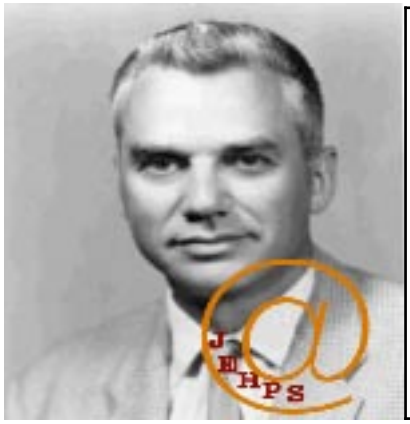

Journ@l Electronique d'Histoire des Probabilités et de la Statistique

Electronic Journ@l for History of Probability and Statistics

Vol 4, n°2; Décembre/December 2008

www.jehps.net

About "Correlations"

BRUNO DE FINETTI[1]

Résumé

Nous introduisons une représentation géométrique possible de le coefficient de corrélation, *r*, qui s'avère utile pour interpréter de façon intuitive les relations entre les espérances mathématiques, écart-types et coefficients de corrélation. Nous commentons des cas particuliers, avant de montrer que, dans le cas général, dès lors que les lois de probabilité de deux variables aléatoires *X* et *Y* ont été définies, on ne peut plus attribuer arbitrairement au coefficient *r*(*X*, *Y*) n'importe quelle valeur entre -1 et +1. Nous prouvons que les deux bornes ne peuvent être atteintes que si les fonctions de distribution Φ_1 et Φ_2 sont toutes deux similaires, "anti-similaires", ou (afin que les deux bornes soient atteintes) similaires et symétriques. Enfin, nous clarifions le fait qu'un indice de concordance peut ne pas avoir le même signe que *r*, et nous proposons ce qui nous semble être l'indice de concordance le plus simple et le plus significatif (et qui, à notre connaissance, n'a encore jamais été suggéré).

Abstract

We sketch one possible geometric interpretation of the correlation coefficient, *r*, which turns out to be useful to make the relationships among mathematical expectations, standard deviations and correlation coefficients intuitive. Then, after adding some remarks on special cases, we show that, in general, when *the probability laws* of two random variables *X* and *Y have been assigned*, it is no longer possible for *r*(*X*, *Y*) to assume an arbitrary value between –1 and +1. We show that the two bounds can be reached only if the two distribution functions Φ_1 and Φ_2 are, respectively, similar, "anti-similar", or (in order to attain both bounds) similar and symmetrical. Finally, we clarify that a concordance index can have a sign different from that of *r*, and suggest what seems to us the simplest and most intrinsically meaningful index of concordance (which has not yet been considered, as far as we know).

[1] Bruno de Finetti, "A proposito di correlazione", *Supplemento Statistico Nuovi Problemi*, A. III, nn. 1-2, 521-38, 1937. English translation by Luca Barone (Goldman Sachs International, Peterborough Court, 133 Fleet Street, London EC4A 2BB, luca.barone@gs.com) and Peter Laurence (Università degli Studi di Roma "La Sapienza", Dipartimento di Matematica "G. Castelnuovo", Piazzale Aldo Moro, 2 - 00185 Roma, laurence@mat.uniroma1.it).

The abstract, which is not present in the original article, has been added. The same applies to section headers. The original notes of the Author are at the bottom of the page, while some notes of the editors are at the end of the document. All the equations have been numbered to make references easier.

Introduction

In the case of "correlations", as indeed is often the case, many discussions arise from confusion between different concepts. For a long time we have seen papers demonstrating and repeating that is necessary to distinguish between the concept of "correlation" that is measured by the "correlation coefficient" r (of Bravais), and the concept of "stochastic dependence" defined in probability theory. If the need to clarify this point is so widespread, it certainly means that the confusion was also pretty widespread, probably because the two concepts are equivalent in the case of Gaussian distributions and we have the unjustified and harmful habit of considering the Gaussian distribution in too exclusive a way, as if it represented the rule in almost all the cases arising in probability and in statistics, and as if each non-Gaussian distribution constituted an exceptional or irregular case (even the name of "normal distribution" can contribute to such an impression, and it would therefore perhaps be preferable to abandon it).

I would not see any need to add a word clarifying the distinction between the two concepts — which should never had been confused and which (as far as I know) have always been kept well distinguished by researchers in the calculus of probabilities — if it didn't seem to me that, discovering that the correlation coefficient doesn't have the meaning that, for an incomprehensible misunderstanding, some attributed to it, we are left to think that this has deprived the correlation coefficient of *every* meaning. It would be like deciding that thrashing machines are useless because we proved to someone who confused them with mills that they don't serve to grind the wheat!

Therefore, I think it is not useless to give a brief exposition of the meaning of the correlation coefficient seen from the vantage point of the calculus of probabilities, in particular because some considerations may perhaps appear new, at least in the form and in the light in which they are cast from the standpoint of the theory of random variables. In addition, I will address in this article the issue of finding the most proper terminology to avoid perpetuating the aforementioned misunderstanding and to obviate other drawbacks that we will run into along the way: however, the matter is so entangled that, whenever I am unable to introduce satisfactory solutions, I will do no more than clarify what is the requirement we should fulfill to make the appropriate decisions.

1 Mathematical expectation of second-order functions

Given a random variable X, we indicate by $M(X)$ the "mathematical expectation" and by $\sigma(X) = \sqrt{M(X - M(X))^2}$ the "standard deviation", as usual.

