WILEY

On the Future of Statistics

Author(s): Harald Westergaard

Source: *Journal of the Royal Statistical Society*, Vol. 81, No. 3 (May, 1918), pp. 499-520

Published by: [Wiley](#) for the [Royal Statistical Society](#)

Stable URL: <http://www.jstor.org/stable/2340859>

Accessed: 28/06/2014 11:35

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>

ON THE FUTURE OF STATISTICS.

By HARALD WESTERGAARD.

IN a recent number of the *Quarterly Publications of the American Statistical Association*, I have tried to explain my views in regard to the theory of statistics ("Scope and Method of Statistics," September, 1916). It was my hope that my Paper would lead to a discussion on this serious question among experts on both sides of the ocean. Some interesting comments in the same number of the *Quarterly Publications* proved that my hope was not in vain, and I have, furthermore, had the privilege of being criticised in this *Journal* (July, 1917)—even though rather severely—by Professor Edgeworth, whose contributions to mathematical theory of statistics are generally admired. Thus far I have been successful, and I have no right to complain of the criticism, since it will give me the advantage of once more stating the problem as I see it. I confess that Professor Edgeworth has not convinced me of being on a wrong track; I shall make an attempt to set forth at some length why I cannot share his views, at the same time trying to meet objections from other sides.

The great problem in all science is to find the causality, to enable us to trace the causes of a given phenomenon and to foretell coming events, where the causes in action are known. Thus, a chemist is able to tell us exactly under what conditions hydrogen and oxygen will unite as water. But the statistician will very rarely be in this position. He can tell us that one occupation is more dangerous than another, but he cannot foretell when a person belonging to this occupation will die; he can prove that a person who was once sentenced for a crime will more likely come in conflict with the law than others, that males more frequently commit suicide than females, &c., but he is unable to say if a certain person will commit a crime or die from suicide. It will be his aim to single out certain observations or groups of observations, which show a greater frequency of a certain event than others, but he will very rarely reach the point where all elements will give exactly the same result. In fact, there is always an incalculable number of causes in action, which influence the result, and it will be quite impossible to trace them all. Thus one person attacked by a contagious disease will die, whereas another with apparently the same physical constitution will recover; it will perhaps be quite impossible to make out why the latter had a greater power of resistance than the former. So also in the case of games of chance. If I draw a ball out of an urn in which several

white and red balls are lying, I cannot foretell if a white or a red ball will come out. There must be some difference between the causes influencing each ball, for else all balls would be drawn at once, or all would remain in the urn, but this difference escapes our attention. The statistician must be content, if he can foretell the collective result of a mass of drawings, whereas the individual case stands as an unsolved problem. We might say, that in the collective result the mass of individual causes disappears, they are so to speak eliminated, at least to a certain degree, leaving the field to a relatively small number of causes, which we can then proceed to investigate. If only one such influence were left, we should be nearly in the same position as the chemist, but, unfortunately, there are in most cases several causes left, which all contribute somehow or other to the collective result we have in view. Here we have the reason why statistics so often are accused of mendacity. One will ascribe a certain phenomenon to one cause, another statistician will call attention to quite opposite influences. It may be, that both to a certain degree are right, each cause having its share in the result, but no answer is possible, until all the causes can be separated. To illustrate this, I shall choose a simple example from a Danish investigation on the frequency of still-births. We have here the following numbers :—

Age of mother.	Still-births among 100 births.	Number of birth.	Still-births among 100 births.
Under 25	1.5	First birth	2.7
25-29....	1.9	Second	2.0
30-34....	2.9	Third	2.2
35-39....	3.8	4th-5th	2.7
40-44....	5.1	6th-7th	3.4
45	5.9	8th-9th	4.7
Total	2.8	10th and above	5.0

Thus, out of 1,000 births, 28 were still-births, and the rate of still-births increased with the age of the mother, and also—except in the case of the first birth—with the number of previous births of the mother. We are very far from being able to give a prognosis of a still-birth; there must be other important causes left, since the highest rate of still-birth is only 5.9. And, moreover, the two influences here considered are hopelessly entangled, for the elderly mothers have generally had several births, whereas the young mothers rarely will be multiparae. It may be, that the still-birth is more frequent, because the mother is in advanced years, but the reason may also possibly be that she has borne several times before.

In order to answer the question, it will be necessary to divide the material, each subdivision having its own age of mother and its own number of birth. Unfortunately, this can only be done for the total number of bearing women, not for the still-births alone. But we can then proceed in the following way: we can suppose the same frequency of still-birth for all mothers belonging to the same age group, asking how many still-births we might expect among first born, second born, &c., if the said frequency held good in all these sub-groups. In this way we can eliminate the influence of the age, and *vice versa*, we can eliminate the influence of the birth-number in order to see what is the influence of age. We thus find the following results, showing the number of still-births to 100 expected cases:—

1st child 142 | 2nd-7th child.... 90 | 8th child and above 108

and by elimination of birth-number:—

Mother under 25		59	Mother 35-39		116
„ 25-29		79	„ 40 and above		136
„ 30-34		107			

It follows from these numbers, that both age and birth number have influence on the frequency of still-birth, but much less than according to the previous numbers, which were under a double influence; the birth-number in particular—with exception of the first-born—does not show much variation.

