Kinship and Correlation

Francis Galton, FRS

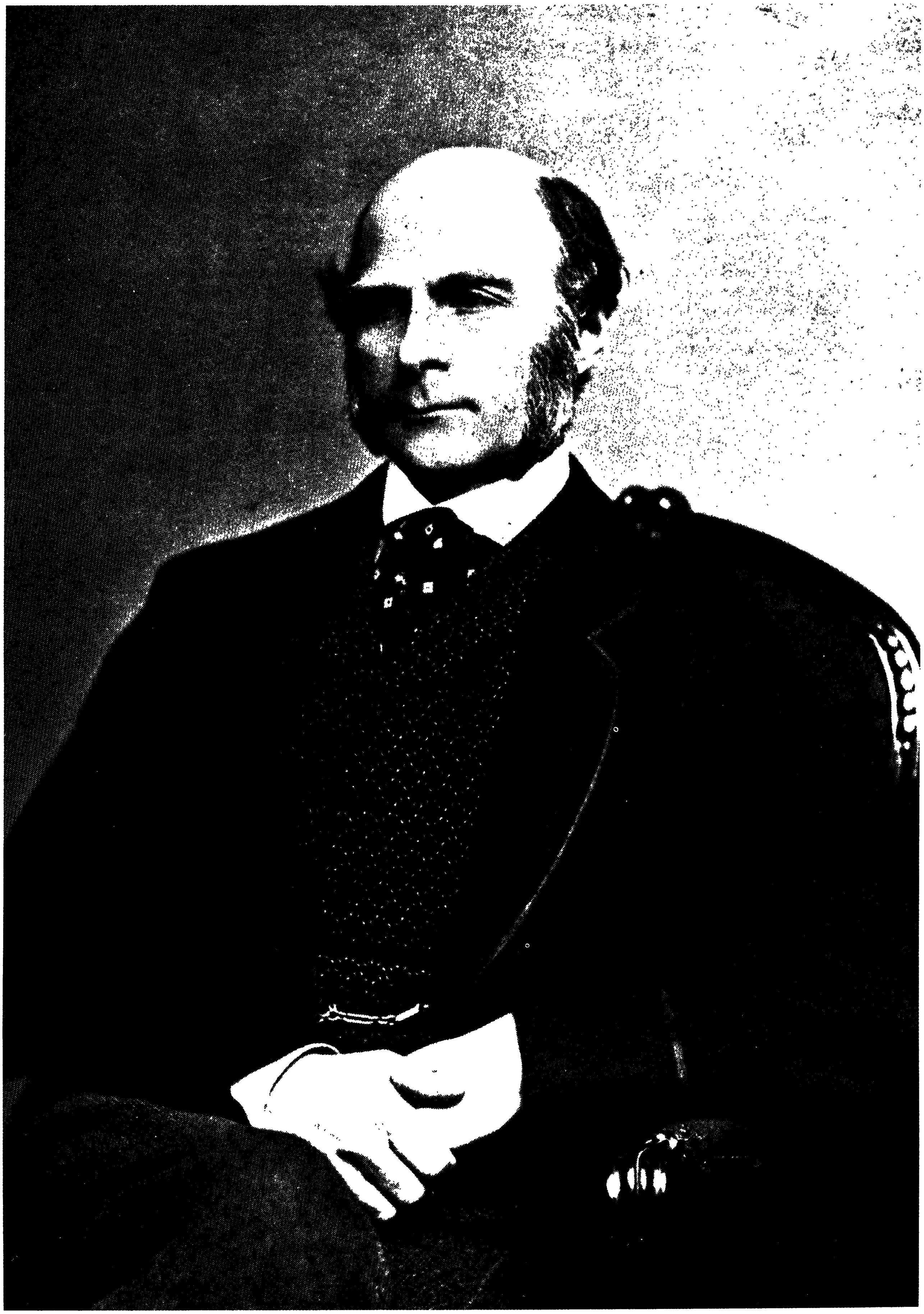
Few intellectual pleasures are more keen than those enjoyed by a person who, while he is occupied in some special inquiry, suddenly perceives that it admits of a wide generalization, and that his results hold good in previously-unsuspected directions. The generalization of which I am about to speak arose in this way.

