STATISTICS

PROF. R. A. FISHER, Sc.D., F.R.S., LL.D. (Calcutta, Glasgow), D.Sc. (Ames, Harvard, London)

Early Sources

ALTHOUGH the rapid progress and wide applications of statistical methods during the last thirty years have made the subject one peculiarly associated with the twentieth century, it is yet to be noted that its roots go back to the Renaissance, like that other great branch of applied mathematics, mechanics or dynamics; that Pascal and Fermat were contemporary with Galileo and Newton, and that the laws of chance were developed in relation to the problems of gaming almost simultaneously with the deterministic mechanical laws, which for some centuries seemed to menace the world with an iron necessity.

The two important tributaries, developed before the twentieth century, to the modern science of statistics and to the corresponding educational discipline are, on the one hand, the theory of probability, developing through de Moivre, Laplace, and Poisson on the mathematical side, through Thomas Bayes, Boole, and Venn on the logical, and reaching out to practical applications in life assurance and demography as the aristocratic pastime of gaming fell into moral disrepute.

A second parallel development, almost independent of the first, came to be known as the theory of errors. Here the emphasis was not upon uncertainty of expectation, but upon the practical business of summarizing and digesting a considerable observational record, of the kind which systematic observations in astronomy or the routine of surveying were already beginning to produce. Gauss's name is particularly to be associated with this development, the importance of which for

From *Scientific Thought in the Twentieth Century*, (ed. A.E.Heath). London: Watts. 31-55, (1951).

mathematical studies is none the less great, although the academic mathematician, developing the subject didactically, is all too prone to lay his emphasis on the more formal development of the theory of probability.

Galton and Statistical Biology

A man who, towards the end of the nineteenth century, played a peculiar part in precipitating modern developments was Francis Galton. A man of means and, had he chosen, of leisure, Galton made his name early in life as an African explorer. In 1869, evidently reacting eagerly to his cousin Charles Darwin's evolutionary theory, he had written *Hereditary Genius*, one of the most remarkable books of the century, and in it had demonstrated how apparently intangible concepts, at first vaguely apprehended, can be made quantitative and relatively precise by the collection and adequate presentation of statistical data. Throughout his life this possibility evidently exercised a fascination on his mind. In a crude way he attempted to collaborate in discussing the numerical results of his cousin's experiments with plants. He tried his hand at the statistical expression of meteorological phenomena, and, towards the end of his long life, armed with much experience, but without adequate mathematical technique, he became convinced that quantitative, and particularly statistical, methods were needed to consolidate Darwin's ideas, and to give confidence to their practical application. In Karl Pearson he found a man of boundless confidence and ambitious energy, and, with the sympathy of W. F. R. Weldon and his wide biological knowledge, Galton believed that a solid foundation could be built for a timely advance in the method and theory of biological research.

So far as Pearson's work is concerned, the immediate

outcome was the appearance from 1894 onwards of a series of extensive memoirs entitled *Mathematical Contributions to the Theory of Evolution.* In all there were twenty-six of these, the first twelve appearing in the *Philosophical Transactions* of the Royal Society. The title chosen must be taken to represent rather Galton's hopes than Pearson's performance, for Pearson was here exploring his own general concepts in mathematical statistics, such as skew-frequency curves, contingency, and the rather numerous statistics to which, without distinction, he applied the term correlation. These developments were accompanied by extensive mathematical tables to facilitate their use. Mendel's laws are discussed in the twelfth memoir, but only to be dismissed as inadequate.

A more enduring consequence was the foundation, in 1901, of *Biometrika: A Journal for the Statistical Study of Biological Problems.* For many years this handsomely produced quarterly was undoubtedly the centre of development of mathematical statistics in this country. It accepted papers from outside Pearson's laboratory and included some of the most important advances of a period of rapid progress. In building up the high reputation of this journal Pearson's labours as editor were constant and indefatigable and constituted the greater part of the scientific activity of his later life.