Very often a division of the observations in sub-groups will be impossible, so that we have no means of finding the causality. The following numbers taken from the municipal statistics of *Budapest* will illustrate this¹:—

District.	Birth-rate, p.m., 1902-13.	Infant mortality, p.m.	Percentage of population living in tenements with one room (1911).
IV.	14·4	85	22·2
II.	22·0	112	37·7
V.	23·2	120	40·6
I.	24·5	125	42·3
VII.	25·8	107	42·5
VI.	26·9	125	47·4
VIII.	27·8	131	56·2
IX.	27·9	145	59·2
X.	30·0	170	74·0
III.	31·7	183	75·9

¹ *Die Sterblichkeit der Stadt Budapest in den Jahren 1908 und 1909.* Budapest, 1916.

The infant mortality (number of births under 1 year to 1,000 births) is—as it appears from the numbers—considerable where the birth-rate is large, and small where there is a low birth-rate, the only exception being the seventh district, where, however, the age-distribution seems peculiar, with relatively many persons in the age 25–50, so that the birth-rate actually is smaller than it appears from the crude birth-rate. There is also apparently a “correlation” between birth-rate and overcrowding, as well as between the latter and death-rate. But it is impossible to disentangle these influences, no one can tell how much a high death-rate is owing to density of population or to a high fertility. In order to draw safe conclusions we should have to divide the districts in sub-groups according to both density of population and fertility.

Still more complicated are the relations between fertility or density of population and tuberculosis of the lungs. Thus, we may separate the districts according to the average number of persons to each room, collecting 5 districts with less than 2.60 per room in one group (with the number of population as weight) and districts with greater density in a second group. We shall then find the following numbers :—

	Average density.	Mortality, p.m., from tuberculosis.	Birth-rate, p.m.	
	2.28	2.4	24.3	
	2.92	2.8	27.7	

A conscientious statistician will be very cautious in dealing with such complicated numbers ; there are evidently so many disturbing causes that no safe conclusion can be drawn. Many questions may be asked :—Will a high birth-rate give increased risk of dying from tuberculosis ? *Vice versa*, is the density of population high on account of the high birth-rate ? And so on. Here, too, it will be necessary to separate the observations in numerous sub-groups in order to trace each influence—I shall later on discuss curious examples concerning secular influences, the relation between time and social or economic phenomena, to show how difficult it is to draw trustworthy conclusions from material of this kind.

To choose a third example, I shall briefly discuss Galton's famous theory of regression, as he described it in his *Natural Inheritance* (1889). The following observations will illustrate his theory. They relate to the correlation between height of “mid-parents” (viz., the average of the father's and the mother's height, the latter multiplied

by 1.08) and of their children (the daughter's height being equally multiplied by the same factor).

Height of mid-parents in inches.	Height of the adult children.	Number of	
		Mid-parents.	Adult children.
72.5	72.2	6	19
71.5	69.9	11	43
70.5	69.5	22	68
69.5	68.9	41	183
68.5	68.2	44	219
67.5	67.6	33	211
66.5	67.2	20	78
65.5	66.7	12	66
64.5	65.8	5	23

The small numbers of observations above 72.5 or below 64.5 inches have been left out of consideration ; even the other observations are hardly numerous enough. It follows from the table that the children show a smaller dispersion from the median value than their parents. Galton saw in these phenomena a consequence of the laws of inheritance, but the question is, whether the observations are clear enough to allow conclusions of this description. In fact, it is impossible to separate the various biological types, of which the material may be supposed to consist. It may be asserted that observations on human stature very often show a rather close approximation to the exponential law of frequency, even if there are various types ; an interesting illustration will be found in the Italian military statistics showing a most remarkable variation down through the peninsula from the tall Lombardian soldiers to the small Sicilians (Professor Edgeworth is also aware of this fact, see, for instance, *Mathematical Representation of Statistics*, p. 32). This heterogeneous structure of material does not exclude us from certain conclusions in accordance with the calculus of probabilities. Thus, we may calculate the mean error of the average height of an Italian soldier. But in other directions it will be impossible to draw any conclusion from material of this kind. In a recent work in Danish on the theory of statistics I have given a curious instance : if we group the Italian soldiers according to height and cephalic index, we shall find the following numbers :—

Height.	Cephalic index.	Height.	Cephalic index.
Under 160 cm.	82.0	165-170 cm.	82.8
160-165 cm.	82.4	170 cm. and above	83.1

It would appear from these numbers, that there is a real correlation, tall persons having a broader shape of their head than small ones, but if we separate the material after provinces, we shall find that no such rule prevails.

Returning to Galton's observations, let us suppose that there are only two biological types, a short and a tall one, and that the height of the type is inherited. The only means to divide the observations will be to separate according to stature, as in the table above; but then the groups will actually be very mixed. Persons belonging to a short type may in some circumstances *de facto* be of tall stature, and persons with the tall type may, by chance, be short. It must be borne in mind, that human stature is very much influenced by circumstances. Thus the Danish soldier has, during a half century, shown an increase of stature of no less than 4 cm., probably in consequence of economic and hygienic progress. In Galton's observations, among parents with a height of $71\frac{1}{2}$ inches, there may be several who only by chance reached that stature, but who really belonged to a low type. Consequently, the children of the tall parents who, according to the hypothesis, on an average inherit the typical height, will have a lower average stature than these, whereas the parents with a small height will have children with a greater average height. Thus Galton's law of regression may be of entirely formal nature, and it seems impossible here to disentangle the causes. If the observations had gone the other way, if there had been no apparent "regression," then it would have been safe to conclude that tall parents had tall children, and to maintain that this influence would appear still greater, if the observations were unmixed. In fact, the same material may allow a conclusion in one direction—in spite of its inaccuracy—and not in another direction. The statistician must be extremely critical in dealing with any series of observations.