In a book of mine called "Natural Inheritance," published about a year ago, I showed that the problems of family likeness fell entirely within the scope of the higher laws of chance; that we were thereby rendered capable of defining the average amount of family likeness between kinsmen in each and every degree, and of expressing the frequency with which the family likeness will depart from its average amount to any specified extent. It followed, very unfortunately for the general reader, that the problems of family likeness do not admit of being properly expressed except in the technical language of the laws of chance, and that it is impossible to discuss them adequately except through the medium of mathematics.

After the proofs of my book had been finally revised and had passed out of my hands, it happened that there was a delay of a few months before its actual publication. In the interim I was busily at work upon a new inquiry that had been suggested to me by two concurrent circumstances. One was a renewed discussion among anthropologists as to the information that the length of a particular bone—say a solitary thigh-bone dug out of an ancient grave—might afford concerning the stature of the unknown man to whom it belonged. It seemed to me that the anthropologists had not discussed their facts in the best statistical manner, and that they ought to have adopted a different form of treatment to any they had hitherto tried. The other circumstance arose out of the interest excited by M. Alphonse Bertillon, who proved that it was feasible to identify old criminals by an anthropometric process. The man who was suspected of having been convicted before was variously measured, and his measures were compared with those of all the criminals who had previously passed through the same process. By a contrivance analogous in principle to that on which a dictionary is constructed, the search through a register containing many tens of thousands of measures was performed with unexpected ease and precision.

Then a question naturally arose as to the limits of refinement to which M. Bertillon's system could be carried advantageously. An additional *datum* was no doubt obtained through the measurement of each additional limb or other bodily dimension; but what was the corresponding increase of accuracy in the means of identification? The sizes of the various parts of the body of the same person are in some degree related together. A large glove or shoe suggests that the person to whom it belongs is a large man. But the knowledge that a man has a large glove *and* a large shoe does not give us very much more information than if our knowledge had been confined to only one of the two facts. It would be most incorrect to suppose that the accuracy of the anthropometric method of identification increases with the number of measures in anything like the same marvellous rapidity that the security afforded by the better description of locks increases with the number of wards. The depths of the wards are made to vary quite independently of each other; consequently the addition of each new ward *multiplies* the previous security. But the lengths of the various limbs and bodily dimensions of the same person do not vary independently; so that the addition of each new measure adds to the security of the identification in a constantly-lessening degree. It seemed important, as well as highly interesting, to investigate this subject.

These two problems—namely, that of estimating the stature of an unknown man from the length of one of his bones, and that of the relation between the various bodily dimensions of the same person—are clearly identical. I was able to attack them at once, from happening to possess a sufficient number of sets of measures of different persons, each of whom had been measured in various ways. My first step was to take a large sheet of paper, ruled crossways; to mark a scale appropriate to the stature across the top and another appropriate to the left cubit (that is, the length from the bent elbow to the extended finger-tips) down the side. Then I began to "plot" the pairs of observations of stature and cubit in the same persons. Suppose, for example, an entry had to be dealt with of stature 69 inches, cubit 19 inches; then I should put a pencil mark at the intersection of the lines that corresponded to those values. As I proceeded in this way, and as the number of marks upon the paper grew in number, the form of their general disposition became gradually more and more defined. Suddenly it struck me that their form was closely similar to that with which I had become very familiar when engaged in discussing kinships. There also I began with a sheet of paper, ruled crossways, with a scale across the top to refer to the statures of the sons, and another down the side for the statures of their fathers, and there also I had put a pencil mark at the spot appropriate to the stature of each son and to that of his father.



Francis Galton

Reflection soon made it clear to me that not only were the two new problems identical in principle with the old one of kinship which I had already solved, but that all three of them were no more than special cases of a much more general problem—namely, that of Correlation.

Fearing that this idea, which had become so evident to myself, would strike many others as soon as "Natural Inheritance" was published, and that I should be justly reproached for having overlooked it, I made all haste to prepare a paper for the Royal Society with the title of "Correlation." It was read some time before the book was published, and it even made its appearance in print (Proc. Roy. Soc., Vol. 45) a few days the earlier of the two. Unluckily, through the hurry of preparation, I now find a sad number of numerical blunders in its tables, though none in the theory or formulas.

I hope to be able to give in this brief notice a just idea of the law of correlation, but it is quite out of the question to do more than explain its first and principal result. I trust it will soon be perceived by the reader that a great variety of important questions can be approached only through its methods.