Although in following his own bent Pearson undoubtedly wandered far from Galton's intention, yet he may be regarded as ploughing the ground in preparation for later developments. That the huge mass of his writings have now little value must be ascribed to two circumstances—first, that his mathematics were on the whole clumsy and lacking in penetration and, second, that without the power of self-criticism he was unable and unwilling to correct his numerous errors or to

appreciate the work of others, which would have been of the greatest assistance. He seems to have regarded observational material principally as a means of illustrating his *a priori* concepts, not as a means of correcting them or as providing problems of interpretation in which statistical methods might be of service. He seems to have felt a contempt for the work previously done in the theory of errors and to have known little about it. He certainly regarded the skew-frequency curves he invented as improvements to be substituted for the normal law of errors.

The Emergence of Modern Statistics

It was in the study of experimental data that the characteristics of modern statistics first began to display themselves, notably in that series of methods which are known as the tests of significance. The logical situation behind these is exceedingly simple, though it has been seriously obscured in recent times by the elaboration of a highly sophisticated mathematical background. This, however, is unnecessary, and all the practical progress that has been made was achieved without its aid. In general, if, in connection with a given observational record, a hypothesis is considered which is well defined in the sense that from it can be derived definite expectations, we may use the observations to test whether these expectations have been realized, or whether, on the contrary, the observations depart so far from expectation in some relevant respect that the hypothesis under consideration must be deemed to be contradicted by the data, and must be abandoned. In the latter case the deviations from expectation are said to be significant, while in the contrary case, if the observations are such that with reasonable probability they might have arisen on the

hypothesis under test, this hypothesis, though not proved, has at least so far been confirmed, and, pending further and more stringent observations, may be accepted. Obviously this process of comparing observational data with more or less vaguely conceived hypotheses has been subconsciously inherent in experimental work from the time of its inception. The hypotheses conceived, however, by the active experimenter are more or less fluid, and his judgment from the results is personal and subjective. The possibility of an exact test of significance relative to a properly defined hypothesis, and to its necessary consequences in the frequencies of the different observational records which might have arisen from it—the possibility of an objective test appropriate to a given class of cases, and specifying the level of significance of the judgment imposed by the data—is of quite recent origin, and was first clearly exemplified by a paper published in 1909 by W. S. Gosset, writing under the pseudonym "Student."[1] One feature in particular should be borne in mind when modern tests of significance are spoken of as exact; namely, that the observational material is assumed to be finite, and may consist of only a small sample. Realism and practical applicability obviously require this. Yet it was the greatest obstacle to clarity of thought to pass from the vague and necessarily inexact approximations of "the theory of large samples" to a computational procedure appropriate to real cases in which the sample is finite. It is characteristic of this difficulty that when methods appropriate to "small" samples were developed, the protagonists of earlier methods should have exerted themselves to stress the smallness of the sample as a ground for disparaging the

[1] "Student" (1908). *The Probable Error of a Mean.* Biometrika, vi, pp. 1–25.

conclusions which could be based upon it. The study of sampling problems appropriate to finite, and therefore sometimes small, samples was inseparable from the development of exact and objective tests of significance, and the underlying mathematics consisted almost wholly in the realistic examination of such problems.

Student's Test

The particular test to which Student's name has been attached concerns the precision of the mean of a sample of observations, such as a series of analyses of random samples of a bulk, the average value of which is of interest. The kind of information which such a series of analyses can give may be made clear by dividing the whole series of possible or hypothetical values of the unknown average into two classes, one of which is contradicted by the data, while the other is consistent with it, at a defined level of significance, or of stringency of the est. Such a subdivision is achieved by a test of significance by which we can determine whether or not any chosen hypothetical value is rejected by the test. Since the hypothetical value chosen is known we may subtract it from each observation and reduce the problem to that of testing whether a given sample from a normally distributed population is or is not consistent with the population having zero for its mean.

If the population sampled were of known variance it was well known that the sampling distribution of the mean of any number, N, of observations was a normal distribution with the same mean, and the variance divided by N. The probability of our mean exceeding by chance the value observed would then be easily calculable, and we might fix our level of significance by saying that deviations were acceptable if they fell within the limits fixed by definite values of the prob-

ability of falling outside these limits—*e.g.*, a not very stringent but convenient level of significance would be fixed by choosing a deviation of a magnitude which would be exceeded in one direction or the other just once in twenty trials. This comes to 1·96 standard deviation. The method therefore in use in the century before the appearance of Student's paper was to obtain as good an estimate as possible of the variance of the population sampled and to apply this calculation on the assumption that the value estimated was in fact correct.

