SCIENCE

VOL. LXIX JANUARY 18, 1929 No. 1777

CONTENTS

SCIENCE: A Weekly Journal devoted to the Advancement of Science, edited by J. McKeen Cattell and published every Friday by

THE SCIENCE PRESS

New York City: Grand Central Terminal.

Lancaster, Pa. Garrison, N. Y.

Annual Subscription, $6.00. Single Copies, 15 Cts.

SCIENCE is the official organ of the American Association for the Advancement of Science. Information regarding membership in the Association may be secured from the office of the permanent secretary, in the Smithsonian Institution Building, Washington, D. C.

THE AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

THE RELATION OF STATISTICS TO MODERN MATHEMATICAL RESEARCH[1]

THE annual meeting of Section A, in conjunction with the other mathematical organizations, and in affiliation with the bodies devoted to progress in all departments of science, gives us occasion as mathematicians to pause and look about us, to be stirred by the proximity of the major activities of the rest of the scientific world, and to form a renewed estimate of our place in the general scheme of things. During most of the year we are occupied with our own problems, or in consultation with a colleague in a neighboring field about a particular aspect of some special investigation, the further implications of which he has time to explain only incompletely, and we imperfectly to understand. Occasionally it is well to reassemble the detached items which have engaged our attention from time to time, and to judge the broad tendencies of which they are significant.

We are constantly reminded that our affiliations are no longer so exclusively with physics and astronomy as was formerly the case. Not only chemistry, but also the biological and social sciences are taking us into their confidence, and welcoming us to the discussion of their projects and their difficulties. A former pupil of mine appears on the records of the graduate school as a candidate for the degree of Ph.D. with psychology as major subject, and a minor divided between mathematics and botany. One can imagine the derision with which such a combination would have been greeted a few years ago. His thesis is a statistical investigation of a psychological problem, in support of which he has taken courses in the mathematical theory of statistics and in biometry. The fact that derision among workers in different fields has been so largely replaced by respectful attention is one of the most wholesome features of the situation.

The diversion of a part of our energies into new channels does not mean that the problems of mathematical physics are any less challenging than before. Within our own recent memory the doctrine of relativity has constituted a major upheaval in the realm of pure mathematics as well as in physics. There has

[1] Address of the vice-president and chairman of Section A—Mathematics, American Association for the Advancement of Science, December, 1928.

perhaps never been a time when current physical discussion was more prevailingly mathematical than in its present analysis of atomic structure. But changes in the mathematical basis of this discussion are so rapid that only those in the front line of investigation can hope to keep pace with them. Some one has said that anything printed on the subject of the quantum theory is *ipso facto* out of date. How much more is it true that by the time the ordinary mathematician has mastered any particular phase of the theory with anything like what he regards as finality, that phase has become a doubtful contender even for historical interest. In turning our attention to the newer applications of mathematical method, we are both responding to the needs of the sciences concerned and finding for ourselves a welcome outlet for some of the products that we are more readily able to supply.

It is not to be expected that we shall soon have the same power of prediction in the new fields that we have in the exact sciences, traditionally so called. The erection of a logical structure to any considerable elevation requires a firm foundation in the shape of an adequately comprehensive system of postulates, and vital phenomena are too complex to be accounted for by any manageable formulation. There is a constant danger of over-refinement bringing conclusions into discredit. A speaker before a local meeting a good many hundred miles from New York had gathered comparative statistics on the price of hogs and the number of students electing Latin in a certain group of colleges, or something of the sort, and reported "a coefficient of correlation of .3999, or very nearly four tenths." Just what important social consequences would have ensued if the correlation really had been as high as four tenths, instead of a little less, he omitted to say. The endeavor to describe human affairs in mathematical terms is an old one. Many essays in that direction have come to grief by attempting from the outset to accomplish the impossible. The present revolutionary extension of the domain of the applications of mathematics is due partly to the development of an appropriate technique, partly to social conditions which have made practicable the assembling of adequate statistical data, and partly to the acceptance of a working balance between aspiration and the possibility of achievement.

A necessary condition for the development of the classical dynamics was not only the creation of the mathematical tools, but also the isolation of the physical concepts which are susceptible of satisfactory mathematical treatment, and the recognition of the quantities which are fundamental in the controlling equations, such as acceleration, energy and the like. This was a long and temporarily uncertain process. The boy in your mechanics class who tells you that the velocity of a projectile in flight is one of the forces acting on it, is only repeating the gropings which the pioneers in the science must have gone through. Having been impressed at an uncritical age with the principle of the conservation of energy, I awoke to a painful realization that I did not know what energy meant, and the more I tried to find out, the more I was puzzled by the alleged importance of the conservation of so remote an abstraction, until it dawned on me in a flash of illumination that the importance of energy lay in the fact that it was the thing that was conserved.[2] The discovery of the existence of such a unifying permanence by the creators of the science was one of the things that made mathematical physics possible.

