Qual Quant (2008) 42:757–777
DOI 10.1007/s11135-006-9066-y

ORIGINAL PAPER

The historical construction of correlation as a conceptual and operative instrument for empirical research

Juan Ignacio Piovani

Published online: 25 January 2007

Abstract This article is meant to reconstruct—from the standpoint of sociology and history of science—the development of the concept and the operative instruments of statistical correlation. The starting point is the discussion of some key mathematical aspects of the Error Theory, including a detailed analysis of the various positions regarding its contributions, if any, to the theory of correlation. Then proceeds to examine how the concept (and its relative instruments) emerged in its modern sense, by the late Nineteenth century, thanks to the work of Francis Galton. Finally, it considers the numerous contributions that rendered possible the formalisation and generalisation of both Galton's concept and methodological tools, in particular those of Karl Pearson, but also those of Walter Weldon, Francis Ysidro Edgeworth, George Udny Yule and Charles Spearman.

> "Co-relation or correlation of structure" is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree.
>
> Francis Galton (1888, 135)

1 The error theorists and the mathematical foundations of correlation

The name of Francis Galton is inextricably bound up with the theory of correlation. In fact, he is usually regarded as its mentor. Nonetheless, it is difficult to consider any intellectual enterprise, even one regarded as revolutionary, as not rooted in tradition, and so, as is only to be expected, many historians of Statistics, starting with Karl Pearson, have attempted to trace the traditional roots of correlation. But individuat-

J. I. Piovani (✉)
National University of La Plata, Calle 48 entre 6 y 7, 1900, La Plata, Argentina
e-mail: jpiovani@unibo.edu.ar

ing a sole intellectual antecedent has been difficult, since it too can always be related to prior developments.

Pearson (1896: 261) offers one of the earliest reconstructions of the origins of correlation, singling out the essay, *Analyse Mathématique sur les Probabilités des Erreurs de Situation d'un Point* by Bravais (1846) as the place where the notion was first discussed. Further developed by Yule (1897b, 1909), Pearson's idea then spread with the publication of the famous handbook, *An Introduction to the Theory of Statistics* (Yule 1911). In 1920, Pearson presents an alternative version of the story, tracing the origins of correlational calculus back to Gauss's essay, *Theoria Combinationis Observationum Minimis Obnoxiae* (Gauss 1823), in which "the normal surface of *n*-correlated variates" was somehow reached (Pearson 1920: 27). Denis (2000) argues instead that the mathematical roots of the technique are to be found in Adrien Marie Legendre's work on least squares, published in his *Nouvelles Méthodes pour la Détermination des Orbites des Comètes* (Legendre 1805).

All apparently agree that the mathematical instruments—the starting point for correlation—are founded on the intellectual tradition known as error theory, to which all the above-mentioned authors belong. However, a comprehensive consensus as to the concept's main precursor has yet to be reached. In the following paragraphs I will analyse the various positions.

The attribution of paternity to Bravais is probably the most widespread. He had already received credit for discovering bivariate and trivariate normal distributions when Galton was still working on the issue:

> The fundamental theorems of correlation were for the first time and almost exhaustively discussed by Bravais […] He deals completely with the correlation of two and three variables […] The 'Galton's function' or coefficient of correlation […] indeed appears in Bravais' work, but a single symbol is not used for it (Pearson 1896: 261).
>
> Of the numerous memoirs on the theory of error the most important in […] connection [to correlation] is that of A. Bravais, who as long ago as 1846 discussed the theory of error for points in space, regarding the errors as either independent or correlated, from the standpoint of the normal law of errors. He did not, however, use a single symbol for a correlation coefficient, although the product-sum[1] formula may be regarded as due to him (Yule 1909: 722).

Similar viewpoints are expressed by Walker (1929), Westergaard (1932), Seal (1967) and Lancaster (1972). All the same, these authors began to notice that Bravais was not aware of the consequences of his formula for the treatment of the relations between variables: "Bravais […] was not thinking of the strength of the relation between the variables, nor did he ever single this out for attention" (Walker 1929: 482).

In 1920, Pearson wrote a memoir entitled *Notes on the History of Correlation*, reconsidering his original appreciation of Bravais's work. In addition, he presented an explanation of the process by means of which Bravais came to be considered the father of correlation. Pearson was reportedly captivated and "immensely excited" by Galton's work on correlation and heredity, and consequently he began to develop the concept and its relative instruments: "the field was very wide and I was far too excited to stop to investigate properly what other people have done. I wanted to reach new results and apply them […] and when I came to put my lecture notes on correlation

[1] Yule uses the expression 'product-sum' instead of the more frequent 'product-moment'.

into written form, probably asked somebody […] to examine the papers and say what was in them (Pearson 1920: 29). It was only many years later that he had the time to go back to Bravais's memoir, realising that he had advanced misleading interpretations later spread by the text-book writers, particularly his former assistant Yule.

Reviewing Bravais's essay, Pearson (1920: 29 Ss.) arrived at the conclusion that his objective was solely to measure the errors in the determination of the coordinates x, y, z of a point in space. These coordinates were not measured directly, but rather were functions of the observed elements a, b, c. Furthermore, Bravais assumed that these observed quantities were not correlated in the modern sense of the term. Pearson argues that there was nothing of any substance in Bravais's work that could not be found in a preceding memoir by Gauss (1823), whose problem was to express the variability of x in terms of the observed quantities a, b, c…and of the differential coefficients A, B, C. This was precisely Bravais's problem, and the only really valuable aspect of the latter's analysis lay in the deduction and discussion of the "properties of a surface of which the contours are […] the familiar ellipses of [the] normal surface", also arriving at a line corresponding to Galton's regression line. However, this was not "a result of observing x and y and determining their association, but of the fact that x and y [were] functions of certain independent […] quantities". Bravais remained loyal to the Gaussian tradition; the observed quantities were considered *absolutely independent*, and he had no notion of their possible correlation. The product terms of his expressions arose not out of the organic relationships between quantities directly observed, but rather from the geometrical relationships which existed between the observed quantities and those deduced. In short, Pearson's conclusion was that Bravais did not make any significant contribution to the subject of correlation.

According to MacKenzie (1981: 69–70), Bravais's innovation consisted of his attempt to estimate the position of a point on a two-dimensional plane or in three-dimensional space.[2] "The form of his joint law of error […] is identical to that constructed by Galton and Hamilton Dickson" (see Sect. 2), but he "had only the *form* of his law". In order to fully discuss the dependence of x and y, he should have proceeded to analyse the coefficients. But his reference to the correlation of x and y was not followed by any attempt to study or to measure it, which, basically, for him would not have made sense.

For Seal (1967) and MacKenzie (1981) an essay concerning a problem analogous to Bravais's, written by Schols and published in Holland in 1875,[3] merits a somehow different evaluation. He did not assume independence; actually, he thought that the errors in the coordinates of a point in space were the result of a large numbers of small errors. Nonetheless, the axes of the coordinates could be determined *as if* the errors were independent. Having reached a conclusion appropriate to his interests, Schols "made no attempt to formulate an expression for the degree of influence of the error in one direction on that in another" (MacKenzie 1981: 71).

Bravais's and Schols's work must be assessed in the light of the scientific tradition prevalent in their time; in other words, they represented a given normal science—in Kuhn's (1962) sense—in which variability and correlation were not yet legitimate questions for consideration. In fact, the then-dominant astronomic paradigm regarded

[2] Walker (1929) considers the American astronomer Adrain to be the first who dealt with the probability of occurrence of two simultaneous errors in the position of a point. Moreover, she comments on Laplace, Plana and Gauss's contributions to this very issue. Gauss's work, as already mentioned, had been highlighted also by Pearson (1920).

[3] In the references a French translation of 1886 is reported.

the determination of errors of measurement as its fundamental problem. To this end, it relied on a theoretical approach based on the mathematical tradition of probability calculus, materialised in the error theory and one of its main tools, normal distribution. Its orientation, which was rather formal and deductive, was closer to French rationalist tendencies. Accordingly, "for neither of them was statistical dependence in itself a focus of attention" (MacKenzie 1981: 71).