In this critical revision of the observations we often feel safe when we reach a distribution approaching to the binomial law, though even here, as mentioned above, great care must be taken. But, as a rule, statistical observations are not grouped according to this law, but they show quite other forms of distribution. It will be sufficient to mention the case of income statistics where the distribution generally has a pyramidal form (Pareto's law). What shall we do with statistics of this kind?

Many statisticians will reply that we must try to find a mathematical formula which these observations may be said to obey. This has been the case as long as there has been a theory of statistics: several attempts to find a "law of mortality" will illustrate this.

Some of these formulæ are excellent points of departure for an interpolation, but we are not entitled to look upon them as an expression for physiological laws. It is evident that any population in countries with Western civilisation—nay, even in much more uniform societies—is a conglomerate, where each element has its own mortality. This is clearly seen at a glance in the reports of the Registrar-General of births, deaths and marriages on the mortality of various professions. The immense secular changes of mortality, too, require constant changes in the formulæ. The law of mortality was good coin a century ago, when it was a common creed that mortality was constant, that if a cause of death like the small-pox disappeared, other diseases would take its place, but now the coin has been much depreciated.

But what can be said about the mortality formulæ like Makeham's, Thiele's and others, will hold good of all other formulæ invented with the aim of describing the irregular distribution of the actual observations. I do not deny their usefulness as formulæ of interpolation nor the genius of their authors, but I cannot see that they will enable us to draw conclusions with regard to the causality. They have a peculiar form, because there is a multitude of causes, and we cannot prove anything till we have investigated all these causes. Thus the distribution, according to income or to wealth, depends on innumerable causes influencing the social and economic conditions of the country, and they are constantly changing; we need not refer to the present war for instances of those changes. I have tried elsewhere to show that it seems possible to divide the population into several groups, each one more or less obeying the binomial law of frequency, and we have no chance of really understanding what is going on until we have reached this goal of separating the material. So also with regard to the statistics of wages, or the grouping of farms according to size, &c., &c.

If I am right in looking on these various formulæ only as instruments of interpolation, I feel justified in maintaining that it is not always necessary to have any formula, the observations themselves often giving a sufficient basis. Thus we may take the question of finding the distribution of a population according to years of age, if we know only the distribution according to decades. This is a very puzzling problem, in spite of its apparent simplicity; most mathematical formulæ of interpolation (as for instance the Newtonian) will generally give absurd results in extreme old age, *e.g.*, negative values in the ages above 90 or 95. Of course mathematical formulæ might be found which would always give positive values, but the results would generally not repay the great trouble of calcu-

lating all the constants which enter into the formulæ. But if we choose a common population, where the age-distribution is known, as standard, we shall usually be able to obtain fairly good results. Let y_x be the unknown number of persons above a given age, u_x the corresponding number of a population with detailed statistics, then we may put $y_x = u_x \phi_x$, where ϕ_x is a simple formula of interpolation, for instance, $a + bx + cx^2 + \dots$, knowing $y_1, y_2, \&c.$, we can easily find the constants a, b, c . In fact, if we were finding the numbers by graphic interpolation, and we began by drawing the curve representing the known population, we would always follow this line in drawing the curve representing the group of population we want to distribute. u_x is a function, perhaps of a very complicated form, but we need no calculation of constants, as we have a sufficient number of observations to find all the values of y_x required.

I have quoted another example relating to the decrease of the birth-rate in my article "On the Scope and Method of Statistics." We find that the Danish statistics on fertility show great differences between various classes of society; in one group there is hardly any reduction of birth-rate, in the other there is a conspicuous reduction. Comparing these two groups we shall find, however, that the distribution of families with a large number of children, for instance, above 6, will correspond very nearly in the two groups. We might, consequently—as an experiment—calculate the number of families with less children than 6 according to the same scale, so that we separated a sub-group with no reduction of birth-rate. The remainder forms a sub-group with very few children, families with modern ideas on fertility; this sub-group having, on an average, 2.2 children per family, the other 4.3. Of course, it is a very hypothetical calculation, though the results perhaps are convincing enough. At all events, it would be waste of time to look out for complicated formulæ of interpolation in order to get a smooth distribution, if we can get tolerably safe results without these formulæ. It would not pay to apply the whole machinery of modern mathematical statistics on a side question like this; too often mathematical formulæ, for such a purpose, might even act as the bed of Procrustes.