It is not apparent that there exists any small group of variables sufficient to characterize the phenomena of a living organism. A physical body is more or less homogeneous, or can be so considered for purposes of argument. With an organism this is not the case. Statistical prediction deals with organisms in the mass, while our vital human concerns are with organisms as individuals; and the individual has a complexity of parts, which demand separate consideration with regard to their several functions. There is no prospect, humanly speaking, that we shall ever have methods for the practical handling of complicated differential equations in several dozen or several million variables. In the controversy between mechanist and vitalist, between behaviorist and subjectivist, the part of the mathematician must occasionally be the inglorious one of counseling moderation. He is too familiar with the limitations of calculation to see in them, or in the present limitations of any science, need for invoking supernatural intervention. But when he considers that an exhaustive discussion of the dynamics of a single atom of helium transcends our powers for the time being, and lifts his eyes from an atom of helium successively to an atom of carbon, a molecule of sugar, a molecule of protoplasmic substance, a living cell, a toadstool, a tree, a guinea pig and a professor of mathematics, he does not look for any early reduction of the laws of life to a system of differential equations which he can integrate.

Nevertheless, the extension of the scope of quantitative methods through the medium of statistical analysis is one of the most significant things going on in the scientific world at the present time. It is noteworthy that it has gone so far with comparatively little attention from mathematicians, and that statisti-

[2] It may be noted that the substance of this passage was written before the appearance of Professor Swann's article in SCIENCE for November 2, 1928; *cf.* especially page 418.

cians have independently reconstructed so much that a simple adaptation of existing mathematics would have given them outright. Even now there is a prevalent idea that statistics, if properly regarded as a branch of mathematics at all, is a remote and highly specialized one. It is my purpose here to show how some of the fundamental notions of statistics fit directly into the frame of modern mathematical research. In doing this I shall not have anything to offer that is essentially new. I shall only be repeating things that are already well known in substance, in a form and with an emphasis appropriate to the occasion.

Professor E. B. Van Vleck (in volume 39 of SCIENCE) has called attention to the dominant influence of the study of Fourier series on the progress of mathematics during the past hundred years, as far as it relates to the theory of functions of real variables. The theory of the Fourier series itself is still a living issue, and within the memory of most of us has produced in the work of Fejér one of the most beautifully simple and satisfying theorems of modern analysis. Investigation of other analogous series of mathematical physics, such as those of Legendre and Bessel, is actively in progress at a less advanced stage. The broader generalizations relating to the series associated with differential systems and integral equations have occupied a large part of the attention of European and American mathematicians in our time. And the still wider horizon of the general function concept, which we owe in large measure to these researches, extends itself from year to year.

It is interesting, as an aside, to recall a passage from the early pages of Les Misérables, describing the state of French civilization in the year 1817: "Il y avait a l'académie des sciences un Fourier célèbre que la postérité a oublié et dans je ne sais quel grenier un Fourier obscur dont l'avenir se souviendra." One wonders whether the author, if he could return to the scene, would be mildly surprised by the verdict of subsequent history, or would dismiss the actual facts as irrelevant and serenely reiterate his original assertion.[3]

Whatever scale of values the novelist might set up for mankind generally, a systematic account of the properties of Fourier series and their successive implications can be made to embrace a very large domain of our technical concerns as mathematicians. And in this domain, by intimate association and abstract identity, are contained the leading ideas of mathematical statistics.

Nothing in the theory of Fourier series is more striking than the least-square property. The partial sum of the series for a given function $f(x)$, through terms of the nth order, is a linear combination of the functions $1, \cos x, \cdots, \cos nx, \sin x, \cdots, \sin nx$. It is characterized among all such linear combinations as the one which gives the best approximation to $f(x)$, in the peculiar sense that the integral of the square of the error, extended over a period, is a minimum. This relation to the classical method of least squares for the adjustment of observations, appearing at first perhaps as a mere superficial analogy, gradually assumes a fundamental significance.

It is speedily recognized that the least-square property has a wide range of generality. Its proof in the case of the Fourier series depends clearly and directly on the orthogonality of the sines and cosines that make up the series, the fact that the integral of the product of any two of them over a period is zero, the same fact which leads to the familiar determination of the coefficients. It can therefore be extended immediately to other series in which the coefficients are similarly determined. Not only the Legendre series, but also the series of orthogonal functions obtained from linear differential equations with boundary conditions or from integral equations, in fact all series of orthogonal functions generally, have the property that their partial sums give a closest approximation in the sense of the method of least squares.