Those who defend Bravais's role as precursor of correlation usually stress his (apparent) consideration of this specific issue, while those who propose the names of Legendre and Gauss point instead to their development of the fundamental mathematical tools for correlational calculus—in particular, the method of least squares—and not to correlation itself. In his reconstruction of pre-Galtonian contributions to correlation, Pearson (1920) acknowledges its close connection to the method of least squares. This point had already been expressed by Yule:

> The method of correlation is only an application to the purposes of statistical investigation of the well-known *method of least squares*. It is impossible, therefore, entirely to separate the special literature of the theory of correlation from that of the […] method of least squares (Yule 1909: 722).

So, the controversy over Legendre's or Gauss's priority in the history of correlation is simply a reflection of the dispute surrounding the origins of the method of least squares. All too often authors not interested in this issue tend to portray Legendre as its inventor, commenting only marginally, if at all, on the above-mentioned controversy (see, for example, Denis 2000).[4] Other authors suggest instead that the method did not come into being suddenly and assess Legendre's work in the light of some of the most relevant scientific problems of his time (see, for example, Micheli and Manfredi 1995). Stigler (1986) regards this method as the culmination of certain developments initiated during the mid-eighteenth century, and considers that even though Legendre was the first to communicate it effectively, Gauss—as he himself emphatically claimed—was probably already using it by the end of the century (Stigler 1981, 1999).

Leaving aside the ticklish issue of who was the inventor of least squares, it seems clear that the method was developed within the tradition of error theory sometime between the end of the eighteenth century and the beginning of the nineteenth in connection with the measurement problem in astronomy, which was already relevant by the mid-eighteenth century. Astronomers acknowledged the unlikelihood of measuring an object with complete accuracy. However, any measurement could be repeated, and soon they realised the usefulness of repetition for reducing error and estimating its probable amount (MacKenzie 1981). At that time it was customary to employ the arithmetic mean for the determination of a quantity from various measurements (with values typically assuming the shape of a normal distribution). The estimate error of this determined quantity was called the 'probable error':[5] "A deviation from […] the mean of the observations, of such magnitude that, if the number of observations be indefinitely increased, one half of the errors may be

[4] This is not necessarily due to a lack of historical sensitivity, but to the fact that this issue might have been rather marginal to the authors' interests.

[5] According to Walker (1929), the expression 'probable error' was introduced by German astronomers at the beginning of the nineteenth century. For Denis (2000) it can be traced back to the mid-eighteenth century.

expected to be numerically greater and one half numerically less than this value" (Walker 1929: 50).

Nevertheless, at the beginning of the eighteenth century there was not yet a systematic method for "balancing" the positive with the negative deviations of each measurement to the mean. The solution proposed by Legendre, in a section, entitled *Sur la Méthode des Moindres Quarrés* of his book published in 1805, consisted of the sum of the squared deviations to the mean. It was a mathematically simple method assuring a coherent result, always providing the "best" solution in terms of error reduction and symmetry (Denis 2000). The method was immediately accepted in astronomy and later spread to other disciplines, including the human sciences (Stigler 1986). Surprisingly, notwithstanding its rapid spread, Galton was ignorant of the method proposed by Legendre, as well as of Bravais's work (Pearson 1896, 1920).

As shall be illustrated in the following section, the influence of error theory on the Galtonian development of regression and correlation only manifests itself in the use of the normal curve and the concept of probable error for the purposes of analysing variation. This latter aspect implied a change in perspective: the error theorists worked predominantly with univariate distributions, or, at most, with mutually independent variables, in order to reduce measurement error. Galton, on the contrary, was genuinely interested in variability: his concepts of regression and correlation provided a revolutionary general approach to the treatment of dependency between variables. The existing technical instruments of least squares (Legendre and Gauss), and the bivariate normal surface (Bravais) thus became reinterpreted in the light of this new perspective.

Therefore, even if the error theorists developed mathematical tools formally identical to those of modern regression and correlation, they cannot be regarded as a source of inspiration for Galton—who did not know of the existence of these developments—nor can they be considered to have pioneered the formulation of a problem that science had yet to acknowledge. It is evident that Galton's work did not emerge from a vacuum: his intellectual antecedents are part of a far more intricate and complex story, only some aspects of which error theory—privileged by the concreteness of its instrumental contributions—can explain. Although less obvious, other influences such as the tendency towards quantification in empirical research, the work of Quetelet and his use of statistical tools in the human sciences, and finally the British naturalist tradition—in particular Darwin's investigations—are not necessarily less important. And yet, even if reconstruction be broadened to include these and other intellectual trends of the times, the emergence of the modern concept of correlation cannot be fully comprehended without due appreciation of Galton's original contribution: the selection of variation as a legitimate problem for scientific research.

2 A true turning point: Galton's concepts of regression and correlation

From the late 1860s to the 1880s, Galton worked on statistical problems in relative isolation, while achieving "a theoretical breakthrough of enormous significance" with the concepts of regression and correlation (MacKenzie 1981: 10).

These developments were intimately bound up with his empirical research on the subject of heredity. There are several hypotheses regarding the origin of this research interest that marked a shift towards evolutionist biology and anthropometrics and

away from former concerns in the fields of geography, naturalist exploration and meteorology. Micheli and Manfredi (1995: 106) point to a personal issue: the nervous breakdown that followed a sterile marriage may have prodded Galton into the study of heredity processes. MacKenzie (1981: 58) proposes a eugenic explanation: the interest in the hereditary transmission of human characteristics was congenial with Galton's project on racial progress. Others generically trace this interest back to the influence of the theory of evolution advanced by Darwin, whose main book *On the Origins of Species* (Darwin 1859) was read by Galton in the 1860s. Actually, his comments on this book by his much-admired cousin show how Darwin's new perspective had encouraged him to conduct quite a few studies focussing on heredity and the possible improvement of the human race (Tankard 1984).

The first result of this new orientation was his book *Hereditary Genius* (Galton 1869), in which Galton followed Quetelet's approach in the use of error theory for research on human beings. Galton's mathematical competence was not extremely refined, but "the basic techniques of error theory were widely known and used in Britain, and it was natural that Galton should turn to them when seeking statistical tools, particularly since […] Quetelet had already successfully applied them to human data" (MacKenzie 1981: 57). A few years later Galton would actually argue:

> The higher methods of statistics, which consist of applications on the law of Frequency of Error, were found eminently suitable for expressing the processes of heredity (Galton 1889: 192–193).

And even earlier, in a presidential address to the Anthropological Institute, he had affirmed:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'law of error'. A savage, if he could understand it, would worship it as a god. It reigns with serenity in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the anarchy the more perfect is its sway. Let a large sample of chaotic elements be taken and marshalled in order of their magnitude, and then, however wildly irregular they appeared, an unsuspected and most beautiful form of regularity proves to have been present all along (Galton 1886: 494–495).

In *Hereditary Genius* Galton proposed no new statistical tools. However, he "envisaged the development of a predictive, quantitative theory of descent", not yet arriving at a "mathematical law connecting parent and offspring generations" (MacKenzie 1981: 59–60):

> I do not see why any serious difficulty should stand in the way of mathematicians, in framing a compact formula, based on the theory of Pangenesis, to express the composition of organic beings in terms of their inherited and individual peculiarities, and to give us, after certain constants had been determined, the means of foretelling the average distribution of characteristics among a large multitude of offspring whose parentage was known (Galton 1869: 371–373).