Of course, we may form hypotheses on the causes which govern a certain series of observations. Professor Edgeworth expresses his view by saying that we require a less simple type of chance—"not the frequency of black balls in a sample taken from a uniform mixture . . . but the frequency of black balls in a composite sample formed by taking a certain number of balls from each of several urns with balls mixed in different proportions." I quite agree that

this would be the next step, leading as it does to Poisson's formula for the exponential law, where there are two or more probabilities prevailing. We can use this formula everywhere without hesitation, under certain suppositions. Thus, if the probabilities are not very different from each other, it would hardly be possible to distinguish the distribution from that which follows from one value of the probability only. Further, where the values differ more widely, the formula may be used so long as the sub-groups keep their relative proportions. If we calculate a common rate of frequency for all groups, we shall get a greater mean error than if we split the material into groups each with its peculiar mean error. This would give us the curious case of a series of observations with an apparently smaller dispersion from the median than according to the formula for the mean error calculated from the common rate of frequency. In fact, it brings us closer to the ideal of separating the material into two groups, one in which the event *never* occurs, and another where it *always* will occur, the mean error thus being reduced to zero. But this case of a constant proportion of the groups is extremely rare. Commonly the proportion will vary incessantly, urban population will gain in relative strength, the numbers of alcoholics and criminals or of paupers will increase or decrease, wealth will change, and so on, and no one can predict what will be the future distribution. As a rule the dispersion will be greater than according to the theory.

It therefore seems to me that there is no way out, if we want to reach thoroughly reliable conclusions, and if we are excluded from separating the material into its various principal groups. But the first aim of a statistician, who does not confine himself to a photographic picture of the state of things, must be to separate the observations in order to understand the causality. The deviations which he observes will lead him to look out for their causes. Thus, in mortality statistics he will be led to observe the rates according to age and sex; but he will not stop there; he will find great differences with regard to conjugal conditions, he will calculate death-rates among legitimate and illegitimate infants, he will separate the rural and town population. Further, he will try to find the influence of occupation, he will divide according to wealth, he will ask whether an uncommonly hot summer will influence mortality or not, he will separate the years in which the population is prospering from economically bad years, &c.; everywhere he will be led to these investigations because there are such great irregularities in the numbers. Where this is not the case, his harvest will be smaller. Thus we have a striking example in birth-statistics.

The birth-rate differs as a rule from year to year, much more than the exponential formula would allow, even where there is no constant decrease, such as we can observe at present in nearly all countries with Western civilisation. Numbers of causes for these great differences may be sought for and eventually found. But if we ask about the relative numbers of males and females, we very soon come to a standstill. It is very difficult for a statistician to find causes for a greater or smaller relative number of males or females; age of parents seems rather irrelevant, and even the influences which the late Dr. Geissler found in Saxon families are rather insignificant, however interesting they are. As I have tried to show, the constant changes with regard to mortality prevent us from calculating a real law of mortality. In the case of the sexual proportion of births we very soon reach such a law, but it is none other than the exponential or rather binomial law of frequency.

We are, however, not always excluded from applying the theorems of probability on a mixed material, where several rates of frequency prevail. As I remarked above, so long as there is no change in the relative numbers of the classes or groups of the material, we are entitled to use Poisson's formula. This will be the case where we want to take "samples"¹ in order to find some result which may hold good for the whole material. On the whole, I think "sampling" has a future in all statistics, and I have often tried to explain the advantage of this method. To choose a simple example. Let it be supposed that we want to know the average rate of interest at a given moment. This is the problem which Dr. Bowley has treated so admirably in the presidential address which Professor Edgeworth quotes in his review. In such a case we shall be able to get very fair results by sampling, supposing that the observations really are taken at random with about equal frequency from each group. A very interesting method of sampling was tried several years ago in a Danish hospital for epidemic diseases in order to test the influence of serum on patients suffering from diphtheria. Patients brought into the hospital one day were treated with serum, the next day's patients got no injection, and so on alternately. Here in all probability the two series of observations were homogeneous. It is just the same problem which Laplace treated in his famous attempt to find out the population of France. Professor Edgeworth objects to my criticism of this attempt—far better as it was than any of

¹ I quite agree that this term must be preferred to that which I have used: "representative statistics," though the latter is familiar to statisticians in several other countries. Writing, as I do, in a foreign language, I must appeal to the indulgence of English readers.

the experiments of the old political arithmetic—that the analysis I propose “might involve more trouble than simply to count the “people by way of a census.” Now it must be remembered that at the beginning of the 19th century, and even much later, statisticians, especially in France, looked on a census with the greatest possible distrust. Laplace could not take out the districts examined at random, he was obliged to choose those places where the mayors were to be trusted, and the result that the distribution of these places might be unfavourable to his experiment could not be excluded. At all events no real investigations had at that time been made as to the conformity of statistical observations with the exponential or binomial law of frequency, this conformity being rather considered as a dogma, and if Laplace should really reach the bottom of the problem, it was indispensable to make thorough investigations of this kind. The result might be, that all the single birth-rates oscillated about the same mode; in this case, Laplace would comparatively soon have been able to come to a definite result, but it is more probable that he would have found a larger dispersion, with two or more modes, and he could not then help applying to my “panacea” of separating the observations. Supposing that in one part of France the birth-rate was 4 per cent., in another, with the same population, 3 per cent., and that proportionally many mayors in the latter part were to be trusted, and had given reports of the numbers of inhabitants (say twice as many); supposing, further, that each part counted 10 millions—all these numbers being, of course, quite fictitious—then the average birth-rate would actually be 3.5 per cent., but we would find only $3\frac{1}{2}$ per cent., and, consequently, we would conclude that the population counted 21 instead of 20 millions. This source of error is avoided by the Norwegian sampling statistics, by which, for instance, the number of invalids was calculated, for it was known exactly how many people lived in each district, and to each one a definite frequency could be attributed; that there were other innate difficulties in this case is willingly agreed.

