CORRELATION AND CAUSATION

By SEWALL WRIGHT

Senior Animal Husbandman in Animal Genetics, Bureau of Animal Industry, United States Department of Agriculture

PART I. METHOD OF PATH COEFFICIENTS

INTRODUCTION

The ideal method of science is the study of the direct influence of one condition on another in experiments in which all other possible causes of variation are eliminated. Unfortunately, causes of variation often seem to be beyond control. In the biological sciences, especially, one often has to deal with a group of characteristics or conditions which are correlated because of a complex of interacting, uncontrollable, and often obscure causes. The degree of correlation between two variables can be calculated by well-known methods, but when it is found it gives merely the resultant of all connecting paths of influence.

The present paper is an attempt to present a method of measuring the direct influence along each separate path in such a system and thus of finding the degree to which variation of a given effect is determined by each particular cause. The method depends on the combination of knowledge of the degrees of correlation among the variables in a system with such knowledge as may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them.

CORRELATION

Relations between variables which can be measured quantitatively are usually expressed in terms of Galton's (*4*)[1] coefficient of correlation, $r_{XY}=\frac{\Sigma X'Y'}{n\sigma_X\sigma_Y}$ (the ratio of the average product of deviations of X and Y to the product of their standard deviations), or of Pearson's (*7*) correlation ratio, $\eta_{X\cdot Y}=\frac{\sigma\left({}_{Y}{}^{M}{}_{X}\right)}{\sigma_X}$ (the ratio of the standard deviation of the mean values of X for each value of Y to the total standard deviation of X), the standard deviation being the square root of the mean square deviation.

Use of the coefficient of correlation (r) assumes that there is a linear relation between the two variables—that is, that a given change in one variable always involves a certain constant change in the corresponding average value of the other. The value of the coefficient can never exceed

[1] Reference is made by number (italic) to "Literature cited," p. 585.

Journal of Agricultural Research, Washington, D. C.
wh

Vol. XX, No. 7
Jan. 3, 1921
Key No. A-55

17777°—21——4