DONALD MAINLAND | *Department of Medical Statistics, New York University College of Medicine*

THE RISE OF EXPERIMENTAL STATISTICS AND THE PROBLEMS OF A MEDICAL STATISTICIAN*

THE NARRATOR'S OBSERVATION POINT

If one desired a complete and objective history of a military campaign, one would hardly expect it to be written by a soldier on leave from his regiment which was still in action. His story could not be complete, and it would certainly overemphasize the activities of his own unit. And yet, are not the accounts of eye witnesses, of actual participants, the raw material from which histories are written? If, therefore, the title of this paper were fully descriptive, it would contain the phrase: "A contribution of raw material."

In gathering such raw material from an eye witness one should, of course, find out where he was when the events occurred, and so it is desirable to indicate briefly the experiences on which the remarks in this paper are based. The remarks are those of one who, after graduating in medicine, started research on the embryology of the ferret. He was puzzled by error in cell measurement, by variations in the counts of chromatin particles, and especially by the problems of small samples, because ferrets were expensive. He obtained no answer from biologists, chemists, physicists, or mathematicians, until he was led to the solution by Dr. C. H. Goulden, an agricultural experimenter in Winnipeg, who introduced him to the book by R. A. Fisher[4] (now Sir Ronald Fisher) on *Statistical methods for research workers* which had appeared three years previously (in 1925).

He then saw that the methods prescribed by Fisher for avoiding bias and allowing for chance, experimental error, biological variation, and sample size, were applicable in all fields of medicine, and that in his own field, anatomy, normal variations in nerve pattern, positions of abdominal viscera, and other organs had often greater clinical importance than the so-called "averages" that he had learned and was teaching. He stole time from the contemplation of the fifteen arteries, which (in those days, at least) arose from the hypogastric artery, and spent this time in applying statistics to his own research and the researches of others, some of it during vacations in Fisher's laboratory. Then, after World War II he realized that he could

* Presented at the annual meeting of the American Association of the History of Medicine, May 6, 1954, New Haven, Connecticut.

not run an anatomy department and keep abreast of statistical developments, and he chose the latter effort.

This personal history will perhaps indicate where bias may occur in the remarks that follow, but the bias will, it is hoped, be more apparent than real. No one in medical statistics can ignore the debt to Raymond Pearl and his predecessors, who promoted a critical attitude toward medical data long before the rise of experimental statistics.

The brief autobiography serves also to illustrate what has happened during the past three decades to a number of workers in various fields of science, and here a kind of bias does enter. It can best be described in the words of a representative of a pharmaceutical house who came recently to the Department of Medical Statistics at New York University to discuss his search for a biological statistician. He said: "In my quest I have met many statisticians, and I have been struck by the enthusiasm of those who came into the field from the outside. I do not find the same enthusiasm among the younger people who have been formally trained as statisticians." Another way of expressing this would be to say that the older crowd, whatever their present professional status may be, are still "amateurs" in the original Latin sense of the word. The remarks that follow, therefore, will contain the bias of an amateur.

"STUDENT" AND FISHER

It was the patron saint of the amateur statisticians who wrote the paper that can be looked upon as the starting-point of what is called "experimental statistics," or, more exactly, "experimenters' statistics"—the paper that appeared in 1908 entitled "The probable error of a mean," by "Student,"[18] the pseudonym of W. S. Gosset, an experimenter employed at Guinness's Brewery in Dublin. In brewing experimentation "the variable materials and susceptibility to temperature change and necessarily short series of experiments" (McMullen[8]) showed, as Fisher[4] later expressed it regarding all laboratory investigation, that "the traditional machinery of statistical processes was wholly unsuited to the needs of practical research. Not only did it take a cannon to shoot a sparrow, but it missed the sparrow!"*

The method demonstrated by "Student" was later called the "t" test, and this led to the analysis of variance method for the comparison of two or more means, which was named the "F" test by Snedecor[12] in honor of Fisher who developed it.