A good, and, indeed, the best possible, estimate was available in Bessel's formula

$$s^2 = \frac{1}{N-1} S(x - \bar{x})^2,$$

in which x stands for any observation of the series, $\bar{x}$ for their mean or average value, S for summation over all the values of the sample, and N for the number of observations. Consequently, the permissible limits at the 5 per cent level of significance would on this method be given by

$$\pm \frac{s}{\sqrt{N}} (1{\cdot}96).$$

The important thing to notice here is that the statement that s^2 is the best possible estimate of the unknown variance is not the same as the statement that the variance is in fact equal to s^2. The estimate s^2, like the mean value of $\bar{x}$, will be subject to errors of random sampling, and will sometimes be greater and sometimes less than the true value. It is not obvious, and was not obvious to the mathematical traditions of the nineteenth century, in what way this circumstance should affect our conclusions.

What Student perceived was that although the

sampling distributuion of $\bar{x}$ must depend on the unknown variance of the population sampled, yet that of the ratio

$$t = \bar{x}/s\sqrt{N}$$

must be independent of this unknown variance. Its distribution would depend only on the known size of the sample N, and Student set himself to determine what its actual distribution might be. Using an extremely crude mathematical approach he arrived in fact at the correct distribution, and showed that the frequency was not proportional to

$$e^{-\frac{1}{2}t^2},$$

which would have justified the traditional method, but rather to

$$\left(1 + \frac{t^2}{N-1}\right)^{-\frac{1}{2}N},$$

a distribution for which the frequencies of extreme values fall off much less rapidly than for the normal distribution.

Student gave a simple table for evaluating the probability that t should fall outside any chosen limits. Since then fuller tables have been made available, so that it is easy at any chosen level of significance to find the corresponding limiting values of t. For ten observations, for example, the 5 per cent limits are not $\pm$ 1·960, but $\pm$ 2·262.

It will be noticed that the use of Student's criterion completely removes all doubt or embarrassment due to the size of the sample being finite and the estimate s^2 not exactly equal to the true variance. It is in this sense that the modern tests of significance are spoken of as " exact " in contradistinction to the approximate tests developed without this refinement and commonly

referred to as derived from "the theory of large samples."

This example has been set out in full, not only because its point has been missed in many expositions, but because it displays very simply the intimate integration of the whole class of advances which make up modern statistical methods. It was mentioned in passing that Bessel's formula supplies the best possible estimate of variance available from the data, and this at once raises a series of questions—How is it known to be the best? In what sense is it the best? What criteria exist for judging one estimate to be better than another?—questions which it is the function of the Theory of Estimation to answer.

Again, Student could have had no success without recognizing and solving a problem of a type at that time almost untouched—namely, the problem of the exact sampling distribution of quantities calculable from samples. These problems of distribution looked exceedingly formidable in Student's time, and although since then many have been solved it is still true to say that where there is doubt or difficulty it is usually because the relevant problem of distribution has so far baffled analysis.

Thirdly, if we express Student's t in the form

$$t = (\bar{x} - \mu)/s\sqrt{N}$$

where the unknown mean, μ, of the population sampled has been introduced explicitly, it is seen that t, the sampling distribution of which is known merely from the size of the sample, is expressible as a function jointly of the observed quantities $\bar{x}$ and s, and of the unobservable parameter μ. It constitutes an example of what have been called pivotal quantities, by which we can pass equally to probability statements about $\bar{x}$

and s for given μ, or about μ, given $\bar{x}$ and s. These latter are known as statements of fiducial probability, to avoid confusion with the statements of inverse probability, which are possible when certain *a priori* knowledge is available—*e.g.*, as to the frequency with which different values of μ shall occur in a bulk, such as has been sampled, but which cannot be derived without presuming (or assuming) such knowledge *a priori*. It is the quality of fiducial probability that it is available in the absence of knowledge *a priori*, and as an obvious consequence that its meaning or logical content must be different from that of any statement of inverse probability. Its validity is derived from the complete independence of all unknown or hypothetical elements, which the pivotal distributions enjoy.

