Linearly Independent, Orthogonal, and Uncorrelated Variables

JOSEPH LEE RODGERS, W. ALAN NICEWANDER, and LARRY TOOTHAKER*

Linearly independent, orthogonal, and *uncorrelated* are three terms used to indicate lack of relationship between variables. This short didactic article compares these three terms in both an algebraic and a geometric framework. An example is used to illustrate the differences.

KEY WORDS: Linear independence; Orthogonality; Uncorrelated variables; Centered variables; n-dimensional space.

INTRODUCTION

In the early 1970's, a series of articles and letters discussed the relationships among statistical independence, zero correlation, and mutual exclusivity (see Gibbons 1968; Hunter 1972; Pollak 1971; and Shih 1971). A related set of concepts that are equally confusing to students is firmly rooted in linear algebra: linear independence, orthogonality, and zero correlation. The series of articles cited above deal with statistical concepts in which the link between the sample and the population is critical. On the other hand, the concepts with which we are dealing apply most naturally to variables that have been fixed, either by design or in the sample. Thus, our framework is primarily algebraic and geometric rather than statistical. Unfortunately, the mathematical distinctions between our three terms are subtle enough to confuse both students and teachers of statistics. The purpose of this short didactic article is to present explicitly the differences among these three concepts and to portray the relationships among them.

AN ALGEBRAIC PORTRAYAL

Algebraically, the concepts of linearly independent, orthogonal, and uncorrelated variables can be stated as follows.

Let **X** and **Y** be vector observations of the variables X and Y. Then

1. **X** and **Y** are linearly independent iff there exists no constant a such that $a\mathbf{X} - \mathbf{Y} = 0$ (**X** and **Y** nonnull vectors).
2. **X** and **Y** are orthogonal iff $\mathbf{X}'\mathbf{Y} = 0$.
3. **X** and **Y** are uncorrelated iff $(\mathbf{X} - \bar{X}\mathbf{1})'(\mathbf{Y} - \bar{Y}\mathbf{1}) = 0$, where $\bar{X}$ and $\bar{Y}$ are the means of **X** and **Y**, respectively, and **1** is a vector of ones.

The first important distinction here is that linear independence and orthogonality are properties of the raw variables, while zero correlation is a property of the centered variables. Secondly, orthogonality is a special case of linear independence. Both of these distinctions can be elucidated by reference to a geometric framework.

A GEOMETRIC PORTRAYAL

Given two variables, the traditional geometric model that is used to portray their relationship is the scatterplot, in which the rows are plotted in the column space, each variable defines an axis, and each observation is plotted as a point in the space. Another useful, although less common, geometric model involves turning the space "inside-out", where the columns of the data matrix lie in the row space. Variables are vectors from the origin to the column points, and the n axes correspond to observations. While this (potentially) huge-dimensional space is difficult to visualize, the two variable vectors define a two-dimensional subspace that is easily portrayed. This huge-dimensional space was often used by Fisher in his statistical conceptualizations (Box 1978), and it is commonly used in geometric portrayals of multiple regression (see Draper and Smith

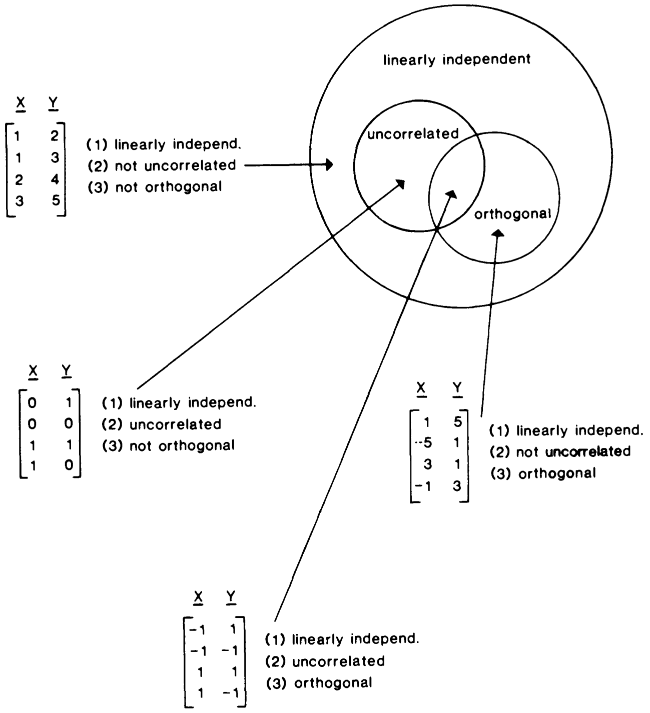

Figure 1. The relationship between linearly independent, orthogonal, and uncorrelated variables.