Galton was perplexed by a fact for which he could still find no reasonable explanation: "if the hereditary make-up that an individual acquires is the result of an enormously large number of factors, all independent from one another [as follows from the error theory] how come a single identical character repeats itself from parent to

child to grandchild without changing at all?" (Micheli and Manfredi 1995: 106). This problem led him to a turning point that, at that time was merely argumentative and rather vague. Nevertheless, Galton began to notice the inadequacies of error theory for his scientific purposes. In his autobiography, he recollects the issue as follows:

> The primary objects of the Gaussian Law of Error were exactly opposed to those to which I applied them. They were to get rid of, or to provide just allowance for errors. But these errors or deviations were the very things I wanted to preserve and to know about (Galton 1908: 305).

According to MacKenzie (1981), Galton's departure from the traditional standpoint of error theory becomes evident in *Statistics by Intercomparison* (Galton 1875). In fact, in this memoir he refers to the concept of error as "absurd": in his opinion, to define an extraordinarily capable man as an enormous error of nature was simply ridiculous. Nonetheless, the interesting aspects of this essay, as regards what the future of correlation would be, go far beyond the critique of the idea of error. Above all, there is another clear position that distanced him from the precedent orientation. The error theorists followed a typically rationalist and deductive perspective. Galton claims instead to pay more attention to observations, anchoring his conclusions in the analysis of "empirical" distributions, rather than aprioristic deductions:

> The law of frequency of error says that 'magnitudes differing from the mean value by such and such multiples of the probable error, will occur with such and such degrees of frequency'. My proposal is to reverse the process, and to say, 'since such and such magnitudes occur with such and such degrees of frequency, therefore the differences between them and the mean value are so and so, as expressed in units of probable error (Galton 1875: 37–38).

Another aspect to highlight is that, by this time, Galton was already interested in developing "a method for obtaining simple statistical results which has the merit of being applicable to a multitude of objects lying outside the present limits of statistical inquiry, and which […] may prove of service in various branches of anthropological research[6] (ibi: 33).

In addition, he explains in detail the measures of central tendency and of variability that a few years later would be used for calculating regression and correlation: the mean—which, according to his definition, would be the modern median[7]—and the interquartile deviance (then called the 'probable error').

In fact, his mean "represents the *mean* value of a series in at least one of the many senses in which that term can be used". For example, "the mean speed of a number of horses would be that of the horse which was the middlemost in the running". Once the mean is determined, "the next great point to be determined is the divergency[8]

[6] Certainly Galton was not referring to correlation, but to a method for comparing distributions that he had already proposed and, up to a certain extent, employed in *Hereditary Genius* (Galton 1869: 26). This search for simple and "universally" applicable methods would always be present in Galton's statistical work.

[7] Since Galton assumes normally distributed populations, the mean and the median coincide. However, empirical observations do not always follow this theoretical pattern, and accordingly, later contributors to the theory of correlation noted the problems arising from the use of the median instead of the arithmetic mean in the calcululation of correlation (see for example Weldon 1892; Pearson 1896). In truth, Galton himself acknowledged that symmetrical distributions, according to his empirical experience, were rather rare (Galton 1875: 35).

[8] Galton uses the word 'divergency' instead of the current statistical term 'deviation'.

[…]—that is, the tendency of individual objects […] to diverge from the mean value of all of them. The most convenient measure of divergency is to take the object that has the mean value, on the one hand, and on the other, those objects whose divergence in either direction is such that one half of the objects […] on the same side of the mean diverge more than it does, and the other half less". The difference between the mean and either of these objects is the measure in question, technically […] called the 'probable error"' (ibi: 34–35).

Two years after the publication of *Statistics by Intercomparison*, Galton delivered a lecture at the *Royal Institution* with the title *Typical Laws of Heredity*.[9] "In retrospect, this […] can be seen as the first stage of Galton's revolution in statistical theory: his first development of the concept that was later to be called linear regression" (MacKenzie 1981: 63). To conduct his research on heredity Galton would have preferred to use anthropometric data,[10] but the difficulties in obtaining them drove him to employ measurements performed on two generations of sweet pea seeds.[11]

Galton started by taking a large sample of seeds that were weighted in order to calculate the mean and the probable error. Then he separated the seeds into seven groups, each one containing seeds of the same weight: one packet had those of the mean value, and the other six had seeds one, two and three times either heavier or lighter than the probable error of the mean. Galton distributed the seeds among his friends, who had to grow them and bring back the descendants. Analysing the data, he discovered that the descent of each class of original seeds had an analogous variability,[12] and that they were distributed normally. Moreover, he realised that the mean of the descendants of the smaller and bigger seeds were not as extreme as that of their predecessors, and that the relation between the weight of the first and second generations was linear (see Fig. 1). In this way he arrived at the concept of 'reversion',[13] later renamed as 'regression':

> 'Reversion' is the tendency of the ideal mean filial type, reverting to what may be roughly and perhaps fairly described as the average ancestral type (Galton 1877: 10)[14]

Encouraged by the results, Galton decided to replicate his research scrutinising the transmission of human traits. In the late 1870s and early 1880s he patiently gathered this kind of data in his Anthropometric Laboratory in South Kensington. In 1884, after offering prizes in return for family records, he managed to obtain the long-desired empirical material.

Galton chose to investigate stature since "it was easy to measure, relatively constant during adult life, its distribution closely followed the law of frequency of error"

[9] Later published in *Proceedings of the Royal Institution* (see Galton 1877).

[10] Galton mentioned this detail in a report read before the *Anthropological Section* of the *British Association* (see Galton 1885).

[11] Note that by then Galton did not know Mendel's work, which spread in British intellectual circles at the beginning of the twentieth century.

[12] This is no less than the modern idea of homoscedasticity (see Pearson 1920: 33).

[13] It may seem that Galton arrived at the "law of reversion" in a purely empirical way, by simply observing the data and deriving the law from them. According to MacKenzie this was rather unlikely: "Galton had a definite prior notion of the kind of law he was looking for: a simple, predictive mathematical statement of the relationship between parent and offspring generation. There is indeed reason to believe that his data did not unequivocally 'suggest' the law of reversion" (MacKenzie 1981: 63).

[14] Quoted in Pearson (1920: 33).

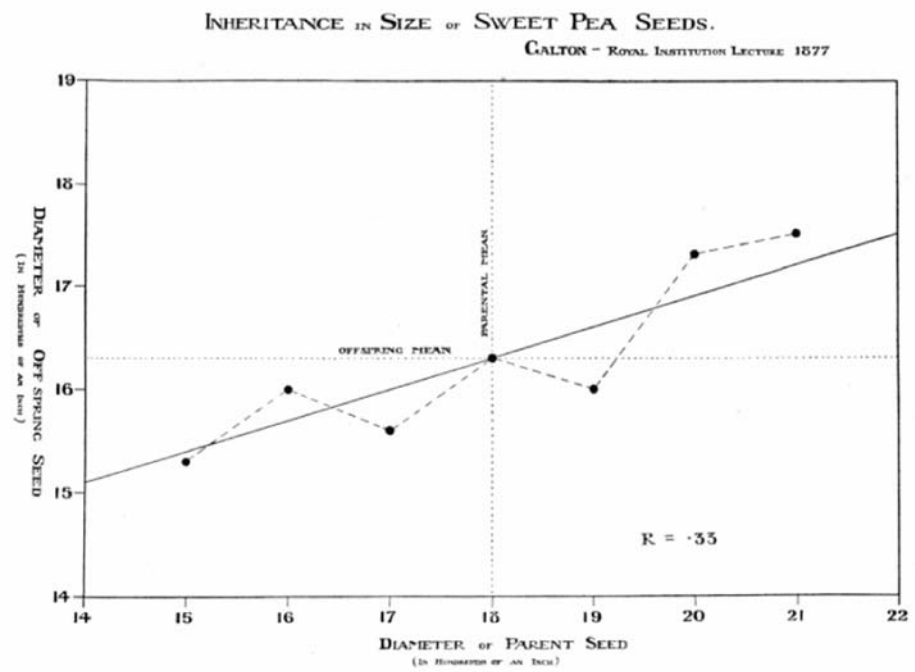

Fig. 1 First regression line drawn by Galton according to Pearson's reconstruction. Source: Pearson (1920: 35, 1914–1930 vol. 3, chapter 14)

(MacKenzie 1981: 64). Galton had data regarding the height of adults and their offspring. For each descendant he calculated what is known as the 'mid-parent', which is the mean of stature of the father and that of the mother multiplied by 1.08. He then examined the relationship between the stature of the offspring and that of the mid-parent. He realised that these human data followed the same model as the seeds, but in a clearer fashion. The first results were communicated in 1885 in his presidential address to the Anthropological Section of the *British Association*. This is a work of fundamental historical relevance: in Pearson's opinion, it contains the first-ever published diagram of the two regression lines, the first correlation table and the figure showing the graphical way in which he reached—based on observations—the bivariate normal surface (Fig. 2).

