USING CORRELATION IN STUDIES OF STUDIES

Alan Safer and Saleem Watson

Abstract. Let X, Y, and Z be random variables. If X is positively correlated to Y and Y is positively correlated to Z, it does not necessarily follow that X is positively correlated to Z. In this article we find ranges for the correlation coefficients r_{XY} and r_{YZ} that guarantee that X and Z have a specified level of correlation. We explore the implications of these results to finding relationships between variables investigated in different studies.

1. Introduction. Suppose X, Y, and Z are random variables with X positively correlated to Y, and Y positively correlated to Z. It is very tempting to conclude that X is positively correlated to Z. In fact, this conclusion is not always valid; it is possible for the correlation coefficients r_{XY} and r_{YZ} to be positive but for X and Z to be totally uncorrelated ($r_{XZ} = 0$), or even negatively correlated. (Recall that the correlation coefficient is a number between 1 and -1 and is a measure of the strength of linear association between two variables.) On the other hand, we will show that if the correlation coefficients r_{XY} and r_{YZ} are very strong (both near 1 or -1), then it is possible to conclude that there is a positive correlation between X and Z. Specifically, we find ranges for the different correlation coefficients that guarantee that X and Z have a desired level of correlation. We also explore the implications of these results in "studies of studies," that is, in attempting to find relationships among variables researched in different studies.

Consider the following example. A logger wishes to estimate the height of a pine tree in a forest. She recalls that in high school she learned a method for finding the height of a tree using the length of its shadow and the angle of elevation of the sun. But in the forest in which she works, it's difficult or impossible to find the shadow of a particular tree. She reasons that it's easy to measure the diameter of a tree, and that the diameter is related to the height. After some research in the library she finds an article that gives a positive correlation between the diameter D and the age A of a pine tree and another article that gives a positive correlation between the age A and the height H. This is rather frustrating because what she wants is to relate diameter to height (D to H). As in many research studies the actual data is not published, so she can't calculate the correlation between D and H directly. But even if the data for each study are available it's probably not for the exact same trees. So what is she to do?

This story points out a common situation. In studying the published research on a specific topic, connections between certain properties of interest may not be directly studied. In this example, the relationship between

diameter and height is not directly studied. Is it necessary for the logger to conduct a field study herself, or can she somehow use the information in the studies already available? In other words, can one make new connections by studying already available studies?

2. Positive Correlation Is Not Transitive. The general situation typified by the above example is as follows: If X is correlated to Y, and Y is correlated to Z, then how is X correlated to Z? In other words, how much information about the data is encapsulated in the single number, the correlation coefficient? To help give an answer to this question, consider the following data.

Study 1		
X	Y	Z
1	5	8
2	0	1
4	8	8
9	7	8
3	7	2

Study 2		
X	Y	Z
7	7	4
4	1	5
5	9	9
6	3	0
5	2	3

For the data in Studies 1 and 2 we have

$$\begin{array}{llll} \text{Study 1} & r_{XY} = .46 & r_{YZ} = .61 & r_{XZ} = .39 \\ \text{Study 2} & r_{XY} = .46 & r_{YZ} = .61 & r_{XZ} = -.36 \end{array}$$

This example shows that for different sets of data X, Y, Z we can have identical correlations for r_{XY} and r_{YZ}, but vastly different values for r_{XZ}. Moreover, it's possible that r_{XY} and $= r_{YZ}$ are positive whereas r_{XZ} is negative. So the property of being positively correlated is not transitive.

3. How is $\mathbf{r_{XZ}}$ related to $\mathbf{r_{XY}}$ and $\mathbf{r_{YZ}}$? In general, r_{XY} and r_{YZ} determine a range of possible values for r_{XZ} [1, 2]. We can see this by considering X, Y, Z as vectors in Euclidean space.

