"Note to the Memoir by Professor Karl Pearson, F.R.S., on Spurious Correlation." By Francis Galton, F.R.S.

Received January 4,—Read February 18, 1897.

I send this note to serve as a kind of appendix to the memoir of Professor K. Pearson, believing that it may be useful in enabling others to realise the genesis of spurious correlation. It is important though rather difficult to do so, because the results arrived at in the memoir, which are of serious interest to practical statisticians, have at first sight a somewhat paradoxical appearance.

The diagrams show how a table of frequency of the various combinations of two independent and normal variables may be changed into one of A/C, B/C, where C is also an independent and normal variable in respect to its intrinsic qualities, but subjected to the condition that the same value of C is to be used as the divisor of *both* members of the same couplet of A and B. In short, that the couplets shall always be of the form A/C_n, B/C_n, and never that of A/C_n, B/C_m.

For the sake of clearness, the simplest possible suppositions, that are at the same time serviceable, will be made in regard to the particular case illustrated by the diagrams, namely, that A, B, and C, severally, are sharply divided into three, and only into three, equal grades of magnitude, distinguished as AI, AII, AIII; BI, BII, BIII; and CI, CII, CIII; also that the frequency with which these three grades occur is expressed by the three terms of the binomial $(1+1)^2$. Consequently there is one occurrence of I to two occurrences of II and to one occurrence of III. Roman and italic figures are here used to keep the distinction clear between magnitudes and frequencies. It will be easily gathered as we proceed, without the need of special explanation, that the smallness of the value of the binomial index has no influence either on the general character of the operation or on its general result.

The large figures in the outlined square, occupying the lower right hand portion of fig. 1, show the distribution of frequency of the various combinations of A and B. The scales running along the top and down the left side of the figure, which are there assigned to the values of A/C, B/C, apply to these entries also. The latter run in the same way as those in Table I below, or when quadrupled, as they will be for purposes immediately to be explained, as in Table II.

Table I.

1	2	1
2	4	2
1	2	1

Table II.

4	8	4
8	16	8
4	8	4

Let us now follow the fortunes of one of the large figures in fig. 1, say that which refers to A = I, B = III, of which the frequency is only 1. When the latter is expanded into the three possible values of the form A/C, B/C, caused by the three varieties of C, it yields $\frac{1}{4}$ case of frequency to (I/I, III/I), $\frac{2}{4}$ case to (I/II, III/II), and $\frac{1}{4}$ case to (I/III, III/III), for entry at the intersections of the lines (I, III), (I/II, III/II), and (I/III, I) respectively.

But, in order to avoid the inconvenience of quarter values, it is better to suppose the original figures in the fig. and in Table I above to have been replaced by those in Table II; then the original entry

Fig. 1.

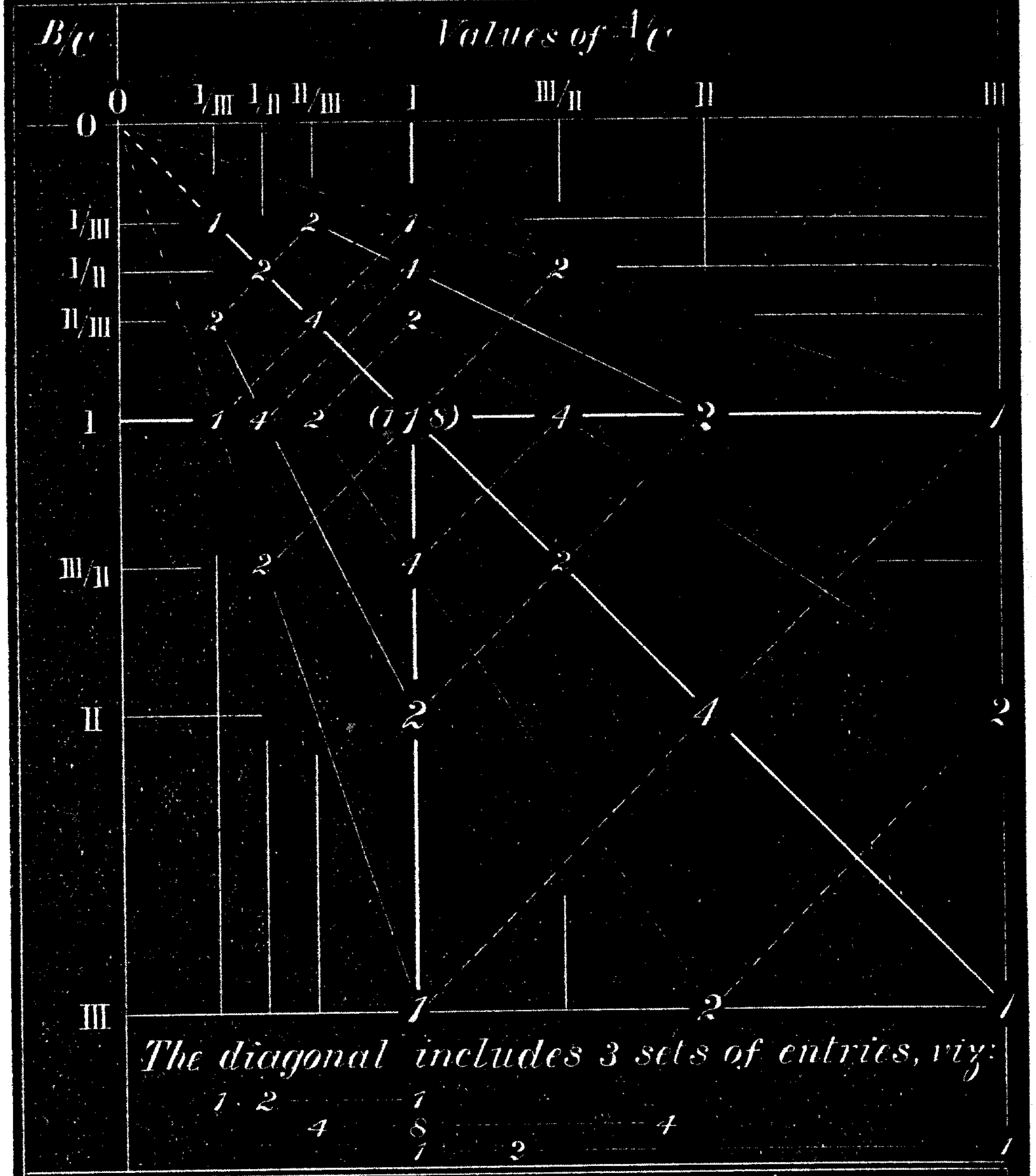
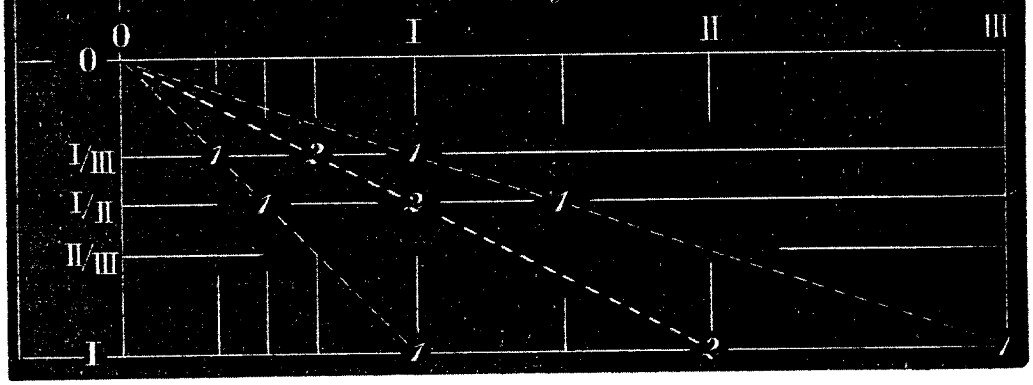