* It is significant that in this quotation the present tense of the original can be changed to the past tense.

The next landmark in the presentation of statistics for experimenters was Fisher's *Statistical methods,*[4] already mentioned, which was produced while its author was at the Rothamsted Agricultural Experimental Station, and is now in its eleventh edition. Ten years later (in 1935) appeared his *Design of experiments,*[5] now in its sixth edition.

The far-reaching effects of these two books, and of Fisher's numerous papers, justify calling the past thirty years of statistics the Fisherian period (or sometimes the "Piscatorial" period); and the logical connection between the books is noteworthy. The first was concerned largely, though by no means wholly, with significance testing, i.e., with proper methods of making allowance for the effects of chance. The second book emphasized the fact that a significance test has no useful meaning unless an experiment has been properly designed. In terms of a simple experiment, the comparison of two treatments on animals or patients, "properly designed" means designed in such a way that, at the end of the experiment, one can say: "Chance would so rarely cause such a large difference in outcome that I shall attribute the observed difference to the treatments." There must be only two possibilities: chance and the treatments; and this situation can be reached only by allocating the treatments to the subjects by what Fisher[5] called "a physical experimental process of randomization," which is now most easily performed by a table of random numbers.

This demonstration of the logical necessity of randomization is one important contribution of Fisher's second book on statistics. Another is his demonstration of how to study multiple factors in the same experiment. He wrote: "In expositions of the scientific use of experimentation it is frequent to find an excessive stress laid on the importance of varying the essential conditions *only one at a time* . . . and it is often supposed that [this] is the essentially scientific approach to an experimental investigation. This ideal doctrine seems to be more nearly related to expositions of elementary physical theory than to laboratory practice in any branch of research." He further called attention to the way in which different factors interact with each other, and demonstrated designs that could most economically reveal not only the main effects but the interactions. In brief, the principles are: (i) balance, i.e., the subjection of equal numbers of subjects to each factor under test, and (ii) randomization of all the factors, known and unknown, that are not under test. (The economy possible by such methods can be illustrated by an experiment recently concluded on seven x-ray technique factors (kilovoltage, developing, fixing, and so on) in which only 128 films were needed, in contrast to 8,192 films that would have been required if the

so-called "scientific" method of varying only one factor at a time had been used.)

These and many other methods that Fisher developed are much more than experimental techniques. They were the first results of a revolution that is still continuing in the hands of others, a revolution by which statistics came to embody the principles of inductive inference—the experimenter's logic.

STATISTICS IN APPLIED SCIENCES

It is important to note where the statistical revolution started and where it spread—in applied sciences. It began in brewing and agriculture, spread to animal husbandry (as in the work of Snedecor[12] in Iowa), to the cotton industry in England, and now, either by research or statistical quality control or both, statistics plays a part at some stage in the production of probably everything that we use or consume. Noteworthy, because of its connection with medicine, is the pharmaceutical industry, where, especially in bioassay, some of the most valuable advances in statistical method have occurred.

In all the applied sciences, inefficient or wrong methods of research or production cause loss of money. Therefore, sound experimentation was profitable; and so applied chemistry and physics adopted modern biological statistics while academic chemists, physicists, and even biologists were disregarding the revolution or resisting it, largely through ignorance.

What, then, about medicine, which in a scale of human values may be considered the highest of the applied sciences? This is the field where quality of work is to be measured by something greater than loss or gain of dollars. It is the field, also, where there is much lamentation over the walls that have arisen between the various "subjects" in teaching and research and over the lack of basic principles of thought, such as are offered by modern statistics.

