Commentary

The use and misuse of statistics in medical publications

Research workers' widespread lack of understanding of the rationale of statistical techniques, and the frequent use of statistical tests as a substitute for thoughtful investigational design, meticulous work, and repetition of experiments, justify the antagonism to statistics exhibited by some experimenters. To one who has had personal experience of the way in which statistical thinking, as distinct from statistical arithmetic, can promote good investigation, this perversion of statistics is lamentable. It appears to be due, not so much to investigators themselves, but to the order in which experimenters' statistics was developed by the pioneers and presented to research workers, because Fisher's "Statistical Methods" (1925), which discussed chiefly significance tests, preceded by ten years his "Design of Experiments," which showed how to plan experiments in order to obtain unambiguous inferences from the tests.

Other causes of misunderstanding are discussed, and, in an effort to promote a more rational attitude to statistics, nine suggestions are presented.

Donald Mainland, M.B., D.Sc.* *New York, N. Y.*
Department of Medical Statistics, New York University School of Medicine

"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." This was written by Major Greenwood,[5] the English medical statistician, in 1932, in the interval between the appearance of the two books by R. A. Fisher (now Sir Ronald Fisher) that have been chiefly responsible, directly or indirectly, for the spread of statistical techniques among experimenters—*Statistical Methods for Research Workers* (1925)[3] and *The Design of Experiments* (1935).[4] The techniques have entered every field of pure and applied science; and yet in 1950 Lancelot Hogben,[7] the experimental biologist who has become a medical statistician, could assert, without fear of serious contradiction, that "less than 1 per cent of re-

This paper was written as part of a project entitled "Promotion of Biometrical Methods in Medical Research," supported by a grant, RG-6100, from the National Institutes of Health, U.S. Public Health Service.

*Address, 550 First Avenue, New York 16, N.Y.

search workers clearly apprehend the rationale of statistical techniques they commonly invoke."

The spread of statistics

During the decade since Hogben wrote, statistical techniques have continued to spread. In medicine two prominent examples are controlled clinical trials, and the search for causal factors in chronic diseases by a method which is called "epidemiologic," but is that of sample-survey statistics.

The spread of statistics is, however, something more than the increased use of experiment designs and arithmetic tests. A book entitled *An Introduction to Scientific Research,* published in 1952 by E. Bright Wilson,[13] a professor of chemistry, is essentially an application of statistical thinking to the performance of laboratory experiments. In *The Common Sense of Science,* Bronowski,[1] an applied mathematician, writes: "This [statistics] is the method to which modern science is moving. . . . This is the revolutionary thought in modern science. It replaces the concept of *inevitable effect* by that of the *probable trend.*"

Perversion of statistics

To a medical research worker who, soon after the first appearance of Fisher's *Statistical Methods,* started using the new ideas and techniques in his own work, saw their relevance to all medical research, and gradually obtained the label "statistician," the triumph of statistics in the past thirty years might be expected to bring unalloyed pleasure. It does not. His impression, from reading medical literature and acting as consultant or collaborator in research, is that, although statistical thinking has spread, the misuse of statistical arithmetic has spread faster. It sometimes occurs to him that he might increase his income considerably by a standing bet with all contributors to certain prominent medical journals that their statistical tests were either unnecessary, misapplied, or misinterpreted; and there would be little additional risk in keeping the bet open for articles in which a statistician's name appeared in the acknowledgments or even among the coauthors. Unfortunately, the evidence on which