I now turn to the most serious objection, viz., that I have been unjust to the theory of correlation, and that it is not too much to say, with one of my commentators, Professor West, that I am apparently not familiar with the properties of the correlation ratio. Of course, a statistician who, according to his conviction, rejects an instrument as not suitable for his purpose, would hardly be so versed in using it as his adversaries, but I shall try to prove why I cannot look upon the theory of correlation as an advance: it seems to me rather more fit to hamper the statistician by involving

more labour than necessary. That I do not believe in the Galtonian law of regression I have already mentioned, but we might, in general, accept the method of correlation, even if differing with regard to this theory. Professor Edgeworth intimates that I have mainly had "in view only statistics in the primitive sense of the term, relating "to *States* or more generally social phenomena." I confess, that I do not see the difference between statistics of this kind and observations concerning physical characters of parents and children, and all the other observations dealt with by the school of Biometrika. So I willingly take up the glove, fully convinced that I shall render future statistics a service, if I can prevail upon students to try the old instruments first before turning to the more complicated machinery of correlation. I beg leave, however, first to discuss at some length a certain group of problems, though I may be accused of taking a circuit.

The most frequent problems of correlation will perhaps arise, when statistical observations are grouped according to time, for instance, yearly birth-rates, mortality-rates, exports or imports, prices, etc. Here the first duty of the statistician will be that of carefully criticising his observations, not only as to accuracy, but also as to their nature. In fact, very often there will be common social and economic causes behind all these phenomena ; they are, so to say, the common root of these phenomena, but between the common root and the branches and leaves which we observe, there is often a great distance. Let me choose a curious example from Finland (taken from the last year-book).

	Exports (million marks).	Suicides.	Divorces.	Number of foxes killed.
1894-98	7,865	604	535	25,748
1899-03	9,858	718	583	18,067
1904-08	12,570	922	722	14,892
1909-13	16,115	1,472	1,184	10,341

Evidently there is a beautiful correlation between these various columns, but who would maintain, that the increase of the value of export is the reason of the increasing number of suicides or divorces, or that the foxes killed will decrease when the suicides increase, etc. Nor can we say, whether the number of foxes killed is smaller now than it used to be, because they are to a great extent exterminated, or because less attention has been paid to their persecution. The first thing to be done is, therefore, closely to study all the conditions of these numbers ; only after having done that can we proceed to

formulæ of correlation. As I have no data on which to base this preliminary discussion of the causality, the following remarks can be looked upon as only tentative, but they will, at all events, be illustrative of the methods available. Having studied the nature of the observations, the first step will be to follow the movements of the numbers in particular, for instance, by means of a graphic representation, in order to see whether there is connection between the curves, if the ups and downs are contemporary, etc., if a conspicuous variation in one column of numbers corresponds to a similar one in the other, if a minimum can be observed for the same year in the two curves, and so on. A graphic representation will perhaps often be sufficient, but many statisticians will wish to find some more definite expression for the connection between the two columns of observations. Unfortunately, it is impossible here to avoid introducing some arbitrary supposition. We can, for instance, suppose a certain smooth movement for one series, according to some more or less simple function with so and so many constants. If a year shows an actual number above this curve, we will ask whether the same holds good with regard to the other series. But what formula will be the best, and how many constants shall it contain? If there are 20 years, then a formula with 20 constants would represent the movement perfectly; but it would give us no ups or downs, as each number would exactly cover the calculated one. The choice is then an arbitrary one, the formula can have several forms, for instance, $a + bx + cx^2 + \dots$ or $e^{f(x)}$, &c., &c., and it will not always be easy to decide which is the best. One of the best means would perhaps be a graphical adjustment; after some practice, the curves which are tentatively drawn will soon approach each other very closely, and after some attempts we may be able to draw a curve which it will be difficult to improve. But here, too, there is something arbitrary in the solution, and it will be very difficult to do away with eventual objections and doubts.

Many statisticians will feel satisfied by a mechanical adjustment, for instance, by taking moving averages of three or five consecutive values as the adjusted values for the years in the middle. Using this method here, leaving out the two first and the two last values, we get adjusted numbers for 16 years from 1896 to 1911 inclusive. Thus, for 1906, the quinquennial average of divorces is 133.6, whereas the observed number is 153, and so on. We can next pick out seven years where the calculated number of divorces is higher than the observed one, and compare these results with the corresponding adjusted and actual numbers of suicides. We thus get the following results as a yearly average. We must, however, bear in mind

that the numbers of observed cases are very small, so that the deviation from the normal curve quite naturally will be rather large, but at all events, the calculation may serve as an example.

	Divorces.		Suicides.	
	Calculated.	Observed.	Calculated.	Observed.
Above the expected number	153	167	164	164
Under „ „ „	139	129	169	165

The years with low numbers of divorces have, on an average, a small diminution of suicides, but the mean error is considerable; at all events, the correlation is very feeble indeed. The corresponding numbers for the *foxes* will be :—

	Calculated.	Observed.
Above the expected number	3,408	3,427
Under „ „ „	3,415	3,386

The difference is evidently very small, there seems to be no correlation whatever, as we might have foretold. *A priori* we would more naturally expect to find a correlation between exports and capital of the savings banks, and here we actually observe some coincidence: though the influence is very small indeed: a high savings-bank capital corresponds to a high export-number, and *vice versa*. The divorces show very trifling movements in the same respect, and this is also the case with the suicides. These calculations will, as a rule, not give essentially different results, if other standards for the general movement are chosen (other formulæ of adjustment). With some caution, we shall be able to see whether there is a conspicuous mutual influence or not.