In fact, Galton had not only limited himself to the exact reproduction of his previous work on seeds; he had gone one step further. Comparing the frequency distributions, he noticed some strange patterns. Using the plotting system already employed in previous research, he drew horizontal and vertical lines. The intersection of each pair of lines defined four adjacent squares. Galton wrote the corresponding frequency in each one of them, and then drew contour lines connecting all the frequencies of equal value. To his great surprise, he "found such contour lines were a system of concentric similar and similarly placed ellipsoids and that the regression lines were what the mathematician terms the conjugate diameters of the variate axes" (Pearson 1920: 36–37). Moreover, the centre of the whole system of ellipsoids was the intersection of the lines representing the mean statures of the parents and the offspring (see Fig. 2).

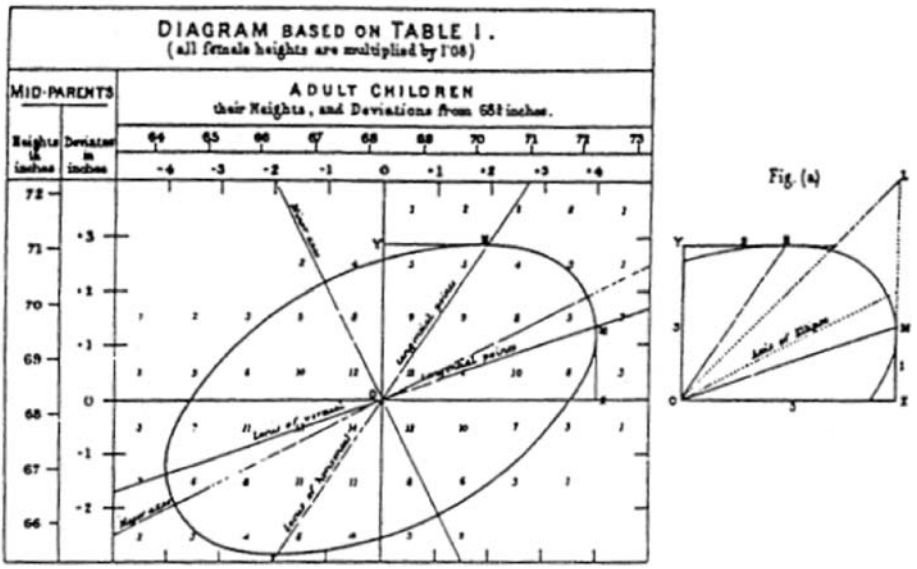

Fig. 2 Joint distribution of the parent and offspring stature showing one of the elliptic concentric contour lines drawn by Galton, and its geometrical relation with the regression lines and the axes of the ellipse. Source: Pearson (1920: 36, 1914–1930 vol. 3, cap. 14)

MacKenzie (1981: 65) argues that "Galton guessed that these patterns might be the clue to a deeper understanding of regression". But he could not yet see his way to express the results in their entirety in a single formula. In his memoirs, Galton remembers the unforeseen emergence of the solution:

> At length, one morning, while waiting at a roadside station near Ramsgate for a train, and poring over the diagram in my notebook, it struck me that the lines of equal frequency ran in concentric ellipses. The cases were too few for certainty, but my eye, accustomed to such things, satisfied me that I was approaching the solution. More careful drawing strongly corroborated the first impression (Galton 1908: 302).

Galton tried to develop an equation in order to represent the joint frequency surface described by the ellipses. Uncertain of his own mathematical skills, he sought the collaboration of Hamilton Dickson, a well-know Cambridge expert. The solution, analogous to that proposed by Bravais in 1846 (see Sect. 1), deeply excited Galton:

> I never felt such a glow of loyalty and respect towards the sovereignty and magnificent sway of mathematical analysis as when his answer[15] reached me, confirming, by purely mathematical reasoning, my various and laborious statistical conclusions with far more minuteness than I had dared to hope, for the original data ran somewhat roughly, and I had to smooth them with tender caution (Galton 1885: 255).

Hamilton Dickson's formulae were included as an appendix to Galton's essay *Family Likeness in Stature* (Galton 1886). Beyond this mathematical formalisation,

[15] He refers to Hamilton Dickson's answer.

the essay presents no innovations in the theory of correlation. The only curious aspect is that Galton abandoned temporarily the notation *r* (already symbolising "reversion" since 1877) and introduced ω.

The next fundamental step was taken in 1888, when Galton begun to realise that the statistical tools derived from his research on heredity had far-reaching implications, and set about developing the concept of correlation. This step from regression to correlation may seem obvious today, but at the time, a new impetus was needed: the stimulus came from the system of personal identification devised by the French criminologist Alphonse Bertillon, which was based on anthropometric measures (MacKenzie 1981; Porter 1986). In his famous memoir *Co-relations and their Measurement, chiefly from Anthropometric data* (Galton 1888), Galton confronted the issue[16] and used the term 'correlation'[17] in its modern statistical sense for the first time.

As a matter of fact, Galton did not invent the term; it was customarily used in the scientific literature of the second half of the nineteenth century. Grove (1846) and Carpenter (1851) made use of it to define the relationships among various natural forces (Winter 1997). In biology it denoted the principle of interdependence of organs: for instance, Darwin (1868), writing about animals and plants, described the interconnections of the different parts of a whole as "correlated together". In his copy of Darwin's book, Galton had underlined the expressions "are to a certain extent" and "so slight",[18] both referring to possible characteristics of this correlation. In *Principles of Science* Jevons used the term correlation, in a sense quite different from the customary usage at the time, as implying things that are so closely related that "*where one is the other is, and where one is not the other is not*" (Jevons 1874: 354). In Galton's personal copy of this book, kept in the *Galton Archives*, this passage is marked and in the margin is written "Nice wd. never so with the common meaning (Grove's)",[19] obviously referring to correlation. According to Stigler (1978) Galton borrowed the term from Jevons, even if the latter opposed the idea of statistical correlation and doubted its applicability. It should be clear though, as Stigler (1978) and Nicholls (1998) suggest, that Jevons' definition *inspired* Galton to introduce the term in its *modern statistical sense*, since he had already used it in some works published before Jevon's book (for example in his 1869 *Hereditary Genius*), and by the early 1870s he was fully acquainted with the biological senses of correlation proposed by Cuvier and Darwin (Porter 1986). Not surprisingly, Galton begins the above-mentioned essay of 1888 with this statement:

> 'Co-relation or correlation of structure' is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequently present than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action in detail, or to show how to measure its degree (Galton 1888: 135).

And a few lines later he actually defines the term:

> Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the

[16] Galton had gathered a vast database in order to study the correlations among the different parts of the human body. To write his article he worked on a sample of 350 males aged 21 or over, obtained at his Anthropometric Laboratory (1888: 136).

[17] Actually, Galton first used the term 'co-relation', later 'correlation'.

[18] Quoted in MacKenzie (1981: 66).

[19] Quoted in Stigler (1978: 288).

same direction.[20] Thus the length of the arm is said to be co-related with that of the leg, because a person with a long arm has usually a long leg, and conversely (*ibidem*).