*Joseph Lee Rodgers is Assistant Professor and W. Alan Nicewander and Larry Toothaker are Professors in the Department of Psychology at the University of Oklahoma, Norman, OK 73019. The authors express their appreciation to two reviewers whose suggestions improved this article.

1981, pp. 201–203). This n-dimensional space and its two-dimensional subspace are the ones to which we direct attention.

Each variable is a vector lying in the observation space of n dimensions. Linearly independent variables are those with vectors that do not fall along the same line; that is, there is no multiplicative constant that will expand, contract, or reflect one vector onto the other. Orthogonal variables are a special case of linearly independent variables. Not only do their vectors not fall along the same line, but they also fall perfectly at right angles to one another (or, equivalently, the cosine of the angle between them is zero). The relationship between "linear independence" and "orthogonality" is thus straightforward and simple.

Uncorrelated variables are a bit more complex. To say variables are uncorrelated indicates nothing about the raw variables themselves. Rather, "uncorrelated" implies that once each variable is centered (i.e., the mean of each vector is subtracted from the elements of that vector), then the vectors are perpendicular. The key to appreciating this distinction is recognizing that centering each variable can and often will change the angle between the two vectors. Thus, orthogonal denotes that the *raw* variables are perpendicular. Uncorrelated denotes that the *centered* variables are perpendicular.

Each of the following situations can occur: Two variables that are perpendicular can become oblique once they are centered; these are orthogonal but not uncorrelated. Two variables not perpendicular (oblique) can become perpendicular once they are centered; these are uncorrelated but not orthogonal. And finally, two variables may be both orthogonal and uncorrelated if centering does not change the angle between their vectors. In each case, of course, the variables are linearly independent. Figure 1 gives a pictorial portrayal of the relationships among these three terms. Examples of sets of variables that correspond to each possible situation are shown.

[*Received April 1983. Revised October 1983.*]

REFERENCES

BOX, J.F. (1978), *R.A. Fisher: The Life of a Scientist*, New York: John Wiley.

DRAPER, N., and SMITH, H. (1981), *Applied Regression Analysis*, New York: John Wiley.

GIBBONS, J.D. (1968), "Mutually Exclusive Events, Independence, and Zero Correlation," *The American Statistician*, 22, 31–32.

HUNTER, J.J. (1972), "Independence, Conditional Expectation, and Zero Covariance," *The American Statistician*, 26, 22–24.

POLLAK, E. (1971), "A Comment on Zero Correlation and Independence," *The American Statistician*, 25, 53.

SHIH, W. (1971), "More on Zero Correlation and Independence," *The American Statistician*, 25, 62.

Kruskal's Proof of the Joint Distribution of $\bar{X}$ and s^2

STEPHEN M. STIGLER*

In introductory courses in mathematical statistics, the proof that the sample mean $\bar{X}$ and sample variance s^2 are independent when one is sampling from normal populations is commonly deferred until substantial mathematical machinery has been developed. The proof may be based on Helmert's transformation (Brownlee 1965, p. 271; Rao 1973, p. 182), or it may use properties of moment-generating functions (Hogg and Craig 1970, p. 163; Shuster 1973). The purpose of this note is to observe that a simple proof, essentially due to Kruskal (1946), can be given early in a statistics course; the proof requires no matrix algebra, moment-generating functions, or characteristic functions. All that is needed are two minimal facts about the bivariate normal distribution: Two linear combinations of a pair of independent normally distributed random variables are themselves bivariate normal, and hence if they are uncorrelated, they are independent.

Let $X_1, \ldots, X_n$ be independent, identically distributed $N(\mu, \sigma^2)$. Let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\quad s_n^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

We suppose that the chi-squared distribution $\chi^2(k)$ has been defined as the distribution of $U_1^2 + \cdots + U_k^2$, where the U_i are independent $N(0,1)$.

Theorem. (a) $\bar{X}_n$ has a $N(\mu, \sigma^2/n)$ distribution. (b) $(n-1)s_n^2/\sigma^2$ has a $\chi^2(n-1)$ distribution. (c) $\bar{X}_n$ and s_n^2 are independent.

Proof. The proof is by induction. First consider the case $n = 2$. Here $\bar{X}_2 = (X_1 + X_2)/2$ and, after a little algebra, $s_2^2 = (X_1 - X_2)^2/2$. Part (a) is an immediate consequence of the assumed knowledge of normal distributions, and since $(X_1 - X_2)/\sqrt{2}$ is $N(0,1)$, (b) follows too, from the definition of $\chi^2(1)$. Finally, since $\operatorname{cov}(X_1 - X_2, X_1 + X_2) = 0$, $X_1 - X_2$ and $X_1 + X_2$ are independent and (c) follows.

Now assume the conclusion holds for a sample of size n. We prove it holds for a sample of size $n + 1$. First establish the two relationships

*Stephen M. Stigler is Professor, Department of Statistics, University of Chicago, Chicago, IL 60637.