EXPERIMENTAL STATISTICS IN MEDICINE

The first example in "Student's"[18] paper of 1908 dealt with the data from a clinical trial of soporifics, but it could hardly have had less effect on medicine if it had dealt with a constellation in outer space. In 1929 there appeared in *Physiological Reviews* a 124-page article by H. L. Dunn[8] who reported that out of 200 quantitative medico-physiological papers in current American periodicals over 90 per cent required statistical methods and did not use them. The rest of the article was occupied with a description of statistical tests, mostly pre-Fisherian, with formulae and examples. This

article was distributed widely and doubtless helped to increase the incidence of statistical tests in medical literature. In 1932 Greenwood,[7] the English medical statistician, remarked that "medical papers now frequently contain statistical analyses, and sometimes these analyses are correct, but the writers violate quite as often as before, the fundamental principles of statistical or of general logical reasoning."

In the past twenty years the increase in the incidence of tests—statistical arithmetic—has continued, and so also, very commonly, has the disregard of the more important contribution of statistics, the principles and methods of sound, economical experimentation and valid inference. Of the various causes of this neglect, one has been clearly demonstrable in some investigators—the fact that medical science had its roots largely in academic physics and chemistry. Another obvious cause is the common human tendency to use gadgets instead of thought. Here the gadgets are the arithmetical techniques, and the statistical "cookbooks" that have presented these techniques most lucidly, without primary emphasis on experimentation and logic, have undoubtedly done much harm.

However, all has not been darkness. Lights have appeared in various laboratories and clinics. Thus, in the 1930's proper attention was given by Gaddum[6] in England and Bliss[2] in this country to the fact that individual animals of the same species, sex, age, and weight differ in their response to drugs and toxins. In the clinical field there appeared in 1948 a beacon or lighthouse beam—the report of the British Medical Research Council's[11] co-operative trial of streptomycin in pulmonary tuberculosis.

Experimentation in medicine is, therefore, improving, not only by devices such as randomization and the balancing of samples, borrowed from other biological fields, but by methods specially appropriate to investigation on human beings, such as the "double blind" method, to avoid bias both from the patient and the observer.

Much research on human beings must, however, remain nonexperimental —attempts to show the causes of disease, interrelationships between diseases, biochemical and other differences between them. Problems of biased samples in such research have long been the concern of public health statisticians, but the light of experimental statistics has shown the difficulties more vividly than before. One example has been called "Berkson's fallacy," to indicate the one who first demonstrated it, Dr. Joseph Berkson of the Mayo Clinic.[1,9] The fallacy is not difficult to grasp, and a research worker in any field, from anatomy to psychiatry, who appreciates its implications, will be perturbed by it, for it may affect all his researches. This is not the place to describe it, or to discuss in detail other recent developments in medical

statistics, but for some discussion of such topics, reference may be made to a section of the latest volume of *Methods in medical research.*[10] In that section an attempt has been made to meet some current problems, especially the problem of a scarce commodity—experienced medical statisticians.

THE STATISTICIAN'S PROBLEMS IN MEDICINE

"Meeting problems." The phrase recalls a remark made by Dean Currier McEwen in 1950 during the discussion of plans for a Division of Medical Statistics at the New York University College of Medicine. Dr. McEwen's remark was: "There will be lots of problems; but it will be fun meeting them." That remark, reinforced by the helpful attitude of colleagues in other departments, has provided the motto for the first four years of the experiment at New York University.

The Division became a separate Department in 1953, and some have said that it is the first such department in a medical school on this continent, independent of a department or school of hygiene, public health, or preventive medicine. Whether that is so or not, the Department is probably a forerunner of other such units and is meeting problems that they will meet. Problems are inevitable when anything is established that does not fit an existing pattern, and such a department does not. It is an academic department, but resembles also a service unit, concerned with maintenance or supplies. The chief problems can be grouped under three heads: Lack of time, lack of money, and lack of personnel. All the problems entail misconceptions of the nature and functions of statistics, misconceptions that are understandable by reference to the foregoing remarks on the history of experimental statistics. In feeling one's way toward solutions of the problems, one must keep constantly in mind the true goal: To help to improve medical research.