Galton noticed that the problem of the relationship between different parts of the human body was similar to others already studied, in particular to that of hereditary mechanisms, and concluded that these analogous issues exemplified an even more general question: correlation.[21] His conclusion materialised in an article published a few months later entitled *Kinship and Correlation* (Galton 1890). Galton presented it as follows:

> 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 belt elbow to the extended fingertips) down the side. Then I began to 'plot' the pairs of observations of stature and cubit in the same persons […]; then 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 […] 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 (Galton 1890: 420).

If variations "were wholly due to common causes, the co-relation would be perfect […] If they were in no respect due to common causes, the co-relation would be *nil*.[22] Between these two extremes are an endless number of intermediate cases, and […] the closeness of co-relation in any particular case admits of being expressed by a simple number" (Galton 1888: 135–136). However, this number—the measure of correlation—should be the same regardless of which variable was taken as independent, and Galton knew by then from his research into stature that the regression coefficients did not satisfy this requisite. Since this was due to the different probable errors in each variable, in order to find an appropriate coefficient, it was necessary to eliminate the influence of their respective scale units.

After having re-scaled the variables, dividing each of them by its own probable error, "Galton found that some simple relationships held. Either variable regressed linearly on the other, and the coefficients of regression were equal […] Galton called

[20] Note that Galton was not aware of negative correlation. In addition, it was not Galton but Weldon who conceived the idea of negative correlation.

[21] In his autobiography, Galton recalls the circumstances in which the idea of correlation first emerged. He was rambling around the grounds of Naworth Castle when "a temporary shower drove [him] to seek refuge in a reddish recess in the rock by the side of a pathway", and there the idea of correlation "flashed across [him], and [He] forgot everything else for a moment in [his] great delight" (Galton 1908: 300).

[22] For Galton this was better understood by means of an example: "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" (Galton 1890: 423).

the mutual value of the coefficients of regression r" (MacKenzie 1981: 67), reintroducing the notation of 1877 that he had abandoned in 1886.

How should the re-scaling be performed? Galton provided a clear example for the length of the left middle finger and 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 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" (Galton 1890: 429–430).

The fundamental characteristics of this "index of correlation" between two variables [y and x], according to Galton, are four: "(1) that $y = rx$ for all values of y; (2) that r is the same, whichever of the two variables is taken for the subject; (3) that r is always less than 1, and (4) that r measures the closeness of co-relation" (Galton 1888: 145).

In 1889, Galton published *Natural Inheritance*, probably his most significant book and corollary of all his previous work on cases of regression and correlation in heredity. He introduced no new developments in this book, but rather presented his concepts in an articulate and systematic fashion, attracting the attention of those who would later formalise and generalise the instruments of correlation.

3 Formalisation and generalisation of Galton's ideas: the modern tools of correlation

Pearson (1920: 41) maintains that one of the most important consequences of the diffusion of *Natural Inheritance* was the attraction that the concept of correlation exerted on three individuals: Edgeworth, Weldon and himself. Beginning in the early 1890s, Edgeworth would be responsible for some relevant developments of the concept in the field of theoretical statistics; Weldon, for his part, was one of the first to use it extensively in empirical research, and Pearson, taking advantage of his mathematical skills, gave the formulae their current form.

In 1890, Weldon wrote his first article on correlation. In it he presents the results of research on the relationship between shrimp organs, calling the coefficients of correlation 'Galton's functions', but without introducing any further innovation. In his second statistical paper he substitutes the median for the arithmetic mean in calculating correlation, thus increasing the accuracy of the results, and he reaffirms the use of r to symbolise the corresponding coefficient (Weldon 1892). In addition, he deals for the first time with negative correlations. Weldon's third contribution, published in 1893, presents the coefficients of correlation between shore crab organs from Plymouth and Naples. In this memoir the need to calculate the probable error or r in order to assess its significance became evident; this issue would later be treated by Pearson, who by that time had just begun to collaborate with Weldon.

Edgeworth wrote on correlation for the first time in 1892, following a suggestion by Galton, who had been unable to mathematically solve the problem of multiple correlation. In *Correlated Averages* (Edgeworth 1892), he develops the so-called 'Edgeworth's expansion', a generalisation of the joint bivariate distribution to the

case of three variables. This article was received favourably,[23] in spite of misprints and omissions that rendered the derivation of the mathematical formulae quite difficult to follow. Edgeworth's most lasting contribution was his proposal to name *r* the 'coefficient of correlation', replacing from then on earlier verbal expressions such as Galton's 'index of correlation' and Weldon's 'Galton's function'.

In 1893, Edgeworth dealt with the problem of how to apply the statistical correlation to social phenomena, but without providing further theoretical innovations. He limited himself to showing how the "double law of error"[24] was fulfilled by Galton's observations on human beings: "there exists a mathematical, as well as an artistic, proportion between the parts of the human frame" (Edgeworth 1893: 671). As far as the debate on whether to use the median or the arithmetic mean for calculating the correlation was concerned—referred to by Edgeworth as the *battle of the means*—he reckoned it was futile in the case of symmetrical distributions, since both of them would yield the same value (which today seems obvious). Nonetheless, particularly when applying mathematical ideas to social phenomena, "regard must be had to the degree of irregularity which may be expected in the subject matter". Contrary to Weldon's and Pearson's opinion, Edgeworth concluded that based both on "theoretical and practical considerations [it should be recommendable] sometimes employing the median" (ibi: 673).

Undoubtedly, the most remarkable contribution to the theory of correlation that followed the publication of Galton's articles in the late 1880s is to be found in a memoir by Pearson: *Mathematical Contributions to the Theory of Evolution: Regression, Heredity, and Panmixia* (Pearson 1896). This article is well-known for introducing the product-moment formula and for demonstrating that it is the "best" mode to calculate the correlation coefficient. However, the relevant theoretical aspects of this essay are numerous and require in-depth consideration.

Pearson begins his argument with a phrase that evokes, by analogy, the words chosen by Galton almost ten years earlier to open his famous *Co-relations and their Measurement, Chiefly from Anthropometric Data*:

> The problems of regression […] have been dealt with by Mr. Francis Galton in his epoch-making work on 'Natural Inheritance', but, although he shows exact methods of dealing, both experimentally and mathematically, with the problem of inheritance, it does not appear that mathematicians have hitherto developed his treatment […] The present memoir will be devoted to the expansion and fuller development of Mr. Galton ideas (Pearson 1896: 254–255).

Pearson defines correlation as the situation in which the mean of a variable—for instance, the size of an organ—is found to be a "function" of the values of another variable. His research on correlation was meant "to reach the necessary fundamental formulæ with a clear statement of *what assumptions are really made*, and with special reference to what seems legitimate in the case of heredity" (ibi: 261). Pearson's basic problem was to determine the best practical way to calculate the coefficient of correlation. While considering previous solutions proposed by Galton and Weldon unsatisfactory, he found that the best-suited method was the one presented by Bravais in 1846, since it always provided the best result and offered no calculation difficulties:

[23] In his *Notes on the History of Correlation* (1920) Pearson, instead, assessed this contribution negatively. For Pearson Edgeworth had left the topic of multiple correlation in an incomplete state.

[24] Bivariate normal distribution.

> It appears that the observed result is the most probable, when r is given the value S (xy) / $(n\sigma_1\sigma_2)$. This value presents no practical difficulty in calculation, and therefore we shall adopt it. It is the value given by Bravais, but he does not show that it is the best (ibi: 265).

Pearson argues that S(xy) is a product-moment,[25] and that its vanishing indicates the absence of correlation.[26] In current terms this is the covariance (Σxy)—the sum of the products of each pair of deviances of x and y from their respective mean—which is zero if the variables are not related and whose upper limit is one. From this starting point Pearson (ibi: 275) arrives at the formula of the correlation coefficient, which he proposes to calculate as follows[27]:

$$r = \frac{S(xy)}{(n\sigma_1\sigma_2)}$$

The product-moment formula was effectively communicated in this article, but Yule maintains that Pearson had already illustrated it at least one year earlier in his statistical lectures at University College London. In fact, in Yule's *Notes on Karl Pearson's Lectures on the Theory of Statistics* for Session 1894–1895 the following can be read: "The 'best' value to give r, deduction of the product-sum formula"[28] (Yule 1938: 201). However, as Yule himself emphasises, Pearson used to tell his students that at that time such formula had not been yet applied in empirical research.