Suppose we have n-data points in a study

$$X = (x_1, x_2, \ldots, x_n) \text{ and } Y = (y_1, y_2, \ldots, y_n).$$

Without loss of generality assume that each of these data sets has mean zero, that is $E(X) = E(Y) = 0$. By definition, the correlation coefficient of X and Y [3] is

$$\begin{aligned} r_{XY} &= \frac{E(XY) - E(X)E(Y)}{\sigma(X)\sigma(Y)} \\ &= \frac{(x_1y_1 + x_2y_2 + \cdots + x_ny_n)/n}{\sqrt{(x_1^2 + x_2^2 + \cdots + x_n^2)/n} \cdot \sqrt{(y_1^2 + y_2^2 + \cdots + y_n^2)/n}} = \frac{X \cdot Y}{||X|| \cdot ||Y||}. \end{aligned}$$

In the last equality $X \cdot Y = x_1y_1 + x_2y_2 + \cdots + x_ny_n$ is the dot product of the vectors X and Y, and $||X|| = \sqrt{(x_1^2 + x_2^2 + \cdots + x_n^2)}$, $||Y|| = \sqrt{(y_1^2 + y_2^2 + \cdots + y_n^2)}$ denote the lengths of the vectors X and Y. It follows that $r_{XY} = \cos\alpha$, where α is the angle between the vectors X and Y (so $0 \le \alpha \le \pi$) [4]. Similarly, $r_{YZ} = \cos\beta$ and $r_{XZ} = \cos\gamma$. Figure 1 shows the vectors X, Y, and Z and the angles between them.

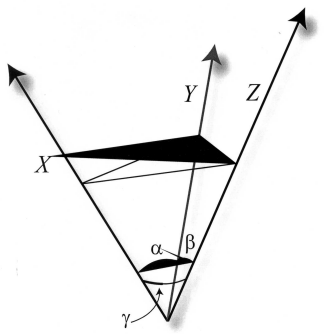

Figure 1.

From Figure 1 we see that for fixed α and β the largest and smallest possible values for γ occur when the vectors X, Y, and Z lie in the same plane. So that the largest possible angle is $\gamma = \alpha + \beta$ and the smallest is $\gamma = |\alpha - \beta|$.

Let us assume that r_{XY} and r_{YZ} are positive. Then $\alpha = \cos^{-1} r_{XY}$ and $\beta = \cos^{-1} r_{YZ}$ are between 0 and $\pi/2$. Using the formula for the cosine of a sum we have

$$\begin{aligned}\cos(\alpha+\beta) &= \cos\alpha\cos\beta - \sin\alpha\sin\beta \\ &= \cos\alpha\cos\beta - \sqrt{1-\cos^2\alpha}\sqrt{1-\cos^2\beta} \\ &= r_{XY}r_{YZ} - \sqrt{1-r_{XY}^2}\sqrt{1-r_{YZ}^2}.\end{aligned}$$

We have used the positive sign for the square roots because α and β are acute angles. Similarly, using the formula for the cosine of a difference we get

$$\cos(\alpha-\beta) = r_{XY}r_{YZ} + \sqrt{1-r_{XY}^2}\sqrt{1-r_{YZ}^2}.$$

Now, since $0 \le |\alpha - \beta| \le \gamma \le \alpha + \beta$, and since cosine is decreasing on $[0, \pi]$ we have $\cos(\alpha+\beta) \le \cos\gamma \le \cos(\alpha-\beta)$, and we get the inequalities

$$r_{XY}r_{YZ} - \sqrt{1-r_{XY}^2}\sqrt{1-r_{YZ}^2} \le r_{XZ} \le r_{XY}r_{YZ} + \sqrt{1-r_{XY}^2}\sqrt{1-r_{YZ}^2}. \tag{3.1}$$

It is easy to check that these inequalities hold in the remaining cases, that is, if both r_{XY} and r_{YZ} are negative, or if one of r_{XY}, r_{YZ} is negative and the other positive.

The inequalities in (3.1) have the following interesting special cases. If one of r_{XY} or r_{YZ} is equal to 1, say $r_{XY} = 1$, then the inequalities reduce to the equality $r_{YZ} = r_{XZ}$. If both $r_{XY} = 1$ and $r_{YZ} = 1$, then the inequalities imply that $r_{XZ} = 1$. In other words, as we would expect, if X and Y are perfectly linearly correlated and Y and Z are also perfectly linearly correlated, then so are X and Z. On the other hand, if one of the correlation coefficients is 0, say $r_{XY} = 0$, then the inequalities in (3.1) become $-\sqrt{1 - r_{YZ}^2} \leq r_{XZ} \leq \sqrt{1 - r_{YZ}^2}$; in particular, 0 is a possible value for r_{XZ}. Finally, if both $r_{XY} = 0$ and $r_{YZ} = 0$, then the inequalities in (3.1) become $-1 \leq r_{XZ} \leq 1$. In other words, if X and Y are totally uncorrelated and Y and Z are totally uncorrelated, then any level of correlation is possible for X and Z.