I may here quote a rather unfortunate attempt to find the correlation between two columns. It testifies to the great risk of false conclusions, if we do not keep the observations strictly in mind. It is to be found in the report from the Swiss Society of Actuaries for June, 1917¹. The author, Mr. E. Hæmig, appears to be a disciple of G. F. Lipps, whose work, *Die psychischen Massmethoden* (1906), he follows somewhat uncritically. He treats the yearly

¹ "Systematische Abhängigkeitsbestimmungen" (*Bulletin de l'Association des Actuaires suisses*, Berne, 1917).

births- and mortality-rates in Canton, Zürich, 1871-1900, asking how often a yearly rate is smaller or higher than the numbers for the remainder of the period. Thus, in 1881, the birth-rate was 31. The same number was observed later for two years (1898-99), whereas the rate of 30 was observed three times, 29 also three times, 28 and 27 each five times, and 26 only once. Giving each of the two years with 31 the weight $\frac{1}{2}$ each, he thus finds that 18 years have a smaller birth-rate, and only one a higher. It would have been advisable to calculate the rates with one or two decimals more, but this objection has, of course, little consequence with regard to the principles. Altogether, he finds for the whole period, 169.5 cases where the rate of the year he considers was higher, 265.5, where it was lower than the following rates, whereas for the mortality the numbers are 59.5 and 375.5. This, according to the author, shows a decrease both of the birth-rate and of the death-rate, the latter having a stronger movement downwards. A simple inspection of the numbers will show the movements much clearer; thus we have in quinquennial averages :—

	Birth-rate.	Death-rate.		Birth-rate.	Death-rate.
1871-75	30.4	23.8	1886-90	26.8	19.2
1876-80	31.8	22.8	1891-95	28.0	18.6
1881-85	29.0	20.4	1896-1900	30.4	17.8

The movement of the birth-rate is undulating, the mortality is decidedly decreasing. Mr. Hæmig's method gives too much weight to the earlier years; thus 1871 is considered 29 times; 1899 only once.

Proceeding now to the contemporaneous movement of the two columns, the author finds that in $12\frac{1}{2}$ cases the birth-rate in 1871 compared to the following 29 numbers, has the same movement as the death-rate, whereas in 16.5 cases the movement is in the opposite direction, and for all the thirty years together, he finds respectively 275.5 and 159.5, thus showing a tendency to parallel movements of the two rates. But the great difficulty is the constant decrease of the mortality, which is not taken in due consideration. A much more practical and simple solution will be to find an expression for the movement by way of some adjustment or a graphical representation, and then to compare the number of cases where the rates are above or below the theoretical values, just as described above. We shall thus find the following results, taking moving averages for five years as theoretical values. Out of twenty-six available years

(1873-98), we find twelve years with a favourable mortality, three years where the mortality was exactly the same as the theoretical value, and eleven years where the rate of mortality was high. In the three groups the actual and calculated values were the following :—

		Average, actual.	Death-rate, theoretical.	Average, actual.	Birth-rate, theoretical.
Low death-rate		19·9	20·8	29·4	29·4
Average	„	19·0	19·0	28·3	28·3
High	„	20·8	19·9	29·4	29·4

A low death-rate has, consequently, not been combined with a low birth-rate, the calculated and observed values of birth-rate are alike in all three groups. If any influence had prevailed, we should have found a marked difference.

These examples will show the route we may choose by investigations on correlation between two series of observation on social or economic phenomena ; if there is any influence, we shall, as a rule, be able to find it by this very simple process.

I shall now turn to the objections by Professor Persons in his comments on my "Scope and Method." He emphatically disagrees with my characterization of the co-efficient of correlation, and refers to the problem of finding the correlation between steel-production and prices, stating as his view, that "such a question "can be answered only by the application of a device such as the "co-efficient of correlation, which enables us to measure the degree "of correlation." He has exposed his theory in one of the following numbers of the Quarterly in a highly interesting article,¹ in which he discusses various methods for comparing the mutual influences of a number of series of economic statistics. Though I run the risk of repeating something of what I have said above, I hope the following remarks will not prove useless.

As far as I can see the problem is a double one. *Firstly*, the problem is to find a "secular" or a "cyclical" movement of the quantities observed ; *secondly*, to discuss the difference between these theoretical or adjusted values and the observed ones.

Now, as to the first side of the question, I quite agree with Professor Persons, that "curve fitting to eliminate the secular trend "of a time series should always be adapted to the problem in hand."

¹ Warren M. Persons, *On the Variate Difference Correlation Method and Curve Fitting* (June, 1917).