It is known that the mathematical expectation enjoys the additivity property $M(X + Y) = M(X) + M(Y)$, and this is enough to resolve all the problems in which we have only linear combinations of random variables: in this case, knowing the mathematical expectation is sufficient to solve the problem. Instead, if $Z = f(X, Y)$ is a non-linear function of X and Y, or if, when such a function is linear, we need to have a deeper knowledge of the probability distribution of Z than can be gained by merely knowing its mathematical expectation, other elements are obviously necessary. The first such element is the correlation coefficient, which is sufficient, together with M and σ, to solve all the "second-

order" problems, in which we need to determine the mathematical expectation of any second-order function of given random variables X_1 X_2, ..., X_n. Among the second-order problems there are two of immediate practical interest, which are sufficient to prove the enormous importance of the correlation coefficient, and to explain why its presence is necessary: these are the problems arising when we need to determine the mathematical expectation of the *product* $X\,Y$ of two random variables X and Y and the standard deviation of their sum $X+Y$ (an obvious generalization of the two problems is the calculation of the mathematical expectation of the product of two linear combinations of random variables, and in particular of the square of such a linear combination).

To tackle the problem from a more general standpoint, let us consider a generic second-order function of n random variables X_1, X_2,..., X_n, and let

$$Z = \sum_{ij} a_{ij} X_i X_j \tag{1}$$

(we can suppose, without loss of generality that $a_{ij} = a_{ji}$).

Indicating by $\overline{X}_i = M(X_i)$ the mathematical expectation of X_i, we can write

$$\begin{aligned} Z &= \sum_{ij} a_{ij}[X_i + (X_i - \overline{X}_i)][X_j + (X_j - \overline{X}_j)] \\ &= \sum_{ij} a_{ij}\overline{X}_i\overline{X}_j + 2\sum_{ij} a_{ij}\overline{X}_i(X_j - \overline{X}_j) + \sum_{ij} a_{ij}(X_i - \overline{X}_i)(X_j - \overline{X}_j)] \end{aligned} \tag{2}$$

and the mathematical expectation of Z is

$$\begin{aligned} M(Z) &= \sum_{ij} a_{ij}\overline{X}_i\overline{X}_j + 2\sum_{ij} a_{ij}\overline{X}_i M(X_j - \overline{X}_j) + \sum_{ij} a_{ij} M(X_i - \overline{X}_i)(X_j - \overline{X}_j) \\ &= \sum_{ij} a_{ij}\overline{X}_i\overline{X}_j + \sum_{ij} a_{ij} M(X_i - \overline{X}_i)(X_j - \overline{X}_j) \end{aligned} \tag{3}$$

because $M(X_j - \overline{X}_j) = M(X_j) - \overline{X}_j = \overline{X}_j - \overline{X}_j = 0$.

Therefore we can always determine $M(Z)$ if we know, in addition to $\overline{X}_i = M(X_i)$, the value of the terms $M[(X_i - \overline{X}_i)(X_j - \overline{X}_j)]$.

When $i = j$ we have by definition $M(X_i - \overline{X}_i)^2 = \sigma^2(X_i)$; instead, when $i \neq j$, knowing $M(X_i)$ and $\sigma(X_i)$ is not sufficient to determine $M[(X_i - \overline{X}_i)(X_j - \overline{X}_j)$ but it is enough to find a bound, from which the introduction of the "correlation coefficient" naturally follows.

2 Basic properties of the correlation coefficient

In fact, given two random variables X and Y, let us consider their linear combination $Z_t = X + tY$, where t is a real number. We have

$$\begin{aligned} \sigma^2(Z_t) &= M[(X - \overline{X}) + t(Y - \overline{Y})]^2 = M(X - \overline{X})^2 \\ &\quad + 2t\,M(X - \overline{X})(Y - \overline{Y}) + t^2 M(Y - \overline{Y})^2 \\ &= \sigma^2(X) + 2t\,M(X - \overline{X})(Y - \overline{Y}) + t^2\sigma^2(Y) \end{aligned} \tag{4}$$

If we consider t as a parameter, then $\sigma^2(Z_t)$ is a second-order function of t. However, $\sigma^2(Z_t)$ is, by its own nature, positive (or zero). Therefore, we must have [i]

$$|\,M[(X - \overline{X})(Y - \overline{Y})]| \leq \sigma(X)\sigma(Y) \tag{5}$$

which is the bound we referred to earlier. We will immediately see that this bound is the most precise we can determine, in the sense that $M[(X-\overline{X})\,(Y-\overline{X})]$ can actually assume all the values between $\pm\sigma(X)\,\sigma(Y)$ (extremes included). So it is natural to introduce the "correlation coefficient", defined by

$$r(X,Y)=\frac{M[(X-\overline{X})(Y-\overline{Y})]}{\sigma(X)\,\sigma(Y)} \tag{6}$$

that always lies between ± 1 and doesn't depend on possible positive coefficients of proportionality which affect X and Y (in other words, if $a>0$ and $b>0$, $r(aX, bY) = r(X, Y)$. More generally, $r(aX, bY) = \pm\, r(X, Y)$ with sign +or – depending on whether the signs of a and b are equal or opposite).

The general formula for the mathematical expectation of the product of two random variables X and Y is thus[ii]

$$M(X\cdot Y)=M(X)\cdot M(Y)+r(X,Y)\cdot\sigma(X)\cdot\sigma(Y) \tag{7}$$

(from which we immediately learn that $M(X\,Y)$ lies between $M(X)\,M(Y) \pm \sigma(X)\,\sigma(Y)$), while for the standard deviation of the sum we have

$$\sigma^2(X+Y)=\sigma^2(X)+\sigma^2(Y)+2\cdot r(X,Y)\cdot\sigma(X)\cdot\sigma(Y) \tag{8}$$

(from which we see that $\sigma(X+Y)$ always lies between $|\sigma(X)-\sigma(Y)|$ and $\sigma(X)+\sigma(Y)$).

Depending on the sign of $r(X, Y)$, we say that X and Y are positively or negatively correlated: they are *positively* correlated when $M(X\,Y) > M(X)\,M(Y)$ or rather $\sigma^2(X+Y) > \sigma^2(X)+\sigma^2(Y)$; *negatively* correlated when $M(X\,Y) < M(X)\,M(Y)$ or rather $\sigma^2(X+Y) < \sigma^2(X)+\sigma^2(Y)$.

