# STATISTICAL METHODS IN BIOLOGY

By Sewall Wright



The applications of statistical methods in biology are essentially identical in purpose with those in other fields of science. There are, however, some differences in form and emphasis, imposed by the kind of material. In all of the characteristics of the individuals of the million or more species of animals and plants there is variability, not the errors of observation of the physicist, but real variability, of interest on its own account. An enormous field is presented here for statistical methods in merely bringing the phenomena of life into an adequate descriptive form which the mind can grasp.

It was necessary for the pioneer biometricians to readapt the methods developed for use in the physical sciences, to their kind of material. The classical normal probability curve, applicable enough as a rule to the treatment of random errors, was wholly inadequate for the description of biological variability. Pearson's system of frequency curves as one solution of this difficulty is familiar. Similarly, the methods of simple and multiple correlation developed by Galton and Pearson met a descriptive need in biology, not encountered in the physical sciences. Mathematically these methods were simply adaptations of the method of least squares, but there was a significant change in viewpoint.

The second type of application to which I shall refer is that of the determination of the significance of differences, whether between statistics of different natural populations or between results of experiments. Here a more direct borrowing of the methods of physical science was possible. But in addition to use of the classical probable error, we have Pearson's $\chi^2$ method for comparing systems of frequencies and the methods of "Student" and R. A. Fisher for dealing accurately with probabilities in the small number of paired observations characteristic of biological experiment. Genetics has been dependent on statistical methods from the first. More recently, physiologists, anatomists, ecologists and others are coming to realize their importance.

A third application is analogous to the physicist's use of the theory of probability in the kinetic theory of gases and in the more recent developments of statistical mechanics. Mendel's interpretation of his experiments in heredity was a simple example of this sort. The interpretation of the properties of populations, including the theory of evolution, as statistical consequences of the genetics of individuals is perhaps the most important example.

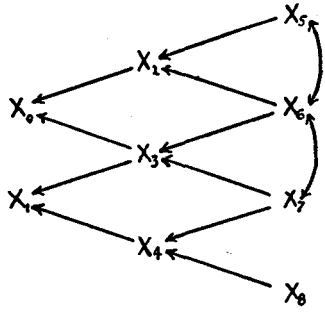
I shall devote most of my time to an application which mediates between two of the above fields, viz., the field of interpretation of statistical descriptions. Biology differs from physics and chemistry in dealing with real variability and thus in having a problem of statistical description. In this it resembles the social sciences, but differs from those and approaches physics and chemistry in the degree to which laboratory experiments can be conducted. The two modes of approach, statistical and experimental, should supplement each other in giving insight into natural phenomena. Actually they are apt to be conducted from such different philosophical viewpoints that they lead to seemingly antagonistic interpretations. We have such unhappy situations as the existence of two sciences of heredity, Mendelian and biometric, scarcely on speaking terms. In biology, at least, we need a technique for interpreting the statistical relations of systems of variables in terms of our knowledge of causal relations, derived in the laboratory.

In connection with such interpretation, there is a certain contrast between the kind of interrelation of variable quantities which the physicist encounters and that frequently encountered by the biologist. The variables of the former are usually in a movable equilibrium, dependent on reversible processes. One speaks of the functional relations of the components of such a system, rather than of causation. The tendency of physics, emphasized recently by G. N. Lewis, is to insist on the complete symmetry of its time.

The biologist, on the other hand, is to a large extent concerned with variables which at his level of observation are related in irreversible sequence. He deals with the development of individuals from egg to adult, and with the evolution of species. His hereditary units affect the characteristics of individuals which possess them but are not themselves affected. Most of the environmental factors with which he is concerned, act upon organisms without being acted upon to an important extent. Thus the conception of one-way causation is a useful one at the biological level and any treatment of systems of biological variables which disregards sequence (where present) omits the very part in which the biologist is most interested. Our technique of interpretation of statistical systems must then take account of sequential relations as well as of symmetrical relations.

A qualitative interpretation of a system of variables (or if one pleases, a mere arbitrary point of view) is conveniently represented by a diagram in which arrows are used to indicate which variables are to be treated as functions of which others. Such a diagram is especially adapted to representation of one-way causation, but is not limited to such relations. Unanalyzed correlations may be represented by two-

CHART I




$$r_{04} = p_{02}r_{24} + p_{03}r_{34}$$
$$= p_{47}r_{07}$$
$$= p_{01674} + p_{03674} + p_{0374}$$
$$= p_{02}p_{26}r_{67}p_{47} + p_{03}p_{36}r_{67}p_{47} + p_{03}p_{37}p_{47}$$

headed arrows, to indicate connection through common factors (Chart I).