The part played by the relations of orthogonality is so conspicuous that it tends to mask a certain further significance of the results. Renewed consideration shows that these relations are a convenience in calculation, rather than a condition essential to the treatment of the problem. An arbitrary linear combination of the first $n+1$ Legendre polynomials, for example, is nothing more nor less than an arbitrary polynomial of the nth degree. The partial sum of the Legendre series for a given $f(x)$ is merely the polynomial of corresponding degree which gives the closest possible approximation to $f(x)$ according to the least-square criterion. It can be defined and studied as such, without any reference to the special properties of the Legendre polynomials at all. A similar procedure is possible in the case of approximation by linear combinations of any given linearly independent functions. The existence of the Legendre polynomials, fulfilling the condition of orthogonality, does not depend on any peculiarity of the functions $1, x, x^2, \cdots$, of which they are built up. If any set of linearly independent functions is given, it is possible to construct a sequence of linear combinations of them which shall be orthogonal to each other. The construction of such a set of orthogonal functions, familiar now in connection with the theory of integral equations, is commonly spoken of as "Schmidt's process," with reference to the thesis of Erhard

[3] It may be worth recording in this connection that the birthplace of the sociologist Fourier and that of Victor Hugo are a few doors apart in the same street of the same French provincial town.

Schmidt, a classic among original memoirs in mathematical research for clearness and elegance of presentation, though Schmidt himself ascribes the essence of the construction to Gram, and one suspects that Gauss would have smiled complacently at a suggestion that the method in Gram's hands was radically novel. The actual carrying out of the process, found thus to be always possible, is for many purposes unnecessary.

It is at this point that one of the far-reaching generalizations of modern analysis becomes pertinent. Reference has already been made to the emergence of the function concept from the study of the types of series that have been under discussion. In the "functional analysis" or "general analysis" of recent decades, crystallizing ideas which had long been current as exemplified in particular cases, not only the manner of dependence of one variable on another is arbitrary, but also the range of values taken on by the independent variable. Instead of ranging over an interval, it may in particular be restricted to a finite number of values. A set of numbers $x_1, x_2, \cdots, x_n$ may be regarded as a function x_k of the index k, which takes on the values $1, 2, \cdots, n$. Many of the relations which are fundamental in the theory of approximation carry over to this type of functional dependence without essential change. The problem of a least-square approximation of a given "function" by a linear combination of other given "functions" becomes the problem of determining a regression formula; and the mathematical processes of statistics begin to take their place in the general outline. The practical cases are excessively simple, to be sure, from the point of view of the general theory, the number of "functions" in terms of which the approximation is to be expressed being usually one, less frequently two, and always small, but even so the characteristic features of the problem are retained to such an extent that the identification is a genuinely significant one. Incidentally, the equations for determining the coefficients, when formulated in general algebraic terms, become the normal equations of the classical method for the approximate solution of a system of linear equations by least squares. Working out the theory of the regression formulas in systematic fashion, one obtains the solution of the older problem at the same time, without extra charge. And while the use of the calculus to obtain a proof of an elementary theorem in algebra may seem *a priori* like a roundabout procedure, it is actually true that for one who is already familiar with the calculus the formulas of integration are often simpler, clearer and more suggestive than those of summation, whether because of a more convenient notation or for a deeper reason, and a proof is often found most easily by thinking first in terms of integrals, and then translating into algebraic notation.

Familiarity with current mathematical notions presently throws light on the formulas of statistics from another angle. For the purpose of the method of least squares, the measure of the discrepancy between a set of numbers $x_1, x_2, \cdots, x_n$ and another set $y_1, y_2, \cdots, y_n$ is the quantity

$$(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2.$$

It is a matter of course then to regard $(x_1, x_2, \cdots, x_n)$ and $(y_1, y_2, \cdots, y_n)$ as the coordinates of two points in n-dimensional space, and $\Sigma(y_k - x_k)^2$ as the square of the distance between them. This does not mean that the mathematician pretends to be physically cognizant of a space in which he can put his right shoe on his left foot, except by an exercise of professorial absent-mindedness. It means only that facts in the geometry of two or three dimensions often correspond to algebraic relations for which the restriction to two or three variables is irrelevant. While it is possible to build up a system of n-dimensional geometry on the basis of postulates, in Euclidean fashion, and while there are mathematicians who have trained themselves to think in terms of such postulates, the language of the geometry of higher dimensions is for most of us just a means of stimulating the imagination to a clearer insight into the structure of algebraic formulas. The notion of distance having received its algebraic expression, the next step is to define the measure of an angle, and this also is effected without difficulty. If the points $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y_n)$ are denoted by P and Q, and if the origin is O, the cosine of the angle POQ is