If $r=0$, or rather $M(X\,Y)=M(X)\,M(Y)$, or rather $\sigma^2(X+Y)=\sigma^2(X)+\sigma^2(Y)$, the two random variables are said to be *uncorrelated*. In particular, two stochastically independent random variables are uncorrelated, because it is known and can easily be seen that, in the case of independence, $M(X\,Y)=M(X)\,M(Y)$. However, the converse statement does not hold.

To show that r can indeed assume *all* the values between –1 and +1 it is sufficient to consider two independent random variables X and X' and, without losing generality, we can suppose that $M(X)=M(X')=0$, $\sigma(X)=\sigma(X')=1$, and then define, for a given r lying between –1 and + l, the random variable Y as follows:

$$Y=rX+\sqrt{1-r^2}\,X'. \tag{9}$$

We have

$$\sigma^2(Y)=r^2+(1-r^2)=1, \tag{10}$$

$$\begin{aligned} r(X,Y)=M(XY)=M(rX^2+\sqrt{1-r^2}\,XX')= \\ rM(X^2)+\sqrt{1-r^2}\,M(XX')=r, \quad \text{QED} \end{aligned} \tag{11}$$

3 Extreme cases

On the other hand, the extreme cases $r(X, Y) = \pm 1$ imply that X and Y are linearly dependent.[iii] Indeed, in this case the equation $\sigma^2(Z_t) = 0$ has in fact the root $t = \pm \sigma(X)/\sigma(Y) = k$ (sign – or + depending on whether $r = +1$ or $r = -1$).[iv] But for the random variable to have a zero standard deviation, the probability of a deviation, greater than an arbitrarily small ε, from the mathematical expectation, must be equal to zero, that is, if we accept the extended principle of total probabilities (i.e. extended to countable classes), the random variable is equal to its mathematical expectation with probability 1. It follows that, in our case, if the probability is not zero then $Z_t = \bar{Z}_t$, that is $X - \bar{X} = k(Y - \bar{Y})$.

However, I would also like to mention here the modification we need to make if we are to abandon the extended principle of total probabilities (as is necessary, in my opinion). In such case we can have $\sigma(X) = 0$ and Prob.$\{|X - \bar{X}| > \varepsilon\} = 0$ for any $\varepsilon > 0$, without necessarily having Prob.$\{|X - \bar{X}| > 0\} = 0$; for instance, if $X = 1/n$, with n being an "arbitrarily chosen whole number" (in the sense that such a whole number has zero probability) we have $\bar{X} = \sigma(X) = 0$, but Prob.$\{X > 0\} = 1$.

To avoid having to repeatedly reformulate a precise statement of this kind, and also to avoid innacuracies, as would have been, in the preceding case, the statement that necessarily $X - \bar{X} = k(Y - \bar{X})$, I think it is appropriate to introduce the symbol $X \doteq Y$, meaning "X coincides with Y", to point out that the inequality $|X - Y| > \varepsilon$ has zero probability for any $\varepsilon > 0$, or rather that $M(|X - Y|) = 0$. We can now express, correctly and in a concise way, our conclusion by saying that, if $r(X, Y) = \pm 1$, then $X - \bar{X} \doteq k(Y - \bar{Y})$,[v] where k is positive or negative together with r. And the above statement will be exact both for those who accept and for those who do not accept the principle of total probabilities, except for the different interpretation of the definition of "coincidence".

4 Geometric interpretation

I will not dwell on the meaning of "correlation coefficient" as a statistical index of "concordance", an aspect which has been often illustrated, including recently in these pages by Pietra ([2]); after all, the formula for $M(X Y)$ contains the whole mathematical aspect of such a meaning.

Rather it may be useful to sketch one possible geometric interpretation, which turns out to be useful to make the relationships among mathematical expectations, standard deviations and correlation coefficients intuitive. Since we can consider linear combinations of random variables, we can interpret them as vectors in an "abstract space". Considering $\sigma(X)$ as the *modulus* of the vector X, and, correspondingly, $\sigma(Y - X)$ as the *distance* $d(X, Y)$ between the vectors X and Y, we define a distance space, or space "D" in the sense of Fréchet, under the hypothesis that all the random variables whose difference is given by the same variable are represented by the same vector. In fact X and $X + a$ have zero "distance", because $(X + a) - X = a$, $\sigma(a) = 0$, and inversely $\sigma(Y - X) = 0$ means $Y - X = a$. In every other case $\sigma(Y - X) > 0$, and (as we basically saw in

[2] See this *Supplemento*, A. II, no. 2-3, 1936 (*L'ostracismo al coefficiente di correlazione?*).

Section 3) the triangle inequality holds:[vi]

$$|\sigma(Y)-\sigma(X)|\leq\sigma(Y-X)\leq\sigma(Y)+\sigma(X)\,. \tag{12}$$

In addition, the space S defined in this way is a *metric abstract space*, or space D_M,[3] since $\sigma(X)\ \sigma(Y)\ r(X, Y)$ may be interpreted as an inner product, being a symmetric linear homogeneous function of X and Y, which reduces to the square of the modulus when $Y = X$. Thus the correlation coefficient is the cosine of the "angle" between vectors X and Y, angle that can be unambiguously defined by setting $r(X, Y) = \cos \alpha(X, Y)$, $0 \leq \alpha(X, Y) \leq \pi$ (that is, α goes from 0° to 180°).

So formula (8) for $\sigma^2(X + Y)$ becomes

$$\sigma^2(X+Y)=\sigma^2(X)+\sigma^2(Y)+2\cdot\sigma(X)\cdot\sigma(Y)\cdot\cos\,\alpha(X,Y) \tag{13}$$

and we can see, recalling the theorem of Carnot, that $\sigma(X + Y)$ is the third side of a triangle when the other two sides, $\sigma(X)$ and $\sigma(Y)$, contain the angle $\alpha(X, Y)$. In the case of zero correlation (or, in particular, of independence) the theorem of Pythagoras holds: $\sigma(X + Y)$ is the hypotenuse of the right-angled triangle whose sides measure $\sigma(X)$ and $\sigma(Y)$.