When initiating his collaboration with Weldon, Pearson had the opportunity to read his essays. Particularly those from 1892 and 1893 convinced Pearson of the extreme importance of considering the probable error of r in order to give a sound interpretation of its results. Not surprisingly, already in 1896 he had made an attempt to introduce a formula for its calculation. Nonetheless, it was slightly erroneous; the correct result would be reached two years later by himself and his close collaborator Filon (see Pearson and Filon 1898).

Another fundamental aspect dealt with by Pearson in the memoir of 1896 is multiple correlation. In substance, Galton had solved the problem of the correlation between two variables. Then, he confronted the issue of its generalisation, but without arriving at a mathematically appropriate solution. In 1892, Edgeworth treated the problem of the joint variation of three variables, and suggested that the method could be extended to models involving even more variables. Pearson welcomed this suggestion and became involved in the development of a large part of the modern theory of (linear) multiple regression and correlation, which "consists of forming a linear equation between any one variable x_1 of a group, and the other variables x_2, x_3, x_4, &c.; this equation being so formed that the sum of the squares of the errors made in estimating x_1 from its associated variables x_2, x_3, x_4, &c., is the least possible" (Yule 1897b: 817).

However, the solution remained confined to rather theoretical grounds, since "in the general case of n variables, the mathematics of the subject [became] somewhat

[25] Pearson had already used the term 'moment' in a letter to *Nature* dated 26 October, 1893.

[26] Note that the coefficient r can be zero even in the case of two variables non-linearly correlated (see Blalock 1960/1986: 395). Yule (1897b: 821) had already emphasised that if a given regression departed substantially from the linear model, then r should be used with extreme caution.

[27] Using contemporary notation, the Statistics textbooks present the following formula: $r = \Sigma xy/\sqrt{(\Sigma x^2)\ (\Sigma y^2)}$, which can be interpreted as the relation between the covariance and the square root of the product of the deviance of x and y (Blalock 1960/1986, chapter xvii).

[28] As already mentioned, Yule used the expression 'product-sum' instead of 'product-moment'.

complicated" for those times (Yule 1909: 722; see also Yule 1897b: 837–838). It was not until 1909 that Yule introduced a special notation which permitted a remarkable simplification both in the algebra and the arithmetical processes involved in multiple correlation. It should be noted in passing that the first comprehensive treatment of the application of any form of mechanic calculation to correlation, Wallace and Snedecor's *Correlation and Machine Calculations*, was published in 1925.

In addition, Pearson soon realised that all too often it made no sense to include many variables in regression and correlation analysis, and consequently, the practical problems involving calculation difficulties became less significant. All the variables under consideration could be associated or correlated among themselves in various degrees; but experience shows that introducing more variables into the model usually yields no better results:

> The theory of multiple correlation shows that freedom to vary is quite compatible with an indefinite number of determining variables, and actual experience of correlation shows it is only a few highly correlated variables that matter (Pearson 1892/1957: 172–173).

Finally, in the 1896 essay appears "one of the first discussions of spurious correlation" (Melberg 2000: 2), an issue to which Pearson would return in *Mathematical Contributions to the Theory of Evolution: On a Form of Spurious Correlation which May Arise when Indices Are Used in the Measurement of Organs* (Pearson 1897). According to Aldrich (1995), Pearson uses the expression 'spurious correlation' in order to distinguish between scientifically important and unimportant correlations, the latter not being indicative of organic relationships. This new concept caught Galton's attention, and he wrote a brief note stressing the great significance of understanding the genesis of spurious correlation, since it can have *prima facie* a paradoxical appearance. He considered that, in order to avoid fallacious conclusions, it was fundamental to sensitise statisticians regarding this problem (see Galton 1897).

Yule resumed Pearson's considerations on spurious correlations, often calling them "illusory" (Aldrich 1995) or "absurd" (Kendall and Buckland 1976) so as to highlight the fact that correlations are not spurious in themselves; what is spurious is the inference of a significant relationship drawn acritically from the face value of a coefficient of correlation (Melberg 2000). As an example of spurious correlation Yule (1926) reports the case of mortality and the proportion of weddings celebrated at the Church of England: the coefficient was 0.95; however, it was absolutely senseless to rationally connect the two phenomena and establish a causal relationship between them.[29]

In 1897 Yule's *On the Theory of Correlation*, another classic article on the subject, was published. He presents his work in an unpretentious fashion, making clear that his objective is not to introduce any innovations. His declared intention is didactic in nature: to gather together previous developments on the theory of correlation in their entirety and to illustrate them with profuse numerical examples.

Yule presents the topic in an educational and clear style. He starts by reviewing previous work on the theory of correlation; he then proceeds to define the most elementary terms and assumptions, stressing the fact that correlation has been applied

[29] Other interesting examples are reported by Ricolfi (1993) and Melberg (2000). Hendry (1980) showed that the correlation coefficient between two time series, inflation and rain in the UK, was astonishingly 0.998.

to "necessarily *numerical quantities*".[30] Following the typical statistical trend of his time, he proposes to use the term 'correlation' instead of 'causal relation'[31] when considering the connection between variables.[32]

His treatment of correlation begins with regression and the application of the method of least squares. Let us consider two correlated variables x and y—for example, the age of men and that of women at the moment of marriage.[33] In a scatter plot, the paired values of x and y will appear more or less closely distributed round a smooth curve that is customarily called the *curve of regression*. "In many cases this curve does not diverge very seriously from a straight line; in a few cases it may be said to be a straight line within the limits of probable error; we will then speak of the *line of regression*". This straight line can be fit to the curve "subjecting the distances of the [paired values of x and y] from the line to some minimal condition", by applying the method of least squares. "If the slope [of the line] be positive, we will say that large values of x are on the whole associated with large values of y. If it is negative, large values of x are associated with small values of y, and *vice versa*".[34] The slope of the line is a practical measure: assuming the regression of y on x, it indicates the changes in y given a change in x. The equation to the line consequently gives a concise answer to two most important statistical questions: (1) if the values of y are in general associated with either large or small values of x, and (2) the magnitude of the quantitative shift in y corresponding to any given numerical change in x (ibi: 814).

The use of standard deviation instead of the arbitrary units that measure x and y makes it possible to calculate the coefficient of correlation r. The best value for r is obtained by means of the product-sum formula proposed by Bravais (1846) and fully demonstrated by Pearson (1896). "When r is unity we may say that the two variables are perfectly correlated, but when it is zero we cannot say that they are strictly uncorrelated […] The condition $r = 0$ is necessary but is not sufficient" (Yule 1897b: 821). Considering this latter statement, it can be affirmed that even though Yule had declared not to be in a position to introduce new elements into the theory of correlation, he was actually highlighting a crucial issue that had been neglected thus far: the assumptions of the normality of the distributions and of the linearity of the relationships underlying former developments on the theory of regression and correlation.[35]

[30] This fact shows that Yule was, by then, already aware of how the nature of the properties under analysis limits the use of certain statistical techniques (in this case, correlation). Not surprisingly he was the first one to conceive alternative tools suitable for investigating the relations between non-quantitative (or non-measurable) properties. (see Yule 1900).

[31] Actually, British statistical thought of the late nineteenth century—following a positivistic perspective—energetically rejected the notion of causality. For a detailed review of this and other philosophical ideas informing statistical thinking, in particular Pearson's, see Piovani (2004).

[32] This is an important issue to mention: Yule may have been the first to speak of correlation between *variables* instead of correlation between *organs* or *characters*. In effect, he wrote: "The quantities […] whose relations it is desired to investigate will be spoken of as the *variables*, since their magnitude varies" (Yule 1897b: 812).