This is exactly what I have constantly argued, one of the first duties of a statistician being that of keeping in contact with the fundamental data. Professor Persons tries various methods of curve fitting for the secular movement, thus the straight line, or even two straight lines each for one sub-period, a moving average, a parabola or a compound interest curve. It will be seen on inspection of the observations that there are very striking differences, each series having its own characteristic movements. Thus, the wholesale price indices move slowly from a maximum in 1882 (1,327) to a minimum fifteen years later, again rising slowly to 1,352 in 1913, whereas the pig-iron production increases moderately till 1896, and afterwards rises immensely. Now, as I explained above, there must be some arbitrariness in the choice of the curve, even if we try to keep in intimate contact with the fundamental data. For in no one of the series will a straight line or a couple of straight lines be a true expression of the movement, nor will a compound interest curve do ; a parabola will appear less arbitrary, at least, a parabola with several constants, but how far shall we proceed in our adjustment ? Nor will the moving average give us an ideal picture of the movement, as it is a too mechanical adjustment of the observations. Thus, the pig-iron production in 1908, which was uncommonly small, will influence the average numbers too much, and so on. If we could find a curve which, from the nature of the observations, gave a true representation of the general movement, the deviations from this curve being rather accidental, then it would be a relatively easy task to study these deviations and to find eventual correlations. But this is impossible, it would be so to say superstitious to expect that a typical curve could be found.

Having found a curve which satisfies us more or less, we proceed to the *second* part of the problem, viz., to compare the deviations between the adjusted and the observed values. Here we may choose the method of comparison just as we like, we can use a formula of correlation, or quite an elementary method. We are by no means confined to the application of a co-efficient of correlation, in fact, the choice of a method of comparison is quite subordinate ; when I prefer the most elementary method, it is for the practical reasons I have explained above. Taking, for instance, moving averages of price indices for five consecutive years, we may ask, how many years gave higher prices than this level, and what was the aggregate deviation, and, further, we can ask, what are the adjusted and observed values in the other series. The following average numbers will be the result of this comparison :—

Prices.	Observed prices.	Adjusted prices.	Corresponding numbers for pig iron (100,000 tons).	
			Observed.	Adjusted.
Fifteen years above the price level	1,183	1,159	129	122
Sixteen „ under „ „	1,091	1,114	123	121

It will be seen that years in which the prices were under the level chosen as representative of the movement have had a somewhat reduced production of iron. But an inspection of the numbers will show us very striking exceptions; still less is it possible to predict how large the values will be, even if they move in the same direction in both series. It will be necessary to study the single years very closely in order to understand these variations thoroughly. This will hold good whatever formula of adjustment has been chosen. Thus, in 1886, the price index was low, whereas there was a considerable rise in the pig-iron production; in 1893 prices were high, but the pig-iron production was under the level, and again, in 1905, just the opposite took place.

It may be, that the movement of the prices is not quite contemporary with that of the production. A stimulated production may follow a high price-index, etc. This does not exclude the application of the same simple method, we need only give one of the columns a push backwards or forwards, comparing, for instance, the prices 1886 to the iron production 1887, etc.

If it is advisable to use elementary methods of comparison here, it is still more so with regard to a problem treated in a previous number of the *Quarterly*, by Professor Gehlke.¹ The author puts the question, whether woman suffrage is opposed by those groups in the electorate who are favourable to the continuance of the legalised manufacture and sale of alcoholic liquors. The records of the respective elections in the voting precincts of a certain county were written on cards, out of which two sets of 100 cards were taken at random. Each set was sorted according to the number of votes in 1912 on suffrage and on licensing the sale of liquor, and in the same manner for 1914 concerning suffrage and prohibition. The tables are of the usual form, with double entry, just like that of the age distribution of brides and bridegrooms. I have, by a very simple process, extracted the following numbers of one of these

¹ *On the Correlation between the Vote for Suffrage and the Vote on the Liquor Question* (March, 1917).

sets, supposing that all precincts practically are alike with regard to number of votes—

	Percentage of votes for suffrage.	Percentage of votes for licence in the same precincts.	
	Under 25 per cent.	31.1	
	25-45 per cent.	24.8	
	45 per cent. and upwards	18.9	
	Total	24.5	

Thus, we see that where there are many votes for women's suffrage, the voters are less inclined to go in for licence than in the opposite case. Professor Gehlke thus seems right in his hypothesis that many voters consider women's suffrage a doom of the legalised sale of liquor. In order to reach this conclusion, it is, however, not necessary to use the complicated process of calculating correlation co-efficients, a method which will require much more labour than the above-described method.

By the same simple process we shall find the following results :—

	Percentage of voters for suffrage.	Percentage of voters for prohibition in the same precincts.	
	Under 25 per cent.	15.3	
	25-45 per cent.	32.8	
	45 per cent. and upwards	49.7	
	Total	30.7	

Here, again, we find the hypothesis supported. Where there are many voters for suffrage there will be many who prefer prohibition to licence. Of course, it will be necessary first of all to test the reliability of the material. And, further, it may be remarked that the time spent in working out the figures according to the more complicated method might have been used to treat all the 500 cards instead of 100 or 200. But these remarks may be laid wholly aside here where the question is one of principles ; we are entitled to use the numbers as an illustration without criticising their reliability.