$$\frac{\Sigma x_k y_k}{\sqrt{(\Sigma x_k^2)(\Sigma y_k^2)}}.$$

But this is just the formula for the coefficient of correlation between the variables (x) *and* (y). The term coefficient of correlation, to be sure, is applicable only if the x's and the y's are measured as deviations from their respective means, so that $\Sigma x_k = \Sigma y_k = 0$; but the algebraic identities underlying the theory are wholly independent of this restriction. If there are three variables $(x_1, \cdots, x_n)$, $(y_1, \cdots, y_n)$, and $(z_1, \cdots, z_n)$, represented by three points P, Q, R, the various coefficients of correlation measure the trihedral angle O-PQR, or the corresponding spherical triangle. The coefficient of partial correlation between (x) and (y), when (z) is "held fast," is the cosine of the dihedral angle between the planes POR and QOR, and the coefficient of double correlation between the variable (x) and the pair of variables (y), (z) is the cosine of the angle between the line OP and the plane QOR.

If a critic demands logic to supplement analogy, there is no difficulty in putting the correspondences in such form that geometrical reasoning in the three-dimensional space of experience applies with cogent force. When more than three statistical variables are involved, the geometrical interpretation demands a plurality of dimensions in earnest, but even then geometrical analogy is genuinely helpful. While the fact that coefficients of correlation can be regarded as cosines has been recognized for some time, stress has usually been laid on a comparatively elaborate setting up of the correspondence for the special case of normal correlation, and the fundamental directness and simplicity of the more general interpretation has been largely disregarded.

Some one may have been disposed to object from the beginning that too much emphasis has been placed on the criterion of least squares as a measure of approximation. It may be argued, for example, that the method of moments is often preferable to that of least squares. A preliminary answer is that for the straightforward fitting of a polynomial curve the method of least squares and the method of moments give identical results. This is because the polynomial is made up of the same powers of the independent variable which enter into the definition of the moments. In other cases, to be sure, the two methods will in general work out differently; and still other methods may come into consideration. But to one who has pondered on the geometrical analogies set forth above, it seems that as long as a straight line is the shortest distance between two points, the least-square criterion is likely to retain a fundamental importance, however often it may happen that one or another of a variety of different measures offers an advantage in particular circumstances. In traveling from New York to Chicago, for example, it would be highly inexpedient to attempt to follow a straight line, or even a geodesic on the earth's surface, but one realizes nevertheless that the shortest path from one point to another has a more universal significance than the topography of the Hudson River or the Alleghany Mountains.

Another contention that may be advanced is that the coefficient of correlation is applicable only when the underlying dependence of one variable on the other is linear, and that uncertainty on this point invalidates any general assertion with regard to it. To this the mathematician will reply that it is not the definition of the correlation coefficient that is uncertain when there is doubt as to linearity of regression, and in default when the regression is non-linear, but its adequacy as a measure of the relationship in question. The algebraic consequences of the definition have all the precision and finality of any group of propositions in mathematics, and have merely to share with other abstract propositions the limitation that in a practical issue they may be occasionally beside the point. Incidentally it is to be noted that resort to a correlation ratio does not necessarily imply abandonment of the concept of correlation coefficient; for example, a correlation ratio based on the representation of y by a quadratic function of x is the same thing as the coefficient of double correlation of y on the pair of variables x and x^2.

The importance of the geometrical analogies is not limited to the statistical applications. Formulated with integrals in place of sums, and extended to an enumerable infinity of dimensions, they constitute the elements of the geometry of function space, a fascinating subject which is now under active investigation, and still at so early a stage of development that its ultimate scope is impossible to foresee. If the geometry of statistical measures and the geometry of function space had not received attention independently, either would inevitably have suggested the other, so compelling is the identity of their fundamental concepts.

Without going far afield, one finds the abstract theory of approximation and the theory of statistics merging along another frontier. The theory of the least-square approximation of a given function may be extended by the introduction of a weight function in the integral to be minimized, an extension suggested in the first instance no doubt by the weighting of data in practical problems of the adjustment of observations. Instead of integrating the square of the error directly, one multiplies it first by a given non-negative function of the independent variable, and seeks to minimize the integral of the product. This introduction of a weight function is both a unifying and a generalizing principle. The Legendre series, for example, can be obtained as a special case of trigonometric approximation, by the use of a suitable weight function and a simple change of variable. Apart from particular cases, the arbitrariness of the weights naturally opens up a wide range of generalization. After problems of approximation over a finite interval have been studied in this light, it is natural to proceed to the consideration of an infinite interval, with a weight function chosen so as to make the improper integrals convergent. And one of the most obvious weight functions in this connection leads directly to the Tchebychef-Hermite-Gram-Charlier series which is of importance in the analysis of frequency distributions.