Likewise, the formula for $M(XY)$ says that the corrective term to be added to $M(X).M(Y)$ is the inner product of the representative vectors of X and Y, that is $\sigma(X)\sigma(Y)$ cos $\alpha(X, Y)$.

Zero correlation means *orthogonality*; instead, positive or negative correlation means that α is respectively *acute* or *obtuse*; the extreme cases $r = \pm 1$ say that $\alpha = 0$ and $\alpha = \pi$, respectively, or rather that the two vectors (and therefore the random variables, up to a positive constant) only differ by a multiplicative constant, which may be, respectively, either positive or negative.

Many well known properties of metric spaces can suggest images of help in studying and solving various problems. Every random variable Y can be decomposed into two components, one correlated with X (the parallel component) and one uncorrelated with X (the orthogonal component). More generally, given n random variables $X_1, X_2, \ldots, X_n$ that are linearly independent (that is, such that there are no coefficients $a_0, a_1, a_2, \ldots, a_n$ for which $a_0 = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$), we can express them as linear combinations of n uncorrelated random variables $Y_1\ Y_2\ \ldots\ Y_n$ with unit average standard deviation (orthogonal unit system) with $M(Y_i) = 0$, $\sigma(Y_i) = 1$, $r(Y_i, Y_j) = 0$ ($i \neq j$), while for σ the usual expression of the vector's modulus holds.

If $\alpha(X, Y)$ is the angle between two random variables X and Y, a third random variable Z cannot form two arbitrary angles $\alpha(X, Z)$ and $\alpha(Y, Z)$, but we must have (obviously, if we think of the geometric picture) $\alpha(X, Y) \leq \alpha(X, Z) + \alpha(Y, Z) \leq 2\ \pi - \alpha(X, Y)$[vii] we have the extreme case $\alpha(X, Z) + \alpha(Y, Z) = \alpha(X, Y)$ if and only if $Z = aX + bY$, $a > 0$, $b > 0$ (the coplanar vector, included in the concave angle between the two vectors), and the other $\alpha(X, Z) + \alpha(Y, Z) = 2\ \pi - \alpha(X, Y)$ if and only if $Z = -(aX + bY)$ $a > 0$, $b > 0$ (the aforementioned condition applied to minus the same vector). This shows that there are some constraints for the degrees of pairwise correlation among

[3] See my note *Spazi astratti metrici* (D_M), «Atti Accad. Pontificia», A. LXXXIII, sess. VI, 1930.

different random variables. In particular, if three random variables are all equally correlated, since - two by two - they are unable to form an angle bigger than $2\pi/3$ (that is equal to 120°), the correlation coefficient cannot be smaller than $-½$.

Analogous constraints subsist for four or more random variables, and it may be interesting to extend the preceding search to the case of several random variables with equal pairwise correlations. So, let's consider n equally-correlated random variables $X_1, X_2 \ldots X_n$; for convenience we suppose that they have a zero mathematical expectation and the same standard deviation σ (this alters the moduli but not the angles of the vectors!).We will show that, under this hypothesis, the necessary and sufficient condition for the correlation coefficient to reach the least possible common value is that $\Sigma X_i = 0$. In fact, setting $Y = (1/n)\ \Sigma X_i$, $Y_i = X_i - Y$: if $Y \neq 0$, we get a new set of n equally-correlated random variables $Y_1, Y_2 \ldots Y_n$ with a smaller correlation coefficient.[viii] We will now prove this. In the first place it can be seen that Y is uncorrelated with every Y_i; let's prove this for Y_1:

$$\begin{aligned} n^2 Y Y_1 &= \left(\sum_{i=1}^{n} X_i\right)\left(nX_1 - \sum_{i=1}^{n} X_i\right) = \left(X_1 + \sum_{i=2}^{n} X_i\right)\left((n-1)X_1 - \sum_{i=2}^{n} X_i\right) \\ &= (n-1)X_1^2 + (n-2)\sum_{i=2}^{n} X_1 X_i - \sum_{i=2}^{n} X_i^2 - \sum_{ij=2}^{n}{}' X_i X_j \end{aligned} \tag{14}$$

(where Σ' stands for the sum extended to the terms with $i \neq j$);

$$n^2 M(YY_1) = (n-1)\sigma^2 + (n-2)(n-1)\sigma^2 r - (n-1)\sigma^2 - (n-1)(n-2)\sigma^2 r = 0, \tag{15}$$

and this, given that, $M(Y) = M(Y_i) = 0$, implies $r(Y, Y_1) = 0$.

Therefore $M(Y_i\ Y_j) = M(X_i\ X_j) - \sigma^2(Y)$ (in fact $X_i\ X_j = (Y_i + Y)\ (Y_j + Y) = Y_i\ Y_j + Y(Y_i + Y_j) + Y^2$, and the second term has $M = 0$), and in particular for $i = j$

$$\begin{aligned} \sigma^2(Y_i) &= M(Y_i^2) = M(X_i^2) - \sigma^2(Y) \\ &= \sigma^2(X_i) - \sigma^2(Y) = \sigma^2 - \sigma_0^2 \end{aligned} \tag{16}$$

where $\sigma_0 = \sigma(Y)$, while for $i \neq j$, $M(Y_i\ Y_j) = \sigma^2 r - \sigma_0^2$.

Thus we have

$$\begin{aligned} r(Y_i Y_j) &= \frac{M(Y_i Y_j)}{\sigma(Y_i)\sigma(Y_j)} = \frac{\sigma^2 r - \sigma_0^2}{\sigma^2 - \sigma_0^2} \\ &= r - (1-r)\frac{\sigma_0^2}{\sigma^2 - \sigma_0^2} < r \quad \text{QED} \end{aligned} \tag{17}$$

If $\Sigma\ X_i = 0$ we have $\sigma^2(\Sigma\ X_i) = 0$,

or
$$M(\Sigma X_i)^2 = M(\Sigma_i X_i^2 + \Sigma'_{ij}\ X_i X_j) = n\sigma^2 + n(n-1)\sigma^2 r = 0, \tag{18}$$

so that
$$r = -\frac{1}{n-1}: \tag{19}$$

we thus get the least level of r and the proof that the condition $\Sigma X_i = 0$ is also sufficient.