The first step will be to explain the character of the connection that unites two related events; the next will be to show an unexpected consequence of relationship. Then the conditions will be pointed out under which mathematics may be applied to the discussion of related events, and one or two of the very important results to which they then lead will be described.

It is by the help of a succession of examples, rather than by a formal definition, that the nature of relation will be most quickly apprehended. Consider two men of the same race and country. Their remote ancestry, both human and prehuman, has been the same. There is, therefore, a considerable amount of identity in the sum of the influences under which they came into existence; there are also some few other identical events in the conditions of the climate in which they live, and even in the food they feed on. On the other hand, each of the men has been subjected to a variety of influences that have affected him separately and specially. In consequence, there is a certain likeness between the two men, intermediate between identity on the one hand and complete dissimilarity on the other. It is easy to express the average measure of this likeness in respect to any characteristic that admits of measurement. Stature will serve as an example: thus I found that, if any considerable number of couples of Englishmen are taken at random, the difference between the statures of the two men that compose each couple falls just as often below 2 inches and 4 tenths as above that amount. We may express the same fact in other language by saying that it is an even bet that the statures of two Englishmen taken at random will differ less than 2 inches and 4 tenths.

The relation between brothers is closer than this, because the number of identical influences that affect them is greater. The whole of their ancestry from their parents upward is the same. I found that the difference between couples of English and adult brothers fell as often below 1 inch and 4 tenths as above it.

Let us examine a little more closely the causes of the dissimilarity of brothers. There is room for a great difference in the circumstances of embryonic and preëmbryonic life, which may have helped to determine at each successive stage of incipient existence which one, out of the many conflicting possibilities of hereditary transmission, should become developed in either brother. Experience shows that the various qualities of ancestors do not blend equally in their descendants, but that the prevalence of one of these qualities, and the more or less complete exclusion of the rest, is a principal characteristic of fraternal dissimilarity. The final prevalence of a particular quality in each individual case may justly be ascribed to "accident," because the results, as I showed in my book, were disposed in conformity with the laws of chance.

I fear it is necessary to digress during a single paragraph in order to insist upon the scientific meaning of the words "accident" and "chance," which a rooted perversity of thought, among theologians principally, leads many educated men to misinterpret. There is nothing whatever in the idea to be attached to either word that is in any way contradictory to the regular course of cause and effect. Either word expresses the fact that at the moment when certain causes came into play the particular combination of the independently varying surroundings was such as to produce an unexpected effect. If the same combination of circumstances is experimentally repeated, the causes will again produce the same effect as before; but the recurrence of the combinations without predetermined arrangement is, judging from antecedent experience, so unlikely that a similar accident may never occur again.

The general character of the conditions of which we were just now speaking, that may have had an extremely important influence during the stages of incipient existence, may reasonably be supposed to be connected with the accidental positions of each several element, amid the swarm of ultimate elements, at the moment when any fresh stage of structure was impending. Little as is known about these invisible ultimate elements, it is ascertained, through the rapid changes in the internal appearances of the owner, that they move considerably among themselves during these early stages. Any one of the elements A, B, or C may be equally suitable to become a constituent of the incipient structure; but if it be impossible for more than one of them to enter into it, it is a fair hypothesis that the element which is at the moment accidentally nearest to the line of tension will be included, and the others thereby excluded. Such accidents as these may reasonably be supposed to have differently affected the form and structure of each brother separately, and to have been a chief cause of their observed diversities.

There are, moreover, many causes of a mixed character, neither wholly identical in their action upon the two brothers nor yet wholly different, but which may be treated as if they were divisible into their two contrasted groups without introducing a sensible error into the general problem. They are the greater or less similarity in food, climate, and early nurture.

It follows from all that has been said that the relation between the form and features of two brothers is the result of three groups of influences: (1) those that have alike affected both brothers; (2) those that have affected the first brother and not the second; (3) those that have affected the second and not the first. If there were no causes (2) and (3), the brothers would be identical; if there were none of (1), the brothers would have no likeness whatever, any more than that, say, of a brick to an elephant, or of a measure of hydrogen gas to a peacock. As it is, they are neither identical nor are they wholly unlike. They fall into the intermediate category of being related.