Problems of Distribution

A class of mathematical problems which came into prominence with the initiation of more critical methods of statistics are known as problems of distribution. From the observational data, whether frequencies or measurements, quantities can be calculated aimed at representing properties of the distribution sampled. In the case of measurements there are characteristically symmetric functions of the observations. The whole class of functions calculable from the data are known as " statistics." If we know, or are given by hypothesis, the distribution of each individual observation, and know also that these several distributions are independent of each other, though expressible in terms of one or more unknown parameters, it is theoretically possible to deduce the sampling distribution of any chosen statistic, so as to determine the relative frequency with which it would fall within specified limits were the sampling process repeated indefinitely from

like material. Although this theoretical possibility is obvious, such problems appear at first sight to be extremely difficult of explicit solution. Their exact solution in an increasing number of important cases has supplied a powerful stimulus to the demand for exact tests of significance, and to the development of an adequate theory of estimation.

Speaking generally, problems of distribution can be approached by three methods. The first, which the subject owes to somewhat similar researches in statistical mechanics, consists in the geometric representation of the sample by a point in generalized Euclidean space, having as many dimensions as there are observational values. The entire sample is thus represented by a single point, and the frequencies with which samples of all possible kinds will appear in the random sampling process is represented by a frequency density in the manifold space employed. For any chosen statistic, if we assign it a particular value, the samples which will lead to this value will fall generally on a continuous sub-space, so that the generalized space may be regarded as subdivided in a series of slices by chosen values of the statistic. The total frequencies found in these regions gives the frequency distribution of the statistic in question. In simple cases, for example, space may be divided by a series of parallel planes, or by a series of concentric hyperspheres, and the frequency distributions, such as those of $\bar{x}$ and s^2 in Student's problem, may be found with little difficulty. This geometric method has indeed been found more successful than any other, and to it are due the majority of the exact solutions now known.

A second method is to proceed by " mathematical induction " from a sample of given size to a sample containing one additional observation or set of observa-

tions. If the distribution is known for a sample of size N, it may be possible to infer, knowing that the next observation is independent of those that precede it, the distribution for size $N + 1$, and so to obtain the general formula.

A third method, using a more advanced type of mathematical analysis, is based upon what is known as the characteristic function. It has been widely discussed and is of great theoretical simplicity, but so far has not led to the resolution of difficulties recalcitrant to other methods. If x is a variable quantity of known distribution, then for any real value of t it is demonstrable that the average value of e^{itx} exists mathematically, and has an absolute value not greater than unity. Regarded as a function of t this is known as the characteristic function. Given the characteristic function, the corresponding frequency distribution may be inferred by an inverse process involving integration.

Now, although the distribution of a given statistic calculable from a sample of observations offers a problem requiring some penetration and insight, yet it is possible without consideration to write down formally an expression for its characteristic function in terms involving a finite number of processes of integration. Consequently, by using the inversion mentioned above, we have here a method of the utmost generality for solving problems of distribution.

The Theory of Estimation

During the long period in which the exact sampling distributions of the statistics commonly in use were unknown and thought to be unattainable in practice, the properties of such distributions could not be used as a means of comparing the merits of such different methods of statistical reduction as might suggest them-

selves. Indeed, the belief was expressed that it was impossible to compare the merits of two methods of estimation without reference to the particular purpose for which the estimates were required, and that the notion of a best estimate was only a subjective illusion.

As soon, however, as statistics are regarded as having knowable sampling distributions, a number of comparisons suggest themselves as representing unquestionable *desiderata* in any quantity used in estimation. In the first place it will usually be true for any appropriate statistic, that, as the sample from which it is drawn grows larger and larger, its range of distribution will grow narrower and narrower without limit, so that it is said to converge in probability, in the sense that there is a limiting value T_∞, such that the probability of the absolute value of the deviation of the statistic from this limit exceeding any chosen quantity ϵ, however small, shall tend to zero. If this is true, then the limiting value will depend only on the population sampled, and will necessarily be some function of the parameters by which this population is characterized. The estimate is said to be consistent when the statistic T is used as an estimate of this particular function of the parameters and not of any other. We may say, then, that T is a consistent estimate of a parameter θ if in large samples it converges in probability to the limit θ. Although the use of inconsistent estimates is not unknown in some special cases where the error involved is exceedingly minute, yet I believe the desirability of satisfying this criterion has never been challenged. In what follows it will be assumed that the estimates spoken of are consistent estimates.