B/C
Values of A/C
0
I/III
I/II
II/III
I
III/II
II
III
0
I/III
I/II
II/III
1
III/II
II
III
1 2 1
2 1 2
2 4 2
1 4 2 (1 1 8) 4 2 1
2 1 2
2 4 2
1 2 1
The diagonal includes 3 sets of entries, viz:
1 · 2 1
4 8 4
1 2 1

Fig. 2.


0
I
II
III
0
I/III
I/II
II/III
I
1 2 1
1 2 1
1 2 1

from which we start will have become four, to be expanded into three derivative entries, having respectively the frequencies 1, 2, and 1; these latter figures are entered in fig. 1 at the intersections of the lines just named. Under this arrangement the large figure from which we started, which had been changed from 1 to 4, again assumes its original value of 1. It will easily be understood, that the positions of the three derivative entries necessarily lie in the same straight line, and that this line necessarily runs towards the (O, O) corner of the figure. The same is true for every other set of derivative entries, with the result that whereas the original set of large figures, referring to the combinations of A and B, are symmetrically disposed on either side of the horizontal, of the vertical, and of the diagonal lines passing through their common centre at (II, II), the derivative values of A/C, B/C are disposed symmetrically only in respect to the diagonal line that runs from the (O, O) corner. Their symmetry, in this sense, is well shown by the dotted connections between the corresponding figures on either side of the diagonal. Also, it will be seen that the diagonal passes through the regions of greatest frequency. It follows that the diagonal in question represents the *locus* of average frequency. Now, along that diagonal, each value of A/C is associated with identically the same value of B/C; in other words, a correlation is found to have become established between them, which is solely due to the fact that *each* member in every couplet of A/C, B/C values is divided by the same value of the variable C.

We will now submit the above process to the test of extreme cases.

First, let the variability of A be so small that it may be treated as a constant, and take it = 1.

Then the values of A/C and B/C, that are severally associated with the three values of C, are as follows:—

Table III.

C.	A/C.	B/C.			Corresponding frequencies.		
I	I	I	II	III	1	2	1
II	I/II	I/II	I	III/II	1	2	1
III	I/III	I/III	II/III	I	1	2	1

These frequencies are laid down at their proper places in fig. 2, where the three entries, corresponding to each successive value of A/C, run in vertical lines, but, on connecting the entries of maximum frequency it is seen that they coincide with the diagonal from the O/O corner; also that the entries of minimum frequency are disposed symmetrically on either side of that diagonal and converge towards the same corner. Consequently, the existence of spurious correlation is manifest here. If B be the constant, and A and C the variables, the general results will of course be the same.

Secondly, let both A and B be constant and equal to I, and C the only variable; then there are only three possible combinations of A/C and B/C. In one of them both values are equal to I, in another to I/II, and in the third to I/III, all of which lie along the diagonal from (O, O), and thus testify to intimate correlation.

Lastly, let C be the only constant and equal to 1. Then A/C, B/C, become A and B, and the table of frequency of their various combinations is that shown in Table I and by the large figures in fig. 1, whose symmetrical disposition in all directions proves that there is no correlation.