The following example, thought totally different in its details to that of kinship, affords nevertheless a true example of relation. Two clerks leave their office together and travel homewards in the same and somewhat unpunctual omnibus every day. They both get out of the omnibus at the same halting-place, and thence both walk by their several ways to their respective homes. We must further suppose that neither of the clerks has any fixed appointment or other reason for adjusting his pace of walking to the time of arrival of the omnibus, by hurrying when it is late, or dawdling in order to get rid of superfluous minutes when it is slow, but that each clerk "goes his own gait" quite independently both of the omnibus and of the other clerk. The upshot is that when either clerk arrives at his home later than his average time, there is some reason to expect that the other clerk will be late also, because the retardation of the first clerk may have been wholly or partly due to slowness of the omnibus on that day, which would equally have retarded the second clerk. Hence their unpunctualities are related. If the omnibus took them both very near to their homes, the relation would be very close. If they lodged in the same house and the omnibus dropped them at its door, the relation would become identity.

It must be clearly understood that relation only concerns itself with differences, and takes no note of total measures, nor of averages, except as a means of obtaining the required differences. Suppose the average time of the arrival of the first clerk was five o'clock, and that on a particular day he arrived at ten minutes past; it would be the ten minutes *plus* that alone concerns us. If the average time of arrival of the other clerk was fifteen minutes before four o'clock, and if he arrived on the same day at ten minutes before four, then he would be five minutes late; and it is this five minutes *plus* that we have to compare with the ten minutes *plus* of the other. Averages have no more to do with our present considerations than the position of the particular spot on the face of a white wall where a bull's-eye is painted for pistol practice has to do with the way in which the marks are distributed around the bull's-eye that are made by the shots aimed and fired at it. Departure is one thing, and the point departed from is another. The problems of kinship and correlation deal wholly with departures or variations; they pay no direct regard to the central form from which the departures or variations are measured. If we were measuring statures, and had made a mark on our rule at a height equal to the average height of the race of persons whom we were considering, then it would be the distance of the top of each man's head from that mark, upward or downward as the case might be, that is wanted for our use, and not its distance upward from the ground. In speaking of the couples of brothers, and of men of the same race who were not brothers, it was the differences of stature that were noted, and not the absolute statures. Differences of stature are identical in value with differences of the departure of either stature from the average of the race. It is, however, under the latter aspect that the mathematician has to consider it.

Fanciful examples like that of the two clerks are useful, because they thoroughly analyze the causes of relation. I will take another of the same kind of examples in order to emphasize the difference between relation and correlation, of which no explanation has thus far been attempted.

Suppose there are three commercial ventures *a*, *b*, and *c*, whose daily profits vary independently of one another, and that a certain investor, whom we will call R, has one share in *a* and another in *b*, while a second investor, S, has several shares in *a* and others in *c*. The total profits, day by day, of R and of S will be related together because they are partly due to an investment in a common concern, but they will vary on different scales, because the ups and downs of the profits of R, who has only one share in *a*, must be less wide than those of S, who has many shares. The estimate that we may (and shall) find it possible to make of the probable profit of S on any particular day, from a knowledge of the profit of R on that day, would not work backwards without modification.

There is not that *reciprocal* relation between them which is conveyed by the word correlation.

So in respect to the lengths of two limbs or other bodily dimensions of the same person that vary on different scales. A long finger usually indicates a tall person, and a tall person has usually a long finger, but by no means to·the same amount. There is relation between stature and length of finger, but no real correlation. On the other hand, the scale of variation of symmetrical limbs, such as that of the right and the left cubit, is so nearly the same that they can justly be said to be correlated.

The general conditions under which a relation between any pair of events will necessarily be established has now been very fully explained and illustrated. They consist in the concurrence of three independent sets of variable influences, which we have called (1), (2), and (3). The set (1) influences both events, not necessarily to the same degree; the set (2) influences one member of the pair exclusively; and the set (3) similarly influences the other member. Whenever the resultant variability of the two events is on a similar scale, the relation becomes correlation. When it is not the same, and when the variations are of the character shortly to be described as quasi-normal, a simple multiplication will be found to suffice (in a way that I may not now digress to explain) to transform the relation into a correlation. Thus we may speak of the length of the middle finger and that of the stature being correlated together under a recognized understanding that the variations are quasi-normal, and that the multiplication in question shall be made. Henceforth I will use the word correlation subject to these tacit understandings.