Lack of time. In any active research center a medical statistician will, sooner or later, find himself submerged by requests for help. He will find that many investigators have no idea how long it takes to provide proper diagnosis and treatment for a case—a research project—even for one that looks quite simple; and they are naturally disappointed and perplexed when, after days or weeks of inquiry and thought, the statistician has to say: "No treatment is warranted."

It may be asked: "Would it not be more sensible to give a little help to a lot of projects than a lot of help to a few?" This is an extremely dangerous suggestion. The "little help," on superficial acquaintance with the project, is commonly worse than none at all. A significance test, which "rules out chance," makes the investigator feel that he has a clean bill of health, that

his results are "statistically proved." But chance is often the least important cause of wrong conclusions; and the "t's" and "sigmas" and "chi-squares" in medical literature are often just spurious science.

A statistician who refuses to do arithmetic unless he knows that it is justified, who tries, in a phrase reminiscent of the Hippocratic Oath, "to keep his art inviolate," will attempt to do as much work as is necessary on as many projects as possible. He will become harassed and will develop a feeling of guilt because he is unable to keep his promises and holds up other investigators' work, even for months.

Toward his assistants, also, he will feel guilt, which he may express somewhat as follows: "I am too old to care much about my own future, but I do care about the fate of my young assistant who needs to get publications out, to earn a higher degree, to obtain salary increases. His time is taken up by work for others. He plans an experiment, is constantly on call during the months when it is running, then analyzes the results and presents a report; and for all this he gets a half-line of acknowledgment. Even if his name were in the by-line whenever he had helped in the investigation, he should, as a member of an academic staff, have adequate time to work on his own problems. Should I recommend that he join a commerical company? Life there would not be free of pain, but a few thousand dollars' extra salary would provide a good analgesic. What should I do?"

The answer that seems to be emerging from experience is twofold:

1. Find out who, in any other academic department, spends the greatest number of hours in teaching. Subtract the teaching time of the statistician of comparable rank, and let the difference be the time that the statistician devotes to his or her "service" functions.

2. Schedule the "service" work at least six months in advance. Select the projects that you can handle adequately in the available time and that seem most likely to bring you nearest to your main goal, the improvement of medical research. Decline the rest. This will be unpleasant, for you will have to refuse help (apparent help) to your friends in projects that are valuable; but even if it were right to give the whole time of your department to this kind of work you would be far from meeting the need.

Lack of money. When a person starts a statistical unit in a medical school, he is likely to be given a very small staff and a small budget and to be told that, as in other departments, he will have to look elsewhere for further help. He will probably not find that help easily. When he explains what he is trying to do—develop the science and art of statistics, and train workers properly—he may meet a remark like that made in 1951 by a high official in one of the foundations for the aid of medical education and

research: "A statistician cannot show an experimenter how to do an experiment." What, then, should a statistician do to raise money? Should he charge his colleagues in other departments for his advisory services? No; even a charge for a computing clerk's services, to help pay her salary, can at times injure the relationship between colleagues.

One recommendation, again based on experience, would be: "If you have been an experimenter in any field, carry on experimental work, applying for grants as do other departments." (Before accepting an appointment it is well to stipulate that facilities be available for research of the kind contemplated.) Even if there were no need to raise money, the same recommendation would be made, for thereby is created the environment which, as a glance at recent history has shown, is best for the development of the science and art of statistics.

Lack of personnel. There is no simple or speedy solution for the problem of the scarcity of medical statisticians; but in seeking for a solution three points should be borne in mind:

1. The first point is essentially a repetition of the last remark in the preceding section. The most important single element in the training (and continuous education) of any statistician is practical experience—experience of investigations for which he himself is responsible, with all their difficulties and disappointments. It is, therefore, very satisfactory to note that in the new scheme of training at Yale—a most important experiment—the emphasis is on actual investigation and not on mathematics.

2. If medical schools hope to increase the supply of statisticians, they must see clearly what is implied. They should no longer imagine that any one person can be competent in vital statistics, experimental statistics, epidemiology, cost accounting, mathematical theory, and the analysis of hospital records (if they are worth analyzing).

3. Although, as time goes on, the supply of suitable statisticians will increase, they will not be the whole answer to the need for statistics in medicine, because the need is for better statistical reasoning by investigators. A course in statistics for medical students can go only a very short way toward meeting this need. It could go somewhat farther if students, in college and medical school, were shown the difference between a proper experiment and what is called an experiment in their laboratory courses, or if some of the elementary exercises in chemistry and physics were transformed into real experiments, or, perhaps, if children in kindergarten or elementary school became really familiar with chance and bias through games with marbles or disks.

A rather more hopeful method of meeting the need for statistical reasoning in medical research, which has apparently not been systematically tried, would be the development of a "statistical nucleus" in each department (or group of cognate departments in a small school). This nucleus would be one of the regular staff members (biochemist, physiologist, physician, or pathologist, according to the department) who would have a sound grasp of the general principles and of a few suitably chosen investigational designs, along with the methods of analyzing the resulting data. The designs would have to be simple, but that would, in itself, be a blessing, because many experiments in medicine are too complicated, too "ragged," and very wasteful. The success of the scheme would depend largely on two things:

1. The attitude of the head of the department and other investigators.

2. The characteristics of the person who was to be the "nucleus." He should really desire to learn and understand and should be able to rid his mind of mistaken ideas and prejudices. Although probably young, he should be mature enough to be patient with the notions of his seniors, but at the same time firm when he knew he was right. If in doubt, before committing himself he should call in a professional statistician.

DISCOURAGEMENTS AND REWARDS

The scheme just outlined would bring difficulties and discouragements, and indeed much of this paper may seem discouraging to one who would like to become a medical statistician. Therefore at the end let the voice of an amateur be heard again. The problems are numerous and heavy; but they seem trivial in face of the satisfaction—the thrill—that can come when several people are planning together a piece of research, when the distinctions between physician and surgeon, chemist and statistician, fade away, because all are trying to practise the difficult art of straight thinking—all are striving to create a good investigation.

REFERENCES

1 Berkson, J.: Limitations of the application of fourfold table analysis to hospital data. Biometr. Bull., 1946, *2,* 47.

2 Bliss, C. I.: The calculation of the dosage mortality curve. Ann. Appl. Biol., Lond., 1935, *22,* 134.

3 Dunn, H. L.: Application of statistical methods in physiology. Physiol. Rev., 1929, *9,* 275.

4 Fisher, R. A.: *Statistical methods for research workers.* 1st ed. Edinburgh and London, Oliver and Boyd, Ltd., 1925.

5 Fisher, R. A.: *The design of experiments.* 1st ed. Edinburgh and London, Oliver and Boyd, Ltd., 1935.

6 Gaddum, J. H.: *Reports on biological standards: III. Methods of biological assay depending on a quantal response.* Medical Research Council Special Report Series No. 183. London, H. M. Stationery Office, 1933.

7 Greenwood, M.: What is wrong with the medical curriculum? Lancet, Lond., 1932, *1,* 1269.

8 McMullen, L.: Foreword to *"Student's" collected papers.* E. S. Pearson and John Wishart, editors. London, Biometrika Office, University College, 1942.

9 Mainland, D.: The risk of fallacious conclusions from autopsy data on the incidence of diseases, with applications to heart disease. Am. Heart J., 1953, *45,* 644.

10 Mainland, D. and Herrera, L.: Statistics in medical research. Sect. III of *Methods in medical research.* J. M. Steele, editor. Chicago, The Year Book Publishers, Inc., 1954, *6,* 121.

11 Medical Research Council: Streptomycin treatment of pulmonary tuberculosis. Brit. M. J., 1948, *2,* 769.

12 Snedecor, G. W.: *Statistical methods applied to experiments in agriculture and biology.* 1st ed. Ames, Iowa State College Press, 1937.

13 "Student" [Gosset, W. S.]: The probable error of a mean. Biometrika, Cambr., 1908, *6,* 1.