RECENT ADVANCES IN MATHEMATICAL STATISTICS, II[1]

By Cecil C. Craig

University of Michigan

The statistical theory of the linear relationship between a dependent variable x_1, and a set of independent variables $x_2, x_3, \cdots, x_{t+1}$, is by now quite generally understood. Supposing that the x_i's are measured from their respective means, we determine the coefficients, $b_2, b_3, \cdots, b_{t+1}$, in such a way as to maximize the coefficient of correlation $r_{1.2\,3\ldots t+1}$ between x_1 and $\sum_{i=2} b_i x_i$. This coefficient of correlation, usually called the multiple correlation coefficient, measures the exactness of the linear relationship that exists, and it has the property of being quite unchanged if the origins or the scales for the separate x_i's are changed in any way or even if the set $x_2, x_3, \cdots, x_{t+1}$ should be replaced by any equivalent set of linear combinations of them. That is, e.g., if $t = 3$, the new variables, $v_2 = x_2 + x_3 + x_4$, $v_3 = 2x_1 - x_3 + 3x_4$, $v_4 = x_1 + 2x_3 - 2x_4$ are equivalent to x_2, x_3, x_4, since the latter can be found if the v_i's are known, and the multiple correlation between x_1 and the v_i's is exactly the same as that between x_1 and x_2, x_3, x_4. Moreover, the requisite sampling theory if the variables involved are normally distributed is well established.

I want to discuss briefly an important generalization of this kind of situation that has been the subject of recent research. In particular, in his paper, "Relations between two sets of variables," published in *Biometrika* in 1936 [1] H. Hotelling set forth these ideas in excellent fashion and contributed much to the mathematical theory required for their practical application. We now suppose that we have two sets of measurements, $x_1, \cdots, x_s$, and $x_{s+1}, \cdots, x_{s+t}$, made on the same object and that we are interested in the linear relations that may exist between the members of one set and the members of the other. As an example, $x_1, \cdots, x_s$ might be the prices of s more or less related commodities at a given time, and $x_{s+1}, \cdots, x_{s+t}$ measures of factors which may be thought to be effective in the price situation.

In the more special case I began with, $s = 1$, and a single equation fully expressed the linear statistical relationship of x_1 with $x_2, \cdots, x_{t+1}$. Now there are s dependent variables and now with $s \leqq t$, not one but s distinct linear relations will exist and will be required to fully describe the linear connections between the two sets of variables. We may assume that there is no mere duplication among the variables we are using, i.e., no one of the s x_i's is always exactly given by a linear combination of the others in the set and the same is

[1] This is the second of two papers read by B. H. Camp and the author on "Recent Advances in Mathematical Statistics" before the American Statistical Association, the Econometric Society, and the Institute of Mathematical Statistics, on December 30, 1941, in New York City. The authors selected topics from papers published during the past five years.