In all cases of practical importance it is found that as the sample is made larger without limit, the form of the distribution of the estimate about its limiting value

tends to assume that of the normal law of frequency of error discovered by Laplace and Gauss, and that the variance of the distribution then falls off inversely to the size of the sample. In other words, if v is the variance of the distribution and N the size of the sample, the product Nv tends in general to a finite limit. Evidently this limit measures in an inverse sense the precision of the estimate under discussion, at least in large samples. The second criterion that has been developed, known as the criterion of efficiency, is that this limit shall be as small as is possible. Statistics satisfying this second criterion as well as the first are said to be efficient. Since the minimum of the limiting value of Nv is directly calculable from the data of any problem, it is comparatively easy to test whether any proposed estimate is efficient, and if not to recognize that it must be capable of improvement. The recognition of this criterion swept away the claims of methods of estimation which had been most confidently advocated about the beginning of this century. The utility of this advance was the greater since it was shown that an efficient estimate in single or simultaneous estimation could always be obtained by the method of maximum likelihood, and that, starting with any consistent method, routine calculations would give the numerical values with whatever precision might be desired. On the practical plane our second criterion provides, therefore, a comprehensive solution of the problem of estimation.

On the theoretical side the problem at this point is still far from closed. Many different efficient estimates may be proposed. These will, indeed, tend to equivalence as the samples are made larger, but for any finite sample they will differ in value. They will differ also in the form of their error distributions, and evidently

one may be preferred to another. The analysis, however, of the criterion of efficiency points the way to a further means of discrimination. It has been pointed out above that the limiting value of the product Nv in large samples has a minimum which does not depend on any chosen method of estimation, but on the sampling properties of the data. This property specifies in an inverse sense how valuable each unit of the original data was for the purpose of estimation in view. In fact, the reciprocal of this minimum limit is defined as the quantity of information provided by each unit of the data, or in simple cases by each original observation.

Now, given the exact sampling distribution in a finite sample of any proposed statistic, it is equally possible to ascertain the quantity of information which a single value of this statistic will provide, and to compare this quantity with that provided by any alternative statistic derived from a sample of the same size. The notion of efficiency, based originally on the limits for large samples, may thus be extended to the comparison of statistics derived from finite samples. It may be shown that the amount of information in an estimate can never exceed the finite amount of information supplied by the original data, consequently the criterion of efficiency when extended to small samples implies merely that the loss of information, if any, incurred in the process of statistical reduction, shall be made as small as is possible.

A particular class of statistic was early shown to satisfy a condition more stringent than that of the criterion of efficiency—namely, that no loss of information whatever is incurred. In such cases the statistic is said to be sufficient. What this means in practice is illustrated by a further property of sufficient statistics—namely, that if T is a sufficient statistic and

T' any other statistic whatever, then the simultaneous distribution of T and T' in random samples shall be such that for any chosen value of T the distribution of T' shall be wholly independent of the value of the parameter θ to be estimated. In other words, as soon as we know the value of T, then the value of T' is completely irrelevant to the estimation of θ, and supplies no information whatever about it. Since, in the case of sufficient statistics, this is true of all alternatives which can be proposed, we are using our words consistently when it is said that the sufficient statistic contains all the information about θ originally present in the data. When sufficient statistics exist thay can be found by the method of maximum likelihood, and in such cases we have a method of estimation which is satisfying not only practically but also theoretically.