It is convenient to measure each variable in terms of its standard deviation. Letting $x_0 = \frac{X_0 - \overline{X}_0}{\sigma_0}$, etc., we can write the best linear expression for deviations of a given variable in terms of those from which arrows are drawn to it in the form:

$$x_0 = p_{02}x_2 + p_{03}x_3.$$

The coefficients are abstract numbers, which I have called path coefficients, related numerically to the concrete partial regression coefficients in the same way that the correlation coefficient is related to total regression. They differ from correlation coefficients, however, in having direction. Their usefulness depends on an easily demonstrated relation to correlation. For any two variables of such a system, the correlation can be analyzed into contributions tracing through the represented factors of either one. Letting $s$ stand for the factors of $X_0$ and $t$ for those of $X_1$,

$$r_{01} = \Sigma p_{0s}r_{1s} = \Sigma p_{1t}r_{0t}.$$

By further analysis of the correlation terms, this leads to the easily remembered principle that any correlation can be analyzed into contributions from all of the paths through the diagram (direct or through common factors) by which the two variables are connected, and that each of these contributions is the product of the coefficients pertaining to the elementary paths. One of these elementary paths in each case may be an unanalyzed bidirectional one, measured by a correlation coefficient.

As a special case, the correlation of a variable with itself (unity) may be analyzed in this way, assuming that there is complete determination by the factors represented.

$$\Sigma p_{0s} r_{0s} = 1.$$

(If determination is not complete the sum of such products gives the squared multiple correlation.)

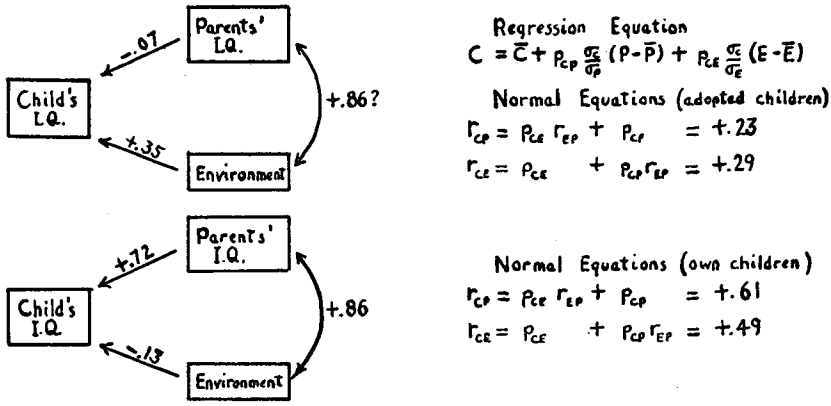
In a system which one wishes to analyze, some of the variables represented may be ones which have been measured, others may be hypothetical. One may deduce unknown correlation coefficients from path coefficients, given by knowledge of the functional relations, or unknown path coefficients from known correlations, or unknowns of both sorts from a mixture of knowns of both sorts. The application is, of course, often limited by inability to make a qualitative interpretation sufficiently definite to be expressed in diagramatic form, and even when such a representation can be made, it is only too likely to turn out that there are more unknowns than knowns, thus giving an indeterminate solution. No quantitative interpretation is then warranted until new facts, suggested perhaps by the attempt at formulation of the problem, have been obtained.

As a geneticist, I have been especially interested in applications in the field of heredity. Let me give as an example a case dealing with heredity in man, as perhaps of more general interest than those dealing with such animals as guinea pigs. I am taking some data presented by Miss D. S. Burks, involving intelligence tests of some 100 California children, tests of their parents, and in addition carefully constructed grades of their home environments. These data were obtained by Miss Burks as a control for similar data for some 200 children, adopted at an average age of three months. The two groups of parents were closely similar. I should say that Miss Burks is not responsible for the use to which I am putting her data.

The observed correlations as corrected by Miss Burks for attenuation are given in the equations, Chart II. Midparents are used for simplicity. The correlation between midparent and home environment (culture index) was not calculated for the foster data. Presumably it was closely similar to the figure for the control data.