[33] Yule exemplifies with a diagram of Italian marriages prepared by Perozzo, apparently in exhibition at the *Royal Society*.

[34] In the equations of regression ($y = a + bx$ and $x = a + by$) the slope is represented by b, termed the 'coefficient of regression' (see Yule 1897b: 819).

[35] In fact, Pearson (1896: 262) had warned: "We shall […] assume that the sizes of this complex of organs are determined by a great variety of *independent* contributory causes […], that the variation in intensity of the contributory causes are small as compared with their absolute intensity, and that these variations follow the normal law of distribution."

In Yule's opinion, the question was not only how to obtain the relationships but also how to interpret them. In the case of two normally distributed variables, the problem had already been solved. But, as he sets forth (ibi: 842), the bivariate normal surface implies that:

(1) the total distributions of the two variables (and the distribution of every array) are normal;
(2) the regressions are truly linear;
(3) the standard deviations of all parallel arrays are equal, and
(4) the contour lines are a system of similar and similarly situated ellipses, the centres coinciding with the mean of the whole surface.

These assumptions of "normal correlation"[36] are usually satisfied when dealing with scientific problems in the fields of biology, zoology and anthropometry, those that interested Galton, Weldon and Pearson. However, Yule was more concerned with specific problems of economics and the social sciences, disciplines in which "normal correlations" are rarely found (ibi: 851). This led him to consider the issue of correlation between variables asymmetrically distributed (*skew correlation*), a topic also treated by Pearson some time later. The *desideratum* was "the discovery of an appropriate system of surfaces, which will give bi-variate skew frequency [and will] free ourseleves from the limitations of the normal surface" (Pearson 1920: 44). Yule's deductions concerning skew correlation (Yule 1897a, b) were later developed by Edgeworth (1902, 1908) and the latter's disciple Bowley (1903).

In 1897, when Yule wrote the memoir scrutinised above, he was a young assistant to Professor Pearson. Nonetheless, he was already making important contributions to the theory of correlation. Actually, the above-mentioned treatment of the correlation between asymmetrical variables was not his first significant input; some time before he had introduced the concept of 'net correlation', later renamed 'partial correlation' on Pearson's suggestion. The rationale underpinning partial correlation was to calculate the coefficient of correlation between any two given variables, eliminating the effects of variations on a third one (see Yule 1896, 1897b). This idea was later considered by Pearson and Lee (1897).

Another important contribution to the theory of correlation was made by Charles Spearman, a pioneer in the application of statistical tools to psychological research. In 1904, in an attempt to "objectively measure" human intelligence, he introduced the rank order correlation and, as its operative tool, the coefficient ρ. According to Spearman (1904b: 222) the search for a suitable method for studying psychological attributes "compelled [him] to enter into a general discussion of the methods universally valid for demonstrating association between two events or attributes". This had actually been the main subject of another classic essay published by him the same year and in the same Journal, *The Proof and Measurement of Association Between Two Things* (Spearman 1904a), in which he presented the coefficient ρ. Above all, Spearman was looking for a measure that could attain "to the first fundamental requisite of correlation, namely, a *precise quantitative measure*" (Spearman 1904b: 222); "one plain numerical value (varying conveniently from 1 for perfect correspondence down to 0 for perfect absence of correspondence)" (ibi: 225):

[36] Yule uses this expression in order to highlight that Pearson's theory and technique were appropriate to very particular situations, and therefore they were not automatically applicable to cases not satisfying the assumptions of normality.

> Our problem is of a very definite objective nature; we wish to ascertain how far the observed ranks in the several [psychological] abilities tend to correspond with one another (ibi: 252).

The coefficient proposed by Spearman was calculated by means of the already well-known product-moment formula; only instead of using measurements (in the strict sense of the term), he employed those numbers denoting the ordinal position of the subjects (their relative rank). In effect the coefficient ρ "measures the intensity of the correlation between two systems of *rankings* or their degree of correspondence" (Kendall and Buckland 1976/1980: 59).

Pearson severely criticised Spearman's approach and rejected both the concept and the coefficient of rank order correlation. The basis of his criticism lay in Pearson's particular conception of all properties as continuous and truly measurable, and his consequently firm opposition to any attempt to develop (and apply) correlation techniques that did not attain to this assumption. Edgeworth and Yule maintained a different position as regards measurement, and therefore favoured Spearman's approach. Actually, in *Correlation Calculated from Faulty Data* (Spearman 1910) Spearman wrote in a footnote that he was pleased to have learnt that Yule "disagrees with Pearson's adverse comments" and finds his proposal "a very important step in the simplification of methods dealing with non-measurable characters". These different ideas regarding the nature of properties were a crucial aspect of Pearson and Yule's bitter controversy over how to define (and to analyse) the most suitable coefficients for treating the association between qualitative variables.

In 1910, the development of the theory of correlation—and in Yule's opinion, also the main direction it would take in the years to come—was oriented towards non-linear relationships (Yule 1909: 723). Pearson had been the first to tackle this issue, but further efforts were still needed (see Pearson 1902, 1905).

Except for this, the theory of correlation and its respective technical instruments were well-established; essentially all the fundamental issues had been dealt with and most already resolved. Therefore, it was now feasible to organise in an articulate and comprehensive manner the diverse contributions made over almost three decades of continuous work (and scattered about in numerous essays) by Galton, Weldon, Edgeworth, Pearson, Yule, Spearman and others. This would have been absolutely impossible some fifteen years earlier, when the concept and the tools of correlation were in their early phases of development (Yule 1938).

In 1908, Hooker wrote a concise elementary article on the theory of correlation, and in 1911 Yule published the first version of his classic handbook of statistics. In the course of its fourteen editions, this book would become an obligatory reference for teachers and students of statistics until the 1950s.

These works introduced no theoretical or technical innovations. Their importance resides in their systematisation of the theory of correlation. Moreover, they bear witness to two interesting facts. On the one hand, there is the enormous spread that these methods had already achieved; in fact, they were being applied in diverse fields even by researchers lacking statistical expertise; at that time, the use of these tools was not limited to those who had contributed—or were in a position to contribute—to their development. On the other hand, they are proof of how specialised and mathematically refined the discipline of statistics had become. Consequently, making educational materials for it more generally available was increasingly necessary.

References

Aldrich, J.: Correlations genuine and spurious in Pearson and Yule. Stat. Sci. **10**(4), 364–376 (1995)

Blalock, H.M.: Social Statistics. Mc Graw-Hill, New York [quotations from the spanish translation, 2nd edition.: Estadística Social. Mexico: FCE, 1986] (1960)

Bowley, A.L.: The Measurement of Groups and Series. Layton, London (1903)

Bravais, A.: Analyse mathématique sur les probabilités des erreurs de situation d'un point. Mémoires présentés par divers savants à l'Académie Royale des Sciences de l'Istitut de France **9**, 255–332 (1846)

Carpenter, W.: The correlation of the physical and vital forces. Br. Foreign Med.-Chir. Rev. **8**, 206–238 (1851)

Darwin, C.: On the Origin of Species. John Murray, London (1859)

Darwin, C.: The Variation of Animals and Plants under Domestication. John Murray, London (1868)

Denis, D.J.: The origins of correlation and regression. In: Proceedings of the 61st Annual Convention of the Canadian Psychological Association, Ottawa (2000)

Edgeworth, F.Y.: On correlated averages. Philos. Mag. Ser. **5**(34), 190–204 (1892)

Edgeworth, F.Y.: Statistical correlation between social phenomena. J. R. Stat. Soc. **56**(4), 670–675 (1893)

Edgeworth, F.Y.: The Law of Error. Encyclopædia Britannica, 10th ed., vol. 28 (supplement to 9th ed., vol. 4), pp. 280–291 (1902)

Edgeworth, F.Y.: On the probable errors of freqeuncy-constants. J. R. Stat. Soc. **71**(3), 499–512 (1908)

