How Negative Can the Product-Moment Correlation Coefficient be?

Question raised by
R Vasudeva
University of Mysore
Mysore 570 006, India

When Pearson's product-moment correlation coefficient is introduced, say in an undergraduate class, it is shown by elementary techniques that its range is [–1, 1]. It is possible to have three random variables X, Y and Z such that any pair has a correlation of +1. For instance, if $Y=2X$ and $Z=X-7$, then the correlation between X, Y, between Y, Z and between Z, X will all be +1. One might wonder if it is possible for X, Y, Z to be such that the correlation between any pair of them is –1. To answer this question, assume without loss of generality that the variances of X, Y and Z are all unity. (We can make a linear transformation of each of them so that the variance is unity and these transformations do not affect the correlation coefficients. Hence no loss of generality.) Then covariance and correlation will be the same for any of these pairs. Now consider a new random variable $X+Y+Z$. Then

$$V(X+Y+Z) = V(X) + V(Y) + V(Z) + 2[Cov(X, Y) + Cov(Y, Z) + Cov(Z, X)] = 3 + 2(\rho_{XY} + \rho_{YZ} + \rho_{ZX}),$$

where ρ denotes the correlation coefficient between the random variables in the suffix. If all these ρ's here are –1, then the variance of $X+Y+Z$ will be –3; hence it is not possible for all the three correlations to be –1. Since $V(X+Y+Z)$ has to be ≥ 0, $\rho_{XY} + \rho_{YZ} + \rho_{ZX} \geq -(3/2)$. Even the values of –0.8, –0.9, –0.7 are not possible for these three correlation coefficients. The average value of these correlation coefficients cannot be less than –(1/2). If the three correlation coefficients are equal, say ρ, then ρ should be $\geq -(1/2)$. Note that the condition $V(X+Y+Z) \geq 0$ is only a necessary condition on the correlations. Suppose that $\rho_{XY} = -0.8$, $\rho_{YZ} = -0.4$ and $\rho_{ZX} = -0.27$. Then $V(X+Y+Z) \geq 0 \Leftrightarrow \rho_{XY} + \rho_{YZ} + \rho_{ZX} \geq -(3/2)$ and this condition is satisfied. But $V(\sqrt{2}X + \sqrt{2}Y + Z/\sqrt{2}) \geq 0 \Leftrightarrow 2\rho_{XY} + \rho_{YZ} + \rho_{ZX} \geq -(9/4)$ and this condition fails to hold, making the variance equal –1.54. Thus

the above numbers cannot represent the correlation coefficients as stated. Thus $V(X+Y+Z) \geq 0$ is not a sufficient condition on the correlations in the general case.

More generally, consider p random variables $X_1, X_2, \ldots, X_p$, each with variance unity. By a similar argument, one can see that the average value of these correlation coefficients cannot be less than $-1/(p-1)$. And, if all the correlation coefficients are the same, say, ρ, then $\rho \geq -(1/(p-1))$. Thus it is seen that the larger the number of random variables, the less negative can the correlations between them be. In the limit, as we let $p \to \infty$, that is, if we consider an infinite sequence of random variables, with the same common correlation coefficient ρ between any pair of them, then this ρ cannot be negative.

The above conclusions were derived by considering only one linear function $X+Y+Z+\cdots$ of random variables concerned. What if we consider other linear functions and exploit the fact that their variances should also be non-negative? Do we then obtain even sharper bounds for the correlation coefficients? Let us consider the class of linear functions of the form $aX+bY+cZ$ of three random variables X, Y, Z, each with unit variance. The non-negativity of the variance of all such linear functions is simply a property of the *correlation matrix*, which is a symmetric matrix with diagonal elements unity and the off-diagonal elements given by the correlation coefficients. The variance of $aX+bY+cZ$ is

$$a^2 + b^2 + c^2 + 2(ab\rho_{xy} + bc\rho_{yz} + ca\rho_{zx}).$$

The property that this is non-negative for all choices of a, b, c is exactly the notion of non-negative-definiteness (*nnd*) of a (symmetric) matrix. Thus in the language of matrices, the bounds for the correlation coefficients are those imposed by the requirement that the correlation matrix is *nnd*. In particular, when all the correlation coefficients are equal, this requirement has been totally exploited by $V(X+Y+Z) \geq 0$ ($\Rightarrow \rho \geq -1/2$) and $V(X-Y) \geq 0$ ($\Rightarrow \rho \leq 1$). Consideration of $V(aX+bY+cZ) \geq 0$ for any other set of constants a, b and c does not add any more

restriction. Although correlation coefficients in the case of several variables cannot all be highly negative, it is possible that the correlation coefficients between X and Y, between Y and Z and between X and Z are all negative dispelling the notion that if X and Y and Y and Z have negative correlation coefficients then X and Z should have positive correlation coefficient. A numerical example demonstrating this phenomenon is given below. Here the random triple (X, Y, Z) can take five possible triples of values given below, each with probability 1/5.

X	Y	Z
1	−1	1
1	0	−1
−2	1	0
−1	0	1
1	0	−1

The correlation coefficients are $\rho_{XY}=-(3/4)$; $\rho_{XZ}=\rho_{YZ}=-(1/2\sqrt{2})$, showing that all three coefficients can be negative.

The algebra of matrices can be used in another way (at least apparently) to establish the bounds obtained in the earlier paragraphs. For this, consider the notion of the characteristic roots (also called latent roots or eigenvalues) of a square matrix. The characteristic roots of a $p\times p$ matrix $\boldsymbol{A}$ are defined to be the roots of the polynomial (of degree p) $\det(\boldsymbol{A}-\lambda I)$ in λ, where I is the identity matrix. In general, not all the p roots are real. However, it can be shown that if the matrix is symmetric like in the case of the correlation matrix, the roots are all real. It can be further shown that if the matrix is *nnd*, then the characteristic roots are non-negative. The characteristic roots in the case of three random variables with equal correlations ρ can be easily worked out, by solving the cubic equation, to be $\lambda_1=1+2\rho$, $\lambda_2=\lambda_3=1-\rho$. Using the fact that $\lambda_1, \lambda_2, \lambda_3 \geq 0$, we obtain the bounds $-(1/2) \leq \rho \leq 1$.

As in the problem discussed above, the basic theory of statistics involving several random variables, their variances, and correlations among them, can be worked out very elegantly and clearly with the use of the algebra of matrices.