4. Bounds for the Correlation Coefficient r_{XZ}. If we have the data that relates X to Y and Y to Z (as in Study 1) then the correlation coefficient r_{XZ} can be calculated directly from the data. In practice we may have different sets of data relating these variables. For example, the studies on pine trees may be made on different trees, possibly with different sample sizes. But if the studies were made on pine trees from the *same population,* then it is reasonable to assume that the calculated correlation coefficients are representative of the population as a whole. Then we can use inequality (3.1) to determine bounds for the correlation coefficient of X and Z.

Inequalities (3.1) have their obvious use: given r_{XY} and r_{YZ} we can find bounds for r_{XZ}. But we use inequalities (3.1) in a different way. Namely, if we want a desired level of correlation for r_{XZ} we can use these inequalities to find the possible pairs (r_{XY}, r_{YZ}) that guarantee that level of correlation. These pairs will be expressed as a region within the square $S = [-1, 1] \times [-1, 1]$ in the coordinate plane. We consider the situation in two cases. For simplicity of notation we let $a = r_{XY}$, $b = r_{YZ}$.

Case 1. Suppose we require that r_{XZ} have a value at least k ($0 \leq k \leq 1$). In this case the "worst case scenario" for the correlation coefficient r_{XZ} is determined by the left-hand side of inequality (3.1). So, we must have

$$k \leq ab - \sqrt{1 - a^2}\sqrt{1 - b^2}. \tag{4.1}$$

When equality holds in (4.1), we can rearrange, square, and simplify to get

$$a^2 - 2kab + b^2 = 1 - k^2. \tag{4.2}$$

This is the equation of a rotated ellipse with eccentricity $2\sqrt{k}/(1 + k)$ and major axis along the line $a = b$ [5]. Note that (4.1) implies that both $|a| \geq k$ and $|b| \geq k$. To see this, write (4.1) as $\sqrt{1 - a^2}\sqrt{1 - b^2} \leq ab - k$, so $ab - k$

must be nonnegative. Thus, $0 \leq= k \leq ab$, and so a and b are either both positive or both negative. Since $|a| \leq 1$ and $|b| \leq 1$, it follows $ab \leq |a|$ and $ab \leq |b|$, and the result follows. So the solution of inequality (4.1) is the region inside the square S, outside the ellipse (4.1), with $|a| \geq k$ and $|b| \geq k$.

Case 2. If we require that r_{XZ} have a value less than $-k$ $(0 \leq k \leq 1)$, then the "worst case scenario" for r_{XZ} is determined by the right-hand-side of inequality (3.1). So, we must have

$$ab + \sqrt{1-a^2}\sqrt{1-b^2} \leq -k. \tag{4.3}$$

The equality in (4.3) determines the ellipse

$$a^2 + 2kab + b^2 = 1 - k^2 \tag{4.4}$$

with eccentricity $2\sqrt{k}/(1+k)$ and major axis along the line $a = -b$. As in the preceding case we have $|a| \geq k$ and $|b| \geq k$. So the solution to inequality (4.3) is the region inside the square S, outside the ellipse (4.4), with $|a| \geq k$ and $|b| \geq k$.

The situation is illustrated graphically for $k = 0.6$ in Figure 2. If we require that $r_{XZ} \geq 0.6$ then (r_{XY}, r_{YZ}) must lie in the first or third quadrants inside the square S, outside the ellipse, with $|r_{XY}| \geq 0.6$ and $|r_{YZ}| \geq 0.6$. This is the shaded region in the first and third quadrants. If we require that $r_{XZ} \leq -0.6$ then the pair (r_{XY}, r_{YZ}) must lie in the corresponding shaded regions in the second and fourth quadrants.

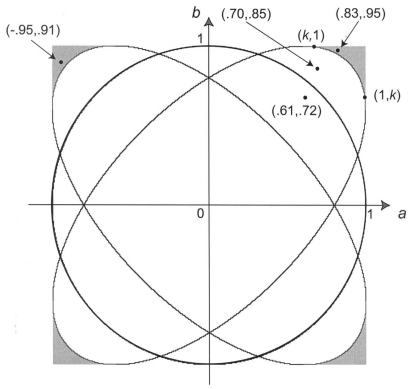

Figure 2. Regions determined by $k = 0.6$.