Galton, F.: Hereditary Genius. An Inquiry into its Laws and Consequences. MacMillan, London (1869)

Galton, F.: Statistics by Intercomparison. Philos. Mag. Ser. **4**(49), 33–46 (1875)

Galton, F.: Typical laws of heredity. Proc. R. Instit. **8**, 282–301 (1877)

Galton, F.: Address to the anthropological section of the British association. Nature **32**, 507–510 (1885)

Galton, F.: Family likeness in stature. Proc. R. Soc. Lon. **40**, 42–73 (1886)

Galton, F.: Co-relations and their measurement, chiefly from anthropometric data. Proc. R. Soc. Lon. **45**, 135–145 (1888)

Galton, F.: Natural Inheritance. Macmillan, London (1889)

Galton, F.: Kinship and correlation. North Am. Rev. **150**, 419–431 (1890)

Galton, F.: Notes to the Memoir by Professor Karl Pearson, F.R.S., on Spurios Correlation. Proc. R. Soc. Lon. **60**, 498–502 (1897)

Galton, F.: Memories of My Life. Methuen, London (1908)

Gauss, C.F.: Theoria combinationis observationum minimis obnoxiae. Dieterich, Göttingen (1823)

Grove, W.R.: The Correlation of Physical Forces. The Substance of a Course of Lectures, London (1846)

Hendry, D.F.: Econometrics. alchemy or science. Economica **47**, 387–408 (1980)

Hooker, R.H.: An elementary explanatin of correlation, illustrated by rainfall and a depth of water in a well. J. R. Meteorol. Soc. **34**, 277 (1908)

Jevons, W.S.: The Principles of Science. A Treatise on Logic and Scientific Method. Macmillan, London (1874)

Kendall, M.G., Buckland, W.R.: A Dictionary of Statistical Terms. Longman, London [quotations from the spanish translation: Diccionario de Estadística, Madrid: Pirámide, 1980] (1976)

Kuhn, Th.: The Structure of Scientific Revolutions. University Press, Chicago (1962)

Lancaster, H.O.: Development of the notion of statistical dependence. Math. Chron. **2**, 1–16 (1972)

Legendre, A.M.: Nouvelles Methodes pour la Détermination des Orbites des Comètes. Courcier, Paris (1805)

MacKenzie, D.A.: Statistics in Britain, 1865–1930. The Social Construction of Scientific Knowledge. University Press, Edinburgh (1981)

Melberg, H.O.: (2000) From Spurious Correlation to Misleading Association: The Nature and Extent of Spurious Correlation and its Implications for the Philosophy of Science with Special Emphasis on Positivism. University of Oslo < http://www.geocities.com/hmelberg/papers/pd_causal2.pdf >

Micheli, G.A., Manfredi, P.: Correlazione e Regressione. Angeli, Milan (1995)

Nicholls, N.: William stanley jevons and the climate of Australia. Aust. Meteorol. Mag. **47**, 285–293 (1998)

Pearson, K.: The Grammar of Science. Scott, London [quotations from the 3rd revision, 1911. New York: Meridian, 1957] (1892)

Pearson, K.: Mathematical contributions to the theory of evolution: regression, heredity, and Panmixia. Philos. Trans. R. Soc. Lon. **187**, 253–318 (1896)

Pearson, K.: Mathematical contributions to the theory of evolution: on a form of spurious correlation which May Arise when Indices Are Used in the measurement of organs. Proc. R. Soc. **60**, 489–498 (1897)
Pearson, K.: On the systematic fitting of curves to observations and measurements. Biometrika **1**(3), 265–303 (1902)
Pearson, K.: On the General Theory of Skew Correlation and Non-Linear Regression. Draper's Company Research Memoirs, biometric series II. Dulau, London (1905)
Pearson, K.: Life and letters of Francis Galton. University Press, Cambridge (1914–1930)
Pearson, K.: Notes on the history of correlation. Biometrika **13**(1), 25–45 (1920)
Pearson, K., Filon, L.N.G.: Mathematical contributions to the theory of evolution, IV: on the probable errors of frequency constants and on the influence of random selection on variation and correlation. Philos. Trans. R. Soc. A **191**, 229–311 (1898)
Pearson, K., Lee, A.: On the distribution of frequency (Variation and Correlation) of Barometric height at divers stations. Philos. Trans. R. Soc. A **190**, 423–469 (1897)
Piovani, J.I.: L'epistemologia di Karl Pearson. Sociol. ricerca sociale **75**, 5–28 (2004)
Porter, Th.: The Rise of Statistical Thinking, 1820–1900. University Press, Princeton (1986)
Ricolfi, L.: Tre variabili. Un'introduzione all'analisi multivariata. Angeli, Milan (1993)
Schols, C.M.: Théorie des erreurs dans le plan et dans l'espace. Ann. l'Ecole polytechnique de Delft **2**, 123–178 (1886)
Seal, H.L.: Studies in the history of probability and statistics. XV: the historical development of the gauss linear model. Biometrika **54**(1/2), 1–24 (1967)
Spearman, C.E.: The proof and measurement of association between two things. Am. J. Psychol. **15**, 72–101 (1904a)
Spearman, C.E.: General intelligence, objectively determined and measured. Am. J. Psychol. **15**, 201–293 (1904b)
Spearman, C.E.: Correlation Calculated from Faulty Data. British J. Psychol. **3**, 271–295 (1910)
Stigler, S.M.: Francis Ysidro Edgeworth, statistician. J. R. Stat. Soc. A **141**, 287–322 (1978)
Stigler, S.M.: Gauss and the invention of least squares. Ann. Stat. **9**, 465–474 (1981)
Stigler, S.M.: The History of Statistics: The Measurement of Uncertainty Before 1900. Harvard University Press, Cambridge, MA (1986)
Stigler, S.M.: Statistics on the Table: The History of Statistical Concepts and Methods. Harvard University Press, Cambridge (1999)
Tankard, J.W.: The Statistical Pioneers. Schenkman, Cambridge (1984)
Walker, H.M.: Studies in the History of Statistical Method. Williams and Wilkins, Baltimore (1929)
Wallace, H.A., Snedecor, G.W.: Correlation and Machine Calculations. Iowa State College Press, Anies (1925)
Weldon, W.F.R.: The variations occurring in certain decapod Crustacea. Proc. R. Soc. **47**, 445–453 (1890)
Weldon, W.F.R.: Certain correlated variations in crangon Vulgaris. Proc. R. Soc. **51**, 2–21 (1892)
Weldon, W.F.R.: On certain correlated variations in carcinus Moenas. Proc. R. Soc. **54**, 318–329 (1893)
Westergaard, H.: Contributions to the history of statistics. P. S. King, London (1932)
Winter, A.: The construction of orthodoxies and heterodoxies in the early victorian life sciences. In: Lightman, B. (ed.) Victorian Science in Context. University Press, Chicago (1997)
Yule, G.U.: On the significance of Bravais' formulae for regression, in the case of Skew correlation. Proc. R. Soc. Lon. **60**, 477–489 (1897a)
Yule, G.U.: On the theory of correlation. J. R. Stat. Soc. **60**(4), 812–854 (1897b)
Yule, G.U.: On the association of attributes in statistics. Philos. Trans. R. Soc. Lon. A **194**, 257–319 (1900)
Yule, G.U.: The application of the method of correlation to social and economic statistics. J. R. Stat. Soc. **72**(4), 721–730 (1909)
Yule, G.U.: An Introduction to the Theory of Statistics. Griffin, London (1911)
Yule, G.U.: Why do we sometimes get nonsense correlations between time-series? A study in sampling and the nature of time-series. J. R. Stat. Soc. **89**(1), 1–63 (1926)
Yule, G.U.: On the correlation of total pauperism with proportion of out-relief. Econ. J. **6**(24), 613–623 (1986)
Yule, G.U.: Notes of Karl Pearson's lectures on the theory of statistics 1894–96. Biometrika **30**(1/2), 198–203 (1938)