We will now apply ordinary common-sense, unaided by mathematical processes, to learn something about the results of relation. They are paradoxical at first sight, and are of the following description: they tell us that a *very* long thigh-bone should lead us to expect that the stature of the unknown man to whom it belongs was not *very* tall, but only tall. Conversely, the knowledge that an ancient worthy was a very tall man should lead us to expect that his thigh-bone would be not very long, but only long. To explain this we must go back to our three groups of causes, (1), (2), and (3), and let the two related events be called C and D, of which C is known and D is unknown. We want to learn something about the expectation that we ought to form concerning D. The size of C must be due either to the concurrent action of (1) and (2) acting together, both concurring in increasing C or both concurring in decreasing it, or else to the prevalence of one set over the other, when they are acting in opposite directions, the one tending to increase C and the other to diminish it. Now, a large departure occurs very much more rarely than a small one, and therefore it is very much more likely that a given departure should be built up of two lesser departures acting in the same direction than by the excess of a large departure over a small one.

It follows that it is much the most likely that set (1), if it acted singly, would produce a smaller departure than R, and this small contribution is all that D can get from set (1). D gets, *on the average*, nothing at all from set (3), because the total effect of that set is just as often in the direction of diminution as in that of increase. Consequently the *average* departure of S must be always less than that of R. In other words, the unknown S is probably *more mediocre* than the known R. Conversely, if S were known and R were unknown, the probability is that R would be more mediocre than S. The unknown brother of the very tall man is probably only tall; the unknown thigh-bone of the very tall man is probably only long; when one of the two clerks arrives home very late, the other clerk is probably only late; and so on. I have called this peculiarity by the name of regression. If there is no regression at all,—that is, if the regression is from 1 to 1,—then the correlation becomes identity. If the regression is complete,—that is, from 1 to 0,—there is no resemblance at all. In all intermediate degrees the ratio of regression is an exact measure of the weakness of the correlation.

We have now taken a general view of the nature of correlation and of its principal result; it remains to show that these general ideas admit of singularly exact interpretation in numerous important cases, and that problems can be worked out, and numerical calculations made, which in many cases admit of being verified by special sets of observation, and are then proved to be exact.

It is now beginning to be generally understood, even by merely practical statisticians, that there is truth in the theory that all variability is much of the same kind. The theory rests on the grounds that all variability is due to an uncounted number of small independent influences, acting variously in different cases. Mathematicians are able on these purely abstract grounds to develop a singularly beautiful law, known as the law of frequency of error. It is the basis of the higher statistics, and is founded upon such laws of chance as those which enable us to calculate the relative frequency of runs of luck of different lengths. The results are as precise as possible. It tells, for example, that if one-half of all the departures in a series of measures lie within 100 units of distance from the common average, three-quarters of them will lie within 171 units of distance. This kind of information is now readily to be obtained in all needed variety from well-known tables that have been calculated for the purpose, and which refer solely to what may be called the standard or the normal form of variability.

Now, when a series of measures are submitted to a competent statistician, it is a very simple matter for him to discover whether they vary normally or not. If they vary normally, then the series of measures is subject to all the numerous and beautiful properties that have been discovered in the law of frequency of error, and the tables just spoken of will apply rigorously to them. If they are quasi-normal, which is the common case, then the laws and the tables will be applied with caution and common-sense prudence; the more so, the more they depart from the normal type. Lengths of limbs vary with very fair approximation to the normal type. In what remains to be said I shall speak only of such variables as may be treated as normal.

A normal system of variables is clustered more closely about its centre than at a distance from it, and it fades away into nothingness on either hand through rapidly-increasing degrees of sparseness. One system differs from another only in its greater or less spread or dispersion. If every measure in the series that has the wider spread were uniformly shrunk, it could be made identical with the other. As soon as the scale of dispersion of a system of variables is known, the whole system is absolutely defined. For instance, we know that such and such a *percentage* of all the measures contained in it will be found between any two distances from its centre that may be named. It is extremely easy to measure the scale of dispersion in different ways that are all mutually convertible (one of which is to ascertain the so-called "probable error" of a single observation), but which I cannot digress to explain.

The numerical value of the scale of dispersion identifies a particular normal system just as completely as that of the length of a radius identifies a particular size of circle. Again, as circles have various properties and relations familiar to readers of Euclid, so normal systems of variables have their own peculiar properties, which enable numerous problems to be worked out concerning them, and make it possible to express in precise and definite language all that has been vaguely shadowed forth in the preceding pages about correlation.