It will be noticed that sufficient statistics do not exist in all cases, and that their existence depends on the functional form of the problem. To complete what is here said on the theory of estimation, it should, therefore, be added that in a number of cases in which sufficient statistics, in the strict sense, do not exist, yet it is possible to render estimation exhaustive, and to avoid all loss of information, by the use of what are known as ancillary statistics, which possess the properties that while they themselves have random sampling distributions independent of all unknown parameters, yet that they, together with the estimates obtained by maximum likelihood, form an exhaustive set. The use of ancillary statistics does not alter the estimate arrived at, but does alter its sampling distribution, and thus supplies additional information which, without their use, would be lost.

The logical form in which the information supplied by observational data about unknown and hypothetical

parameters or about observations not yet actually made, should be expressed, deserves close attention. In the early work of Thomas Bayes, published 1763, a serious attempt was made to bring the problem within the scope of the theory of probability, so that on the basis of the observations, statements of probability could be made as to the value of the hypothetical parameters. It appears that, in general, it is not possible to set the results accurately within this framework of ideas. Bayes' approach was accepted somewhat uncritically by Laplace as the basis of his theory of inductive inference, but from the middle of the nineteenth century a succession of writers, beginning with Boole, pointed out that this approach is open to criticism and leads to contradictions. These criticisms appear to be unanswerable, and the theory of inverse probability, originating with Bayes, is now almost universally abandoned. The existence of pivotal quantities having the properties explained above, with distributions independent of all unknowns, and expressible in terms both of observable statistics and of one or more of the unknowns, does, however, allow probability statements about these unknowns to be inferred from the known distribution of the pivotal quantity. Statements derived in this way are known as statements of fiducial probability, and the corresponding distributions of the parameters as fiducial distributions. In practical applications these are often referred to as providing confidence limits.

Information and Experimental Design

The development of the precise quantitative notion of the quantity of information transformed the statistician's task in two ways. It had been his business to make the most of the data available. So long as grossly defective methods of doing so were current, and recom-

mended by the highest authorities, it was natural that precise or reliable results obtained should stand to the credit of the statistical method employed; in the contrary case, if no tolerably precise results were obtainable, the statistician presented his work with a somewhat apologetic air. When, however, it became manifest and easily demonstrable that any finite body of numerical data contained only a finite amount of information, and that methods of extracting either the whole, or very nearly the whole, of this had been made available, it became obvious that so far as precision was concerned, the statistician as such had no responsibility. His task was fulfilled, if he correctly assessed the amount of information, and presented his results embodying the whole of it relevant to the questions under discussion. At this stage his task resembles that of a chemist making an assay, and it would be absurd for him to be ashamed if the assay is low, or elated should it prove to be high. The precision of the results and their value for all purposes are inherent in the body of numerical data originally available.

Relieved of one responsibility, statistical work is in the position to undertake another. The methods by which quantitative information is assessed reveal also to what particular features of the observational record its limitations are principally due. We are in a position to estimate how many observations and of what kind would be needed to give results of any required precision. It is this very fundamental change of outlook that has turned the attention of so much modern statistical work in the direction of the design or planning of observational records, and of experimental arrangements. Most modern text-books of statistics deal now explicitly with some aspects of experimental design with reference, in particular, to the needs of

quantitative biology. The history of the art of experimentation is, of course, a long and complicated one, and although it is now most readily apprehended by the student through the exposition of formal principles, yet these have come relatively late in its historical development. The driving force and urge to improvement has lain in the diverse enterprise of thousands of workers exercising their curiosity and ingenuity in overcoming particular difficulties. In such fields as biology, agriculture, psychology, it early became evident that the technique of experimentation traditional in the so-called exact sciences of physics and chemistry was not sufficiently penetrating. Exploration of knowledge in the biological fields required in particular *controlled* experiments—*i.e.*, experiments in which it is not the absolute values of the observed quantities which are of primary interest, but comparisons between these. The principles of the complex subject of biological control were, however, very hazily understood and, for more than a generation, the word " control " was used in an almost reverentially mystical attitude.