This reasoning becomes intuitive if we think about its vector interpretation: n vectors of S_n, in order to form – two by two – the same angle (as large as possible), must have the same direction of the rays that connect the centre O of an equilateral simplex to its vertexes; the angle α of two of these vectors is given by $\cos \alpha = -1/(n-1)$. Therefore, if we have an infinite set of random variables $X_1\ X_2 \ldots\ X_n\ \ldots$ (or a set that can be made arbitrarily large, e.g. by multiplying the "observations" of a phenomenon), they cannot be equally correlated unless they are uncorrelated or positively correlated; this conclusion can be interesting because it shows a deep intrinsic difference between the possibilities that a positive correlation and a negative correlation may appear. However, more than for the question's intrinsic interest, the analysis of the present problem is intended to give an example of the usefulness of vector interpretation in suggesting procedures and calculations and making them intuitive. In this regard, we still have to observe that the simple procedure showing that $r(X, Y)$ can assume all the values between –1 and +l was also suggested by a geometrical consideration. Given two unit orthogonal vectors, u and v, it is obvious that $u \cos \alpha + v \sin \alpha$ is a unit vector that forms an angle α with u.

5 On the theorem of Pythagoras

Instead of using the same vector to represent all the random variables whose difference is given by a certain constant, we can adopt a representation in which only the *coincident* random variables are represented by the same vector. If, as the *distance* $d'(X, Y)$ between the random variables X and Y, instead of $\sigma(X - Y)$, we consider $\sqrt{M(X-Y)^2}$, we have $d'(X, Y) = 0$ if and only if $X \doteq Y$. However, this representation is not totally new: since

$$M(X^2) = [M(X)]^2 + [\sigma(X)]^2 \tag{20}$$

we can immediately see that every vector of the new space S' can be decomposed in two components that are orthogonal to one another: the first component represents a fixed number (that is $\bar{X}$), the other a random variable with zero mathematical expectation. In other words, writing $X = \bar{X} + (X - \bar{X})$ the two components $\bar{X}$ and $X - \bar{X}$ are orthogonal. The first is a pure fixed number, or rather the component along the "axis of real numbers", and its "modulus" is the "modulus" or "absolute value" in the usual meaning ($|\bar{X}|$); for the second component we have (by definition) $M(X - \bar{X})^2 = \sigma^2(X)$, so that the hyperplane orthogonal to the axis of real numbers is the space S considered in the preceding section; we can therefore think of it as the projection of the space S' (X and Y are represented on the same vector of S if $X - Y = \alpha$, that is $X - Y$ is parallel to the axis of real numbers along which the projection is made).

Therefore, the observation of Pietra ([4]) that the relationship between $\sqrt{M(X^2)}$, $M(X)$ and $\sigma(X)$ is graphically translatable into the theorem of Pythagoras, assumes a precise geometric meaning in S': the quadratic average $\sqrt{M(X^2)}$ is the modulus of the vector representing X, while $M(X)$ and $\sigma(X)$ are the moduli of the two orthogonal components $\bar{X}$ and $X - \bar{X}$.

[4] See this *Supplemento*, A. II, no. 2-3, 1936 (*Il teorema di Pitagora e la Statistica*).

6 Remarks on special cases

It may be of interest to add some remarks on special cases.

The simplest case, that of *events*, introduces the notable particularity that, for these, *zero correlation* implies stochastic *independence*. Let E' and E'' be two events and $E = E'E''$ their simultaneous occurrence. The probabilities are $p' = P(E')$, $p'' = P(E'')$, $p = P(E)$, and the standard deviations are $\sigma' = \sqrt{p'q'}$, $\sigma'' = \sqrt{p''q''}$ (as usual, $q' = 1 - p'$, $q'' = 1 - p''$). We thus have $p = p'p'' + r\,\sigma'\sigma''$ and $p = p'\,p''$ [ix] (stochastic independence) if and only if $r = 0$. In general, the equivalence between the two concepts of zero correlation and independence subsists when each of the two random variables X and Y is allowed to assume only one out of two possible values, x_1 and x_2, y_1 and y_2 (strictly speaking: if they "coincide" with two random variables X' and Y' that have such a property: $X \doteq X'$, $Y \doteq Y'$). In any other case (we will soon prove it) the concept of independence is actually more restrictive than that of no correlation.

In the case of events, we can give a formula for r that reveals its meaning from another point of view. Let $p'(1 + \rho)$ be the probability $P(E'/E'')$ of E' conditioned by E': then $\rho = 0$ in the case of independence, $\rho > 0$ and $\rho < 0$, respectively, in the case of positive or negative correlation. In addition $P(E''/E') = p''(1 + \rho)$ (in essence, this gives Bayes theorem). We can thus write

$$p = p'\,p''(1+\rho) = p'\,p'' + r\,\sigma'\sigma'' \tag{21}$$

so that

$$r\,\sigma'\sigma'' = \rho\,p'\,p'' \tag{22}$$

In addition, if (using "$\overline{E}$" to denote the negation of "E") we set

$$q' = P(\overline{E}'), \quad q'' = P(\overline{E}''), \quad q = P(\overline{E}'\,\overline{E}''), \quad q = q'q''(1+\tau), \tag{23}$$

we obtain

$$q = q'\,q''(1+\tau) = q'\,q'' + r\,\sigma'\sigma'' \tag{24}$$

and thus

$$r\,\sigma'\sigma'' = \tau\,q'\,q'' \tag{25}$$

By multiplying these two expressions we obtain

$$r^2\,\sigma'^2\,\sigma''^2 = \rho\,\tau\,p'q'\,p''q'', \quad \text{but} \quad \sigma'^2 = p'q', \quad \sigma'' = p''q'', \tag{26}$$

and finally

$$r^2 = \rho\,\tau \tag{27}$$

(and we immediately see that the sign of r is equal to that of ρ and τ, which necessarily have the same sign). Therefore, the correlation coefficient expresses the geometric mean of two coefficients: the coefficient of increase (or decrease) of the probability of an event occurring if the other occurs and the coefficient of increase (decrease) of the event not occurring, if the other does not occur.