For instance, it was said that the statures of a couple of Englishmen, taken at random, were equally likely to differ more or to differ less than 2 inches and 4 tenths. Theory teaches us that it follows from this that the stature of a single Englishman is equally likely to depart more or to depart less from the average stature of his race by that amount divided by the square root of 2, say by 1 and 4 tenths, which gives the result of 1 inch and 7 tenths. Observation confirms this.

A most interesting property of regression is brought into evidence by the theory of normal variability, and is fully confirmed by observation; namely, that the ratio of regression is unchanged, whatever may be the amount of the departure. In the case of brothers, the ratio of regression is as one to two-thirds. Therefore, if a man exceeds (or falls short of) the average of his race by one inch, one foot, or one decimetre, his unknown brothers will probably exceed (or fall short of) the average in question by two-thirds of an inch, of a foot, of a decimetre. In the case of a man and his son, the ratio of regression is as one to one-third, and similarly in the case of a man and his father. So we can now appreciate the completeness with which the ratio of regression measures correlation. A single value suffices to connect the whole of two systems.

When dealing with *correlated* dimensions of the same person, we must take their several scales of dispersion into the account. Thus in respect to the left middle-finger and the stature, observation showed that a departure of 1 inch in the finger was associated on the average with one of 8 inches and 19 hundredths of an inch in the stature; and that a departure of 1 inch in the stature was associated on the average with one of 6 hundredths of an inch in the finger. There is no numerical reciprocity in these figures, because the scales of dispersion of the lengths of the finger and of the stature differ greatly, being in the ratio of 15 to 175. But the 6 hundredths multiplied into the fraction of 175 divided by 15, and the 819 hundredths multiplied into that of 15 divided by 175, concur in giving the identical value of 7 tenths, which is the index of their correlation.

The purpose is now fulfilled that I had in view in writing this article, of giving a notion, that should be true as far as it went, of the chief law of correlation. Those who care to learn more about the subject may refer to what is said in the book and in the memoir already quoted, to which it is likely that I may be able to make additions before long.

The gain that has been now achieved is the discovery of the true and entirely unforeseen method of looking at correlation. The novelty of the idea is well exemplified by the question raised at the outset, of the thigh-bone and the probable stature of the man to whom it belonged. The old notion was that, the average length of the bone being so and so, and that of the stature of men of the same race being so and so, then if the bone were, say, a twentieth part longer than the average of such bones, the stature of the man to whom it belonged should be estimated at one-twentieth more than the average stature (subject to certain corrections). This we now perceive to be doubly erroneous in principle. We have nothing to do with twentieths or other fractional parts of the average length, and there exists no direct proportion between the total lengths of the bone and of the actually associated stature. The idea of regression being a factor in these relations has been hitherto quite unsuspected by anatomists. We now see that it necessarily plays an essential part in them, and that its value affords an admirable measure of the closeness, or weakness, of correlation between any two series that severally vary in a quasi-normal manner. We can also construct tables similar in form to those spoken of in the earlier part of this article, wholly by calculation from the following five *data*: namely, the averages and the scales of dispersion ("probable error") of either of the two quasi-normal series, and the ratio of regression from either of them to the other.

There seems to be a wide field for the application of these methods to social problems. To take a possible example of such problems, I would mention the relation between pauperism and crime. I have not tried it myself; but it is easy to see that here, as in every case of relation, success would largely depend on finding quasi-normal series to deal with. Both pauperism and crime admitting of many definitions, it would be necessary to restrict the meanings of those words for the purpose of the inquiry, so that the cases to be dealt with shall be fairly homogeneous in respect to all important circumstances. To do this is the business of the statistician, who becomes assured of the soundness of his judgment in devising his restrictions when he finds that his statistics are of a quasi-normal character. If he is able to succeed in this task in the present problem, the relation between pauperism and crime would be rigorously expressed by the simple methods already explained.

In conclusion I must repeat what was said before, that it is impossible to go deeper into this subject without using very technical language and dealing freely with conceptions that are, unhappily, quite unfamiliar to the large majority of educated men. I can only say that there is a vast field of topics that fall under the laws of correlation, which lies quite open to the research of any competent person who cares to investigate it.