, etc., y_1, y_2, y_3, etc., be corresponding values of the variables x and y, then if x_1, x_2, x_3, etc., y_1, y_2, y_3, etc. are randomly distributed in time and space, it is easy to show that the correlation between the corresponding nth differences is the same as that between x and y.

Let ${}_nD_x$ be the nth difference.

For $${}_1D_x = x_1 - x_2, \quad \therefore {}_1D_x^2 = x_1^2 - 2x_1 x_2 + x_2^2.$$

Summing for all values and dividing by N and remembering that since x_1 and x_2 are mutually random $S(x_1, x_2) = 0$, we get†

$$\sigma^2_{{}_1D_x} = 2\sigma_x^2.$$

Again, $${}_1D_y = y_1 - y_2, \quad \therefore {}_1D_x \, {}_1D_y = x_1 y_1 - x_2 y_1 - x_1 y_2 + x_2 y_2.$$

Summing for all values and dividing by N, and remembering that x_1 and y_2 and x_2 and y_1 are mutually random

$$r_{{}_1D_x \, {}_1D_y} \sigma_{{}_1D_x} \cdot \sigma_{{}_1D_y} = 2r_{xy} \sigma_x \sigma_y,$$

$$\therefore r_{{}_1D_x \, {}_1D_y} = r_{xy}.$$

Proceeding successively $$r_{{}_nD_x \, {}_nD_y} = r_{{}_{n-1}D_x \, {}_{n-1}D_y} = \dots = r_{xy} \quad \text{.......................(1).}$$

Now suppose x_1, x_2, x_3, etc. are not random in space or time; the problems arising from correlation due to successive positions in space are exactly similar to those due to successive occurrence in time, but as they are to some extent complicated by the second dimension, it is perhaps simpler to consider correlation due to time.

Suppose then $$x_1 = X_1 + bt_1 + ct_1^2 + dt_1^3 + \text{etc.}, \quad x_2 = X_2 + bt_2 + ct_2^2 + dt_2^3 + \text{etc.}$$
where X_1, X_2, etc. are independent of time and t_1, t_2, t_3 are successive values of time, so that $t_n - t_{n-1} = T$, and suppose $y_1 = Y_1 + b't_1 + c't_1^2 + \text{etc.}$ as before.

* The method had been used by Miss Cave in *Proc. Roy. Soc.* Vol. LXXIV. pp. 407 *et seq.* that is in 1904, but being used incidentally in the course of a paper it attracted less attention than Hooker's paper which was devoted to describing the method. The papers were no doubt quite independent.

† The assumption made is that n is sufficiently large to justify the relations

$$S_1^{n-1}(x)/(n-1) = S_2^n(x)/(n-1) = S_1^n(x)/n \text{ and } S_1^{n-1}(x^2)/(n-1) = S_2^n(x^2)/(n-1) = S_1^n(x^2)/n,$$

being taken to hold.

Then $${}_1D_x = {}_1D_X - bT - cT(t_1+t_2) - dT(t_1^2+t_1t_2+t_2^2) - \text{etc.}$$
$${}_1D_x = {}_1D_X - \{bT + cT^2 + dT^3 + \text{etc.}\} - t_1\{2cT + 3dT^2 + 4eT^3 + \text{etc.}\}$$
$$- t_1^2\{3dT + 6eT^2 + \text{etc.}\} - \text{etc.}$$

In this series the coefficients of t_1, t_2, etc. are all constants and the highest power of t_1 is one lower than before, so that by repeating the process again and again we can eliminate t from the variable on the right-hand side, provided of course that the series ends at some power of t.

When this has been done, we get
$${}_nD_x = {}_nD_X + \text{a constant},$$
$${}_nD_y = {}_nD_Y + \text{a constant},$$
so $$r_{{}_nD_x\,{}_nD_y} = r_{{}_nD_X\,{}_nD_Y} = r_{XY},$$
and of course $r_{{}_{n+1}D_x\,{}_{n+1}D_y} = r_{{}_nD_x\,{}_nD_y}$, for ${}_nD_x$ and ${}_nD_y$ are now random variables independent of time.

Hence if we wish to eliminate variability due to position in time or space and to determine whether there is 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.

This process is tedious in the extreme, but that it may sometimes be necessary is illustrated by the following examples: the figures from which the first two are taken were very kindly supplied to me by Mr E. G. Peake, who had been using them in preparing his paper "The Application of the Statistical Method to the Bankers' Problem" in *The Bankers' Magazine* (July—August, 1912). The material for the next is taken from a paper in *The Journal of Agricultural Science* by Hall and Mercer, on the error of field trials, and are the yields of wheat and straw on 500 $\frac{1}{500}$ acre plots into which an acre of wheat was divided at harvest. The remainder are from the three Registrar-Generals' returns.

	I	II	III	IV	V	VI
Correlation between ... and ...	Sauerbeck's Index numbers. Bankers' Clearing House returns per head	Marriage Rate Wages	Yield of Grain Yield of Straw	Tuberculosis Death Rate. Infantile Mortality		
				Ireland	England	Scotland
Raw figures ...	−·33	−·52	+·753	+·63	+·35	+·02
First difference ...	+·51	+·67	+·590	+·75	+·69	+·51
Second difference ...	+·30	+·58	+·539	+·74	+·74	+·65
Third difference ...	+·07	+·52	+·530	—	—	—
Fourth difference ...	+·11	+·55	+·524	—	—	—
Fifth difference ...	+·05	+·58	—	—	—	—
Sixth difference ...	—	+·55	—	—	—	—
Number of cases	41 years	57 years	500 plots	42 years		

The difference between I and II is very marked, and would seem to indicate that the causal connection between index numbers and Bankers' clearing house rates is not altogether of the same kind as that between marriage rate and wages, though all four variables are commonly taken as indications of the short period trade wave. I had hoped to investigate this subject more thoroughly before publishing this note, but lack of time has made this impossible.