The data suggest certain things rather definitely but in other respects interpretation is not obvious. The routine method of treatment is to calculate partial correlation coefficients or the closely allied partial regression coefficients, treating child's IQ as a function of parental IQ and environment. The results are shown in the figures. The solution gives a rather curious result. Environment makes a significant positive contribution in the foster data, but in the control data its contribution is negative, as far as it goes. The partial correlation coefficients differ similarly. How are we to interpret this change from +.35 to

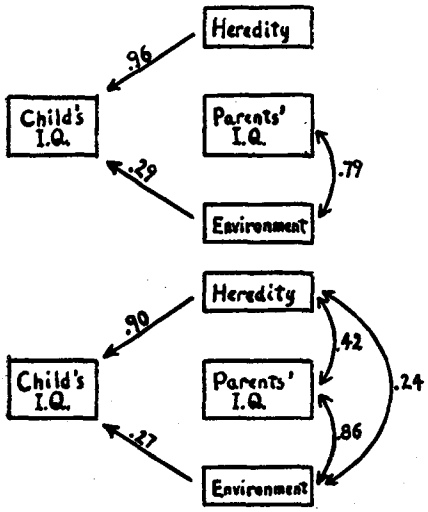
CHART II




−.13 in the environmental regression coefficients? The answer is, of course, that we have no right to put any biological interpretation on them. We have prediction equations, the best which the data yield, but not an interpretation. For interpretation we must take account of the causal relations. The IQ of child, and of parent, and the grade of home environment are not functionally related after the simple fashion of volume, pressure and temperature of the gas law. We know that the characteristics of the child trace to two distinct groups of biological factors, the constant internal factor, heredity, present in the chromatin material of the child's cells, and the external factors to which this heredity and its products have reacted in the developmental process. There is one-way causation by heredity and doubtless the same to at least a first approximation by environment.

The IQ of parents is related to home environment both directly and indirectly. For the moment, it will suffice to indicate this by a double-headed arrow. Heredity of foster children should be independent both of parental IQ and environment since Miss Burks shows that there was no possibility of selective adoption with respect to intelligence. In the control data, parental IQ should be correlated with child's heredity in various ways, all indirect (as being through parental heredity), and the diagramatic representation must be by a two-headed arrow. It is important to recognize that parental IQ is very far from being the child's heredity. Finally, in the case of human intelligence, it is to be expected that there will be some correlation between child's heredity and environment. Good heredity in preceding generations should have built up the conditions for a favorable environment. The simplest interpretations which can possibly be considered adequate biologically are those of Chart III.

CHART III




Adopted Children

$r_{CE} = p_{CE} = +.29$

$r_{CP} = p_{CE} r_{EP} = +.23$

$p^2_{CE} + p^2_{CH} = 1.00$

Own Children

$r_{EP} = +.86$

$r_{CE} = p_{CE} + p_{CH} r_{HE} = +.49$

$r_{CP} = p_{CE} r_{EP} + p_{CH} r_{HP} = +.61$

$p_{CE} = \frac{.29}{.96} p_{CH}$

$p^2_{CE} + p^2_{CH} + 2 p_{CE} p_{CH} r_{HE} = 1.00$

The analysis of the foster data is very simple. If IQ of the foster parents is related to child's IQ only through correlation with home environment, the parent-offspring correlation should be the product of the two intermediary coefficients. This leads to a value of the correlation between midparent and environment (+.79) closely similar to that observed in the control data. This indicates that there was no influence of the parents other than through the home environment as actually measured. There was only about 9 per cent determination of variance by home environment ($.29^2$) leaving a residuum of 91 per cent determination and a residual path coefficient of about .96. How far this traces only to child's heredity and how far to unmeasured environmental factors the data give no answer. But since home environment is presumably much the most important environmental factor (cases in which there had been illness likely to affect mentality having been excluded), one may surmise that the residual group is largely heredity.

In the other group, the situation is more complex. We can at once write three equations representing analysis of the three known correlations. If we assume that the only factor of child's IQ apart from the home environment as measured is heredity, we can write a fourth equation expressing complete determination. But there are five coefficients to be determined. No solution is possible and no quantitative interpretation is possible from the data of the control group. This is not a fault of the method. It is rather a merit that it brings clearly to light the impossibility of any biological interpretation without further data.

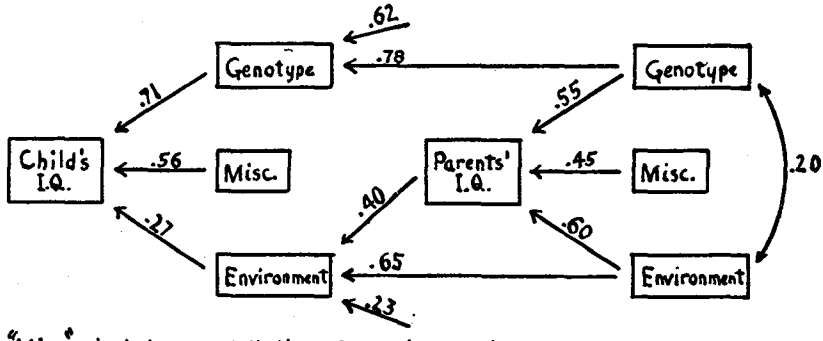
In the present case, however, we have another resource. The control group of parents was carefully selected for comparability with the foster group. Presumably home environment has closely similar effects in the two cases. We should be able to borrow the environmental coefficient from the foster data. Theoretically, however, it is the concrete partial regression coefficient and not the path coefficient which is directly transferable, the latter being affected by the correlation between heredity and environment in the control data. From this it may be deduced that the ratio $p_{CE}:p_{CH}$ should be the same in the two cases, giving a fifth equation.