It follows from what I have remarked here, that I do not look upon the formulæ of correlation as misleading. The only thing which I will maintain is this, that Bravais' formula, and other methods connected with it, do not show anything more than the ordinary

elementary process. In fact, both processes are based on quite the same principles ; Bravais' formula does not introduce a new principle into the calculus of probability. Taking a table of observations with double entry, for instance, on the mutual age of brides and bridegrooms, we may try to group these observations around some curve, each age of the bride having its peculiar centre of gravity for the ages of the bridegroom and its peculiar dispersion from this centre. Now, supposing that this curve is a straight line with the equation $y = kx$, we can find k under the condition that the sum of the squares of all the deviations shall be minimum according to the usual process in the method of least squares. We shall, consequently, have the minimum of $\Sigma(y - kx)^2$, and, further, we must have :

$$k = \frac{\Sigma yx}{\Sigma x^2}$$

If we put $\Sigma x^2 = nS_x^2$, $\Sigma y^2 = nS_y^2$ and $\Sigma xy = nS_{xy} = nsS_xS_y$ we shall have the equation :

$$y = kx = r \frac{S_y}{S_x} x,$$

and for the mean error μ we find :

$$\mu^2 = S_y^2 (1 - r^2)$$

If $r = 1$, the mean error will be zero ; all the observations will be found on the straight line $y = kx$. The smaller r , the smaller the correlation ; if $r = 0$, we have a horizontal line, all ages of the bride give just the same distribution of the bridegrooms, the former numbers have in this case no correlation whatever with the latter.

These formulæ of correlation being directly derived from the formulæ of the calculus of probabilities, I feel justified in my assertion, that they do not introduce a new principle. Some statisticians may prefer the formulæ, looking upon them as a better and more practical instrument than the common method. I belong to the party who is of the opposite opinion, and for a very simple reason. Generally, we get the safest conclusions from a given statistical material, if we can keep the observations directly in view. This, in fact, should always be the case. A trained statistician may be able to see the numerical observations with his inner eye, even after having finished all his manipulations, and after having figured out the co-efficient of regression. But if he has mostly been trained in pure mathematics, having been attracted by all the interesting and complicated problems to which the correlation theory gives rise, if he has not had the opportunity of dealing with many practical statistical problems, he will run the risk of forgetting realities for formulæ. Having mastered the whole technical

apparatus, he will perhaps look on the realities with a feeling of indifference, whereas the mathematical problem claims his interest. Very often we learn too little from statistical investigations of high scientific value, from the simple cause that it is impractical to shoot sparrows with cannons.

I challenge statisticians to show me problems which can be solved with greater facility by means of the modern formulæ than was the case in the good old times. As far as I can see, it will be a great advantage, if all young statisticians first of all will try to make themselves familiar with the observations—rough and inaccurate as they too often are—asking how far they can be used, or within what limits they will allow safe conclusions.

In nearly the whole lifetime of theory of statistics or “political arithmetic” many able authors have made serious efforts to find mathematical laws everywhere. Indeed, much labour has been wasted in that way. I believe that the reason why so many statisticians cling to the formulæ of correlation, is a conscious or unconscious idea of the existence of such laws. They share this idea with numbers of statisticians of past generations. Sometimes a hypothesis of this kind is rather naïve; many years ago a Danish author tried to show that the number of telegrams exchanged between two countries depended on shipping and international trade, according to a rather simple mathematical formula. These quantities entered under a square-root, and he was not aware that if the unit changed, for instance, if he used francs in the formula instead of sovereigns, he would find other numbers of telegrams. The case is instructive in spite of the *naïveté* of the author, for it testifies to the fact that any law which we might discover by statistical investigations must have an exceedingly simple form. If a society consists of two groups a and b , the former having the frequency p of a certain event, the latter the frequency q , then the “law” giving us the number of expected events will be $ap + bq$; if the values of a or b change, then the whole expression will get a new numerical value. Of course, we meet many problems where the things appear much more complicated. Supposing that cancer or tuberculosis disappeared from the bills of mortality, what would be the increase of the expectation of life? And what would be the effect if both disappeared at once? This is a rather curious problem, and not without importance in practical life, but behind it we have the fundamental problem, viz., how many deaths will occur in a given population, if these diseases are brought to vanish, and the solution of this fundamental problem will only require a very simple arithmetical process.

I will by no means deny that these more complicated problems are useless. Consequently, we must apply to mathematicians for help. As we need formulæ for the expectation of life, or annuities on human life, so we need methods of adjustment or interpolation. And so, too, we need thorough investigations with regard to the laws of frequency. But, as far as I can see, the fundamental hypothesis with regard to frequency will be extremely simple, leading, as it does, to formulæ derived from the binomial law, and if we find other distributions of the observations, this is a strong appeal to us to separate the material into various groups in order to find the causes of this peculiar distribution.

In all science, as in all human life, there must be a very great division of labour. Statistics will require much assistance from trained mathematicians, but the main object of all statistics must be that of finding the causality by means of the numerical observations. It is most desirable to be able to split the observations as much as possible. There is an abundance of urgent problems in all branches of statistics, in mortality statistics as well as in statistics of births and marriages, in criminal statistics and other chapters of "moral statistics"; further, there are the statistical problems of the "industrial reserve army," or of invalidity and of prices in relation to wages, etc. The statistician is not in danger of being thrown out of work, and each new problem will, as a matter of fact, give rise to other problems, some of them of very great importance for the present generation, whereas others are far out in the horizon. As far as I can see, this is the horoscope of statistics; everything which can promote this evolution must be greeted with satisfaction.

If I have succeeded in bringing someone to share my view and to turn from abstract formulæ to fertile concrete statistical investigation, these lines have not been written in vain. My life's work as a statistician has been to look out for realities, and I shall willingly adopt Professor Edgeworth's expression in his final remarks. In fact, not even do I wish to bring the philosophy of Laplace down to earth, but I also sincerely wish that all statistics shall remain there, and so also my own humble contributions to this discipline.