Another related remark is that, once the *probabilities* p' ad p'' have been *assigned*, the correlation coefficient r cannot vary between –1 and +1, but only between –

$\sqrt{p'p''/q'q''}$ and $+\sqrt{p'q''/q'p''}$ (under the hypothesis that $p' < p''$, $p' + p'' \leq 1$; in the other cases we should exchange p' with p'' and the p-s with the q-s, respectively). The lower bound is –1 only if $p' + p'' = 1$, the upper bound is +1 only if $p' = p''$ and thus only for $p' = p'' = ½$ are the two bounds ±1. But the two limits, let's call them r_1 ad r_2, *may* be close to zero (it is sufficient to take a very small p'); more generally, once we arbitrarily choose r_1 and r_2, provided, as it is necessary, that $-1 \leq r_1 < 0 < r_2 \leq 1$, we can determine p' and p'' in such a way that

$$r_1 = -\sqrt{p'p''/q'q''}, \quad r_2 = +\sqrt{p'q''/q'p''} \tag{28}$$

(the solution is unique, and is given by $p' = -r_1 r_2 / (1 - r_1 r_2)$, $p'' = -r_2 / (r_2 - r_1)$).

7 Bounds, zero correlation and independence

The conclusions we reached in the special case of "events" show that, in general, when *the probability laws* of two random variables X and Y *have been assigned*, it is no longer possible for $r(X, Y)$ to assume an arbitrary value between –1 and +1; after all it is obvious, since we must have $X = a\ Y + b$ ($a > 0$) and $X = a\ Y + b$ ($a < 0$), that the two bounds can be reached only if the two distribution functions Φ_1 and Φ_2 are, respectively, similar ($\Phi_1(\xi) = \Phi_2(a\xi + b)$, $a > 0$), "anti-similar" ($\Phi_1(\xi) = = 1 - \Phi_2(a\xi + b)$, $a < 0$), or (in order to attain both bounds) similar and symmetrical ($\Phi_1(\xi) = \Phi_2(a\xi + b) = 1 - \Phi_2(-a\xi + b)$).

In general, the two extreme values r_1 and r_2 can be achieved if we consider the two extreme cases in which one of the two random variables X and Y is a decreasing or increasing function of the other; in the latter case, if X assumes the value ξ such that $\Phi_1(\xi) = t$, Y assumes the value η such that $\Phi_2(\eta) = t$; in the opposite case, to ξ such that $\Phi_1(\xi) = t$ corresponds η such that $\Phi_2(\eta) = 1 - t$. Using $\xi(t)$ and $\eta(t)$ to denote the values for which we have, respectively, $\Phi_1(\xi) = t$ and $\Phi_2(\eta) = t$ (inverse functions), the correlation coefficients for the two extreme cases we considered are

$$r = \int_0^1 \xi(t)\eta(t)\,dt \qquad r = \int_0^1 \xi(t)\eta(1-t)\,dt \tag{29}$$

provided that we consider "normalized" X and Y so that their mathematical expectation is zero and their average standard deviation is equal to 1. ([5])

To show that the cases considered truly achieve the extreme values r_2 and r_1, first of all, observe that, if the support of the probability distribution in the plane ξ, η consisted of two points ξ_1, η_1 and ξ_2, η_2 with $\xi_2 > \xi_1$ but $\eta_2 < \eta_1$, to which we assign equal probability, then the correlation would increase by the quantity $p(\xi_2 - \xi_1)(\eta_1 - \eta_2)$ if we displace the support of the two probabilities to the points ξ_2, η_1 and ξ_1, η_2, and this does no change the distributions Φ_1 and Φ_2. Except for some further refinement and corrective terms that can be made negligible, the same reasoning holds for probabilities not concentrated in exactly the two points ξ_1, η_1 and ξ_2, η_2, but supported in their neighborhood. As long as $\xi(t)$ is not equal to $\eta(t)$, we can increase r with the above-mentioned inver-

[5] Note that $\int \xi\eta\,dt$ can be interpreted as cosine of the angle between the two functions ξ and η in the functional space, so that the upper bound of $\cos(X, Y)$ is the cosine of $\xi(t)$ and $\eta(t)$ in the functional space. Similar considerations for the minimum link the metric for the space of random variables to the functional space.

sion procedure, and analogously we can diminish r until $\xi(t)$ is equal to $\eta(1-t)$.

Finally, we close by completing a proof we left unfinished: that zero correlation and independence coincide only if both Φ_1 and Φ_2 have all their probability mass concentrated on two points. Let's choose a value a (between 0 ad 1) in such a way that

$$\begin{cases} \eta(t) & \text{if} \quad t \le a \\ \eta(1+a-t) & \text{if} \quad t > a \end{cases} \tag{30}$$

corresponds to $\xi(t)$.

We have $$r(a) = \int_0^a \xi(t)\eta(t)dt + \int_a^1 \xi(t)\eta(1+a-t)dt, \tag{31}$$

and, as a varies between 0 and 1, $r(a)$ varies continuously between the extremes r_2 and r_1. For a certain value $a = a_0$ we have $r(a_0) = 0$. We immediately see that such a distribution, is uncorrelated and does not coincide, except in the case mentioned earlier, with the distribution in the case of independence. We can also note that, for every r between r_1 and r_2, the distribution is unequivocally determined for Φ_1 and Φ_2 having probabilities concentrated in only two points, while in any other case we have infinite solutions.

8 Terminology

In this note I have always used the term "correlation" in the meaning pertaining to the "correlation coefficient". But, as we noted at the beginning, this term has been used with many different meanings, and there are different opinions on the terminology that would be best to adopt to avoid any possible misunderstanding. We will briefly examine the matter, the proposals that have been made and those that could be made.