Need for control arises from the experimenter's consciousness that he is ignorant of innumerable causes which may affect his experimental results. It is for this reason that controlled experimentation was first developed and refined in sciences which did not think of themselves as " exact." If plants are grown in a greenhouse, the experimenter is aware that the conditions in which they are grown may not be the same as those prevailing in other years, or in other greenhouses. He prefers, therefore, to rely on comparisons between groups of plants grown together in the same greenhouse. If he wishes to know whether the addition of an element, such as boron, to the solution in the water culture affects the growth of plants, he will make

sure not only that a number of his plants are treated with boron, but also that a number grown in parallel with them shall be without this addition. The object of his experimental comparison is to make sure that if a significant difference does appear in the growth of the two groups of plants, it shall properly be ascribed to the addition of boron ; and also, what is a different problem, to make sure that if boron is in fact capable of affecting appreciably the growth of the plants, his experiment shall be competent to detect this effect. Obviously, also, if there are many different questions in the mind of the experimenter, more complex systems of control may need to be elaborated. Although traditionally a particular group of experimental units such as the plants receiving no boron have been spoken of as " controls," in contradistinction to the other groups of plants in the experiment, yet in the design of the experiment, and in the treatment of the results, no special role is played by this control group, for all should be thought of interchangeably as controlling one another.

Naturally, from the beginning of controlled experimentation a great deal of care has been given to render the circumstances and treatment of the different groups of experimental units closely alike in all respects save those under test. It has only more recently been realized that differences in environment between units assigned to the same treatment are equally of vital significance to the success of the experiment, and to rendering it capable of an unequivocal interpretation. At its simplest this general principle arises from the fact that we can only judge of the significance of the difference in performance of groups of units treated differently by comparison with similar differences in performance between units treated alike This being so, it became obviously imperative for the validity of the comparison

that the physical conduct of the experiment shall ensure that the very same uncontrolled causes of diversity which affect units treated differently shall equally act between units treated alike. The modern experimenter, therefore, takes care that nothing in his experimental arrangements shall allow unknown causes of diversity to affect either group of comparisons without equally affecting the other. The most general means to effecting this is by what is called randomization.

A simple example may be used to show what randomization amounts to. In one of his experiments Darwin compared the growth rates from germination of two groups of maize seedlings, one from self-fertilized and the other from cross-fertilized parents. Darwin arranged his seedlings in pairs, one from cross-fertilized and the other self-fertilized, growing each pair side by side in the same pot. He took the greatest care that these two plants should be comparable in, for example, the date of germination. It is probable that all the seedlings from self-fertilized plants were in a line along one side, whereas those from cross-fertilized plants were in a parallel line along the other side of the pots, perhaps one line was on the east and the other on the west. Since it is not known *a priori* which side would be favoured, owing, for example, to differential insolation, or to the convection currents circulating in the greenhouse, the comparison between one group and the other was fair or unbiased, in that neither was known to enjoy any advantage from such causes. On the other hand, such causes are not known not to exist, and if they exist, they will affect all the plants in one line differentially from their opposite numbers in the other line. There was, therefore, in Darwin's experiment the possibility of a cause affecting differentially comparisons

between the different groups of plants, without affecting equally the differences between the relative performances of the different pairs. All that the statistical examination of the results can ascertain is that the difference between the two averages of fifteen plants each is rather greater than could be ascribed to chance, judging from the variation among the fifteen differences in the fifteen pairs of plants. The statistical analysis does not, therefore, exclude the possibility that this apparently significant result was really due, in part at least, to the sites assigned to the two treated groups.

If, however, randomization had been applied in the experimental design,the members of each pair would be assigned to the west or east side of their pots at random, as by tossing a coin. Any significant difference between the performances of the two contrasted groups would then be known to be ascribable to the effect of self-fertilization. Randomization is a device designed to ensure that the laws of chance used in testing the significance shall be validly operative in the physical conduct of the experiment.

To summarize briefly the steps by which statistical studies have gained their present level of usefulness, we may say that during the present century we have learnt

(i) To conserve in its statistical reduction the scientific information latent in any body of observations.

(ii) To conduct experimental and observational inquiries so as to maximize the information obtained for a given expenditure.

In accomplishing these tasks of immediate utility, it has incidentally furthered the task of experimental

science by incorporating the vast amorphous body of research experience, which has been accumulating for centuries, in principles of scientific inference which can be taught to students. The art of adding to natural knowledge by experimentation is no longer the " mystery " of a craft, but, in a sense in which it has not been before, part of the heritage of a scientific education.