In each of the above cases, as k gets closer to 0 the ellipses in Figure 2 have smaller eccentricity. In the extreme case $k = 0$ the ellipses reduce

to the unit circle. So, to guarantee $r_{XZ} > 0$ or $r_{XZ} < 0$ we must have the pair (r_{XY}, r_{YZ}) inside the unit square but outside the unit circle, in the appropriate quadrants. Note that the points inside the unit circle do not belong to either case (for any value of k). So, if (r_{XY}, r_{YZ}) is inside the unit circle no information can be obtained about that r_{XZ} (not even its sign).

Also, in each of the above cases, as k gets closer to 1 or -1, the ellipses in Figure 2 have larger eccentricity and the shaded regions become smaller. In the extreme case $k = \pm 1$, the solutions of inequalities (4.1) and (4.3), are single points (the corners of the square S). In other words, if we require r_{XZ} to be perfectly correlated ($k = \pm 1$) then r_{XY} and r_{YZ} must also be perfectly correlated ($a = \pm 1$ and $b = \pm 1$). This provides the converse of the obvious fact, noted earlier, that if we substitute 1 for r_{XY} and r_{YZ} in inequality (3.1), we get $r_{XZ} = 1$.

5. An Application. As an application of the above observations, consider the following scenario. To encourage high school students to study, a counselor tells students that high school grades H and college grades C are positively correlated ($r_{HC} = 0.61$) and college grades and starting job salary J are also highly correlated ($r_{CJ} = 0.72$). The implication is that starting salary is positively correlated to high school grades. A sharp student sees the flaw in the counselor's pitch. Using inequality (3.1) the student reasons that

$$-0.11 \leq r_{HJ} \leq 0.99.$$

(Since the point $(.61, .72)$ is inside the unit circle in Figure 2, no useful information is obtained from the given correlations.) So, in order to determine whether H and J are actually positively correlated, a separate study must be made. On the other hand, if the counselor had slightly better correlation data, say $r_{HC} = 0.70$ and $r_{CJ} = 0.85$ then we have

$$0.22 \leq r_{HJ} \leq 0.987.$$

(Since the point $(.70, .85)$ is outside the circle in Figure 2, it follows that r_{HJ} is positive). Finally, each correlation coefficient would have to be very high to guarantee that r_{HJ} is really strong (say, at least 0.6). For instance for $r_{HC} = 0.83$ and $r_{CJ} = 0.95$ we have

$$0.61 \leq r_{HJ} \leq 0.96.$$

(Since the point $(.83, .95)$ is in the shaded region in Figure 2, $r_{HJ} \geq 0.60$.)

6. Conclusion. The single number (the correlation coefficient) does not encapsulate enough information to guarantee transitivity of the property of being positively correlated. However, if we know the correlation coefficient r_{XY} and r_{YZ} then we can find upper and lower bounds for r_{XZ}.

Moreover, both r_{XY} and r_{YZ} must be extraordinarily strong in order to guarantee that r_{XZ} is significant.

References

1. M. G. Kendall, *The Advanced Theory of Statistics,* Vol. I, 4th ed., Charles Griffen, London, 1948.

2. E. Langford, N. Schwertman, and M. Owens, "Is the Property of Being Positively Correlated Transitive?," *The American Statistician,* 55 (2001), 322–325.

3. D. Moore and G. McCabe, *Introduction to the Practice of Statistics,* 4th Ed., W. H. Freeman, New York, 2002.

4. J. Stewart, *Calculus: Early Transcendentals,* 5th ed., Brooks/Cole, Belmont, CA, 2003.

5. J. Stewart, L. Redlin, and S. Watson, *Precalculus: Mathematics for Calculus,* 5th Edition, Brooks/Cole, Belmont, CA, 2006.

Mathematics Subject Classification (2000): 62H20

Alan Safer
Department of Mathematics and Statistics
California State University, Long Beach
Long Beach, CA 90840-1001
email: asafer@csulb.edu

Saleem Watson
Department of Mathematics and Statistics
California State University, Long Beach
Long Beach, CA 90840-1001
email: saleem@csulb.edu