As noted, one of the meanings in which the term "correlation" has been used is that corresponding to the "stochastic dependence" of the calculus of probabilities; to measure degrees of "dependence" (or rather, as it is always better to say when dealing with indices, "to determine a number that gives an idea of such a degree of *dependence*") involves defining an index with the property that its extreme values are assumed in the two extreme cases in which X and Y are stochastically independent or are functions one of the other, and intermediate values in the other cases, closer to this or that extreme, depending on how strict, based on a certain (largely arbitrary) criterion, this *dependence* is. Unlike the correlation coefficient r, whose basic element is the *sign*, the dependence index can only vary between 0 and 1 (when functional independence and dependence corresponds to the values 0 and 1).

As regards the terminology, we could:

- decide to preserve the term "correlation" in the sense of "stochastic dependence" in which it was improperly used, on the other hand agreeing to abandon it in every different sense; or
- use the term "dependence", adding on the specification "stochastic", as is the usage in the calculus of probabilities; or finally
- introduce a new and different term.

This last solution seems the best, because the term "correlation" is more appropriate in

the sense in which we used it here, while that of "dependence", if we don't want always add the term "stochastic", may often cause ambiguity with the concept of functional dependence. This drawback has become a sensitive issue since for some time now, in the calculus of probabilities, we apply both meanings of dependence to random variables; besides, it is not desirable for such disparate notions to be denoted by the same word, leaving to an adverb the duty to distinguish between them. This solution is appropriate when we need to distinguish between several particular variants of the same notion, as in the locutions "linearly dependent", "algebraically dependent", that refer to particular aspects of "dependence" in only one sense, that of the analysis.

However to find a new word is not an easy task. On the one hand we want a word that makes the meaning intuitive, as is necessary since the term refers to a concept of current use among practitioners (for instance, in the insurance industry) and should become part of every day language. On the other hand, the word should enjoy many of the grammatical possibilities that are intrinsic in the word *dependence* and that are useful, if not necessary, in the calculus of probabilities. In fact it generates, both for the affirmation and for the negation, the name, verb, adverb, and adjective, usable in transitive or reflexive sense (see the following table)

X dependent on Y	*X independent of Y*
X and *Y* (*mutually*) *dependent*	*X* and *Y* (*mutually*) *independent*
dependence	*independence*
X depends on Y	*X does not depend on Y*
X, dependently on Y,	*X, independently of Y,*

The term "connection", proposed by Gini and Pietra, gives the adjective *connected* that may be used in both ways (*connected* (among themselves), and *connected with*), but the verb and it's negation seem to me unusable (*to connect*, *disconnected*). As for the meaning, *connection* recalls something rigid, and to say that some risks are connected does not seem to me well suited to the idea.

One could propose "influence" (*influenced by*, *influencing*, *mutually influencing*, *influences*, *they influence each other*); as regards the meaning, I would consider such a word as the ideal solution, because it renders impeccably the precise meaning of "stochastic dependence", but it has the serious defect of not possessing a negated form: the negation with the "not" (*mutually not influencing*, etc.) would be rather heavy, even more than *mutually influencing*.

One could propose "tie" (*tied* (to each other), *tied to*) that has grammatical deficiencies analogous to those of "connection", but it seems to me more consistent with the meaning: "to tie" is to unite but in more elastic way than "to connect". Besides, saying for instance that two risks are tied to each other, it seems to me that the meaning is clear independently of the possible convention to introduce such a terminology in the scientific language. After all, the French already use the term "loi liée" to point out the probability distribution of a random variable corresponding to a certain value of another: if such a random variable is said to be "tied to", the expression "tied to distribution" (translation of "loi liée") would be by itself connected to the term explaining the general concept.

As a final comment, we may remark that "correlation" (which we have excluded for reasons of another nature) would not be grammatically opportune.

Does some other appropriate but flexible word exist for our purposes? It would be worthwhile to look for it before taking a decision: however, having to choose among the terms considered, I would give my preference to one of the last two (*influence* or *tied to*).

9 Concordance vs correlation

The other meaning is that for which Gini and Pietra have proposed the term "*concordance*" (respectively, *discordance*, *indifference*). Such a concept is analogous to, but more general than, that of strict correlation (corresponding to *r*). It is analogous because it concerns the tendency of a random variable to generally assume a greater or smaller value according to the greater or smaller value assumed by the other. We thus have to distinguish the *direction* (concordance or discordance) as in the case of correlation (positive or negative). But the concept of concordance is more general because it generically covers all the aspects under which such a tendency can be considered and studied (aspects that can be measured by other indices that have been proposed — as those of *omofilia*).

Given the role that the correlation coefficient *r* plays in the calculus of probabilities (which I have tried to shed some light on), I believe that it would be a good idea to follow the proposal of Gini - Pietra by adopting the term "concordance" for the general meaning, but reserving that of "correlation" only for the meaning corresponding to *r*. Otherwise the term "positively correlated" would not have any defined meaning, because it is not necessarily the case that another index of concordance must always take a positive, negative or zero value depending on whether *r* is positive, negative or zero.

Instead, according to Fréchet, it would be a good idea to call *r* "index of linearity" (he would reserve the term "correlation" for the meaning of "stochastic dependence"). However, such a terminology would not allow us to deduce the equivalent of the expressions "correlated", "positively or negatively correlated", "not-correlated", terms that, as we have seen, are essential in the calculus of probabilities. Besides, it seems to me that it corresponds to a rather limited idea of the meaning of *r*, that would indicate only the tendency towards a "linear regression". Instead, the fundamental meaning of *r*, which we have tried to shed light on, is completely independent of whether the probability distribution tends to be concentrated along a given line, and of whether this, when it exists, is a straight line or a curve. The only difference is that, if such line is straight, *r* can give an idea of the accumulation of the distribution around it; if such line does not exist, no misunderstanding can arise, while if it exists and is curved we need to be warned that *r* does not allow the conclusion that subsisted in the case of linear regression.