These five equations differ from the normal equations of the ordinary application to multiple regression in not all being linear. They are easily solved, however, with the results indicated in Chart III. It may be well to emphasize the point that the fact that it was possible to use the relation between environment and child's IQ of the foster data in the control data and obtain results conforming to the observed correlations in the latter shows that the apparent contradiction in the partial correlations in the two cases was an illusion.

One may be struck by the low correlation (+.42) between midparental IQ and child's heredity in contrast with the high correlation (+.86) observed between the former and home environment.

It appears that midparental IQ is a much better index of home environment than of child's heredity. This is not surprising, however, in the light of genetic theory. Even the correlation between midparental heredity and child's heredity is theoretically only $\frac{2}{3}\sqrt{\frac{1}{2}}$ or .47 under certain common conditions (complete dominance present and no

CHART IV




"Misc." includes contributions to variance due to

1. Nonadditive combination effects of genes (dominance, epistacy)
2. Nonadditive combination effects of heredity and environment
3. Residual environmental factors (uncorrelated with parental I.Q.)

assortative mating). It may be worth while to carry the analysis a generation back, still averaging the parents for simplicity.

For this analysis, it is convenient to deal with heredity in a different way. In place of heredity as a factor in development, we shall use the genotype as the sum of such gene contributions as best approximate the developmental ranking. The two measures are identical only if dominance is wholly lacking and there are no epistatic effects (i.e., if the effects of independent series of genes combine additively). This will give us a minimum instead of a maximum estimate of the genetic factor, compatible with acceptance of the observed correlations. We must now recognize a residual factor in both generations, theoretically composed of three very diverse elements, which, however, cannot be distinguished in the present data. These are the usually important contributions of dominance and epistacy to variance, just referred to, which are purely hereditary factors from the developmental viewpoint; second, environmental factors not included in the measure of home environment, and as indicated by the foster data, not correlated with the parents; and third, possible contributions to variance due to non-additive effects of heredity and environment in relation to each other.

Home environment is treated as in part created by the IQ of the parents (direct path) and in part as tracing from the previous environment of the parents, as would be true of the effects of inherited wealth and family tradition. The possibility of some independent determination is indicated by a third arrow.

Child's genotype traces, of course, to midparental genotype but is not completely determined thereby because of the intervening phenomenon of Mendelian segregation. If there were no assortative mating the correlation (and also path coefficient) is $\sqrt{\frac{1}{2}}$ or .707. In the present data there was strong assortative mating, .55, to be raised to about .70 on correcting for attenuation. This raises the value of the above path coefficient to a slightly varying extent, depending on the nature of the assortative mating. The value, .78, can be accepted as reliable within a smaller range than any of the observed correlations.

The diagram has twelve paths. To make it quantitative at least twelve equations must be found for solution. We have just derived one from Mendelian theory, and three others are given by the three observed correlations. The four residual paths correspond to four equations of complete determination. The environmental effect on child's IQ, borrowed from the foster data, brings the number up to nine. The list could easily be completed, if we could assume that the situation back of the parents was the same as back of the children.

This is not a necessary assumption, however, if only because the parents were tested as adults, the children between 5 and 14 years of age. It turns out that it is a mathematically impossible assumption. To complete the series it was assumed that the ratio of the coefficients in the case of the residual group and genotype is the same in the two generations (expected if the residual group is largely due to dominance and epistacy); second, the environmental influence and the correlation between genotype and environment were adjusted to agree as well as possible; and third, a small arbitrary value was assigned to the residual path back of environment (a value practically immaterial if small). The solution is not strictly determinate, but is so within rather narrow limits. The coefficients are accordingly given only to the nearest .05 (Chart IV).

We have as a result a somewhat roughly quantitative interpretation of the relations of the variables in this population, partly based on observed correlations and partly based on our knowledge of the mechanism of heredity. It illustrates sufficiently, I hope, the difference between a biological interpretation of statistical data and a prediction formula based on the same data.