The above outline is of course only illustrative. Apart from the method of least squares, problems of curve fitting and mechanical quadrature are perennially timely, both on the theoretical and on the practical side, and constitute an important bond between mathematical science and its applications. The alternative geometrical interpretation of coefficients of correlation for the special case of normal correlation, already referred to, involves an interesting application of the theory of quadratic forms. And the whole theory of approximation represents only a part, though in its widest scope a not inconsiderable part, of modern mathematics. Many other topics could be named as of common interest to students of mathematics and of statistics. There will inevitably be a shifting of emphasis as time goes on. It appears likely that the further development of the theory of probability in the next few decades may turn out to be a major chapter in the history of science. And if this prediction is to be verified, one may further surmise that its realization will demand the profoundest insight of the professional mathematician, and will result in an unprecedented extension of exact knowledge of the material world and of the workings of human society.

DUNHAM JACKSON
UNIVERSITY OF MINNESOTA

THE INCREASED EFFICIENCY OF AMERICAN AGRICULTURE[1]

FOOD is the first and most urgent requirement of life. How to provide an adequate supply has been in the past the fundamental problem of mankind. From the dawn of history to our present highly developed western civilization, the food problem in one aspect or another has demanded public recognition. Ancient civilization developed only under conditions favorable for the production of food. Even so, a major portion of the population was engaged in its production. Only a few had leisure to devote to other occupations. Ancient civilization always declined when the production or the securing of food became difficult.

The food problem in one form or another has been to the fore in our own country from the first starvation winter of our Pilgrim Fathers to the present time, when the economic status of the food producer is one of our most urgent national problems. During this period there has taken place a fundamental change in the position which food supply occupies in the consciousness of the nation. In colonial times, nearly the entire attention of the population was devoted to food production. More than 95 per cent. of all producers were farmers. The entire family worked on the land. And yet there was produced scarcely more than enough to feed and clothe the people. The per capita exports in relation to those engaged in farming were far less than they are to-day.

Beginning with a condition when almost the entire population was rural and when the entire family worked on the farm, the efficiency of American agriculture has increased gradually until to-day less than one third of our population is rural, women and small children no longer work in the field, the length of the working day on the farm has been greatly reduced, exports of food and clothing products have increased and the problem of how to avoid the production of a surplus of agricultural products resulting in the lowering of the economic status of the producer is one of our urgent national problems.

The rapidity with which the proportion of rural to urban population has declined is strikingly shown by the census reports. Thus in 1790 rural people (those who live on farms or in towns of 2,500 or less) made up 96 per cent. of the total population; in 1880, 71 per cent.; in 1890, 65 per cent.; in 1900, 60 per cent.; in 1910, 50 per cent.; and in 1920, 48 per cent.[2] The U. S. Department of Agriculture has estimated that on January 1, 1927, only 27,892,000 persons out of a total population of about 119,000,000, or less than 24 per cent., were actually living on farms. In other words, one family living in the country produces the raw material for food and clothing for three families living in towns and cities. In addition the exportable surplus of agricultural products has increased greatly. Thus with the exception of the period of the World War a larger volume of agricultural products was exported in the five-year period 1921–25 than for any similar period in the history of our country.[3]

At no time has agricultural progress as measured by production per man been more rapid than within the past few years. Dr. O. E. Baker, agricultural economist of the U. S. Department of Agriculture, in an address before the joint session of the Farm Economics Association and the Rural Section of the American Sociological Society in 1927 said:

> Agricultural production as a whole was over 14 per cent. greater in the period 1922–26 than in the period of 1917–21, whereas population increased less than 9 per cent. between the midyears of these two five-year periods; in other words, the increase in agricultural production was over 50 per cent. greater than the increase in population.
>
> More surprising than this rapid increase in production and the resultant surplus, however, is the fact that this

[1] Address of the vice-president and chairman of Section O—Agriculture, American Association for the Advancement of Science, New York, N. Y., December 28, 1928. Contribution No. 41, Office of Director, Agricultural Experiment Station of Kansas.

[2] Census Monograph VI, Farm Population of United States for 1920.

[3] Statistical Abstract of the United States, 1926, p. 470.