10 An index of concordance

One final remark is in order. Its purpose is to clarify that a concordance index can have a sign different from that of *r*, and to suggest what seems to me the simplest and

most intrinsically meaningful index of concordance (which has not yet been considered, as far as I know). To express it in the simplest way, let us consider a particular case: given the distribution of marriages according to the ages of the consorts, let us consider the probability that, choosing two couples at random and independently, the younger girl is the wife of the younger boy. Indicating by X_1 and Y_1, X_2 and Y_2 the age of the bridegroom and his bride in the two couples, we have to consider the probability c that

$$(X_1 - X_2)(Y_1 - Y_2) > 0\ , \tag{32}$$

and, if $\varphi(x, y)\, dx\, dy$ is the probability that the ages of the consorts are between x and $x + dx$, y and $y + dy$, respectively, we can write

$$c = \int_C \varphi(x,y)\,\varphi(\xi,\eta)\,dx\,dy\,d\xi\,d\eta \tag{33}$$

where C is the domain defined by $(x - \xi)\,(y - \eta) > 0$, which is bounded by the planes $x = \xi$, $y = \eta$. We can consider c as an index of concordance: we would have concordance, discordance or indifference if c is greater than, smaller than or equal to ½, respectively; we would have $c = 1$ and $c = 0$ only in the extreme cases in which the age of the bride is an increasing or decreasing function, respectively, of the bridegroom's age.

It is easy to see that c does not necessarily have the same sign as r: it is sufficient to note that c remains unchanged if we substitute X and Y with two increasing functions, $f(X)$, $g(X)$ (in fact, in this case $[f(X_1) - f(X_2)]\,[g(Y_1) - g(Y_2)]$ has the same sign as $[X_1 - X_2]\,[Y_1 - Y_2]$) while $r(X, Y)$ and $r[f(X), g(Y)]$ may well have different signs ([6]).

Despite the difference of behavior, c and r share a common meaning that it is worth explaining. If, besides the sign, we also want to consider the value of $(X_1 - X_2)\,(Y_1 - Y_2)$, to give a weight proportional to the differences $X_1 - X_2$ and $Y_1 - Y_2$, we should substitute the integral for c with the following

$$\begin{aligned}\int (x-\xi)(y-\eta)\varphi(x,y)\varphi(\xi,\eta)\,dx\,dy\,d\xi\,d\eta = \\ = M[X_1Y_1 + X_2Y_2 - X_1Y_2 - X_2Y_1]\end{aligned} \tag{34}$$

and it is easy to see that this expression is equal to

$$2\sigma(X)\sigma(Y)r(X,Y). \tag{35}$$

Therefore, the meaning of c can be considered analogous to the meaning of r, when we observe the *direction* of the inequalities but omit their *order of magnitude*.

---

[6] For instance, if the possible values for (X, Y) are

(–1,1)	with probability	1/4
(0, –1)	"	" 1/2
(a, 1)	"	" 1/4

we have $r(X, Y) = 0$ when $a = 1$; changing a (between 0 and ∞, extremes excluded), that is by substituting X with its increasing function $g_a(X) = X + [(a - 1)/2]\,X\,(X + 1)$,
we get $r > 0$ when $a > 1$, $r < 0$ when $a < 1$.

References

de Finetti, B. (1930), *Spazi astratti metrici* (D_M), Atti Accad. Pontificia, LXXXIII, VI, 1930.

Pietra, G. (1936): *L'ostracismo al coefficiente di correlazione?*, Supplemento Statistico Nuovi Problemi, II, 2-3, 1-7, 1936.

Pietra, G. (1936): *Il teorema di Pitagora e la Statistica*, Supplemento Statistico Nuovi Problemi, II, 2-3, 7-9, 1936.

Editors' notes

---

[i] Since the quadratic function $\sigma^2(Z_t) = at^2 + bt + c$ must always be non negative, its discriminant, $D = b^2 - 4ac$, must be non-positive. In this case, $a = \sigma^2(Y)$, $b = 2M(X - \bar{X})(Y - \bar{Y})$, $c = \sigma^2(X)$. By substituting these values in $b^2 - 4ac \leq 0$, the Author gets the inequality (5). The Italian version has an incorrect $|M[(X - \bar{X})(Y - \bar{Y})]| < \sigma(X)\sigma(Y)$. We corrected the typo.

[ii] If, in Equation (1), $n \equiv 2$, $a_{ij} \equiv 1$, $X_i \equiv X$, $X_j \equiv Y$, then $Z = XY$. In this case, Equation (3) gives $M(XY) = \bar{X}\bar{Y} + M[(X - \bar{X})(Y - \bar{Y}) = M(X)M(Y) + M[(X - \bar{X})(Y - \bar{Y})$. Substituting Equation (6) in the last expression gives Equation (7).

[iii] If $r(X, Y) = r = \pm 1$ in (11) then $Y = \pm X$ in (9).

[iv] If $r(X, Y) = \pm 1$ then $M[(X - \bar{X})(Y - \bar{Y})] = \pm\sigma(X)\sigma(Y)$ in (4). Therefore, by using the notation introduced in endnote i, $b = 2M(X - \bar{X})(Y - \bar{Y}) = \pm 2\sigma(X)\sigma(Y)$ and $t = (-b \pm \sqrt{b^2 - 4ac})/(2a) = [\mp 2\sigma(X)\sigma(Y)]/[2\sigma^2(Y)] = \mp\sigma(X)/\sigma(Y)$.

[v] The Italian version has an incorrect $X - \bar{Y} \doteq k(Y - \bar{Y})$. We corrected the typo.

[vi] This formula follows from (8), page 4.

[vii] The Italian version has an apparently incorrect $\pi - \alpha(X, Y)$. We corrected the typo.

[viii] The Italian version has an incorrect $X_i = X_i - Y$. We are grateful to an anonymous referee for pointing out this typo.

[ix] The Italian version has an incorrect $p = p\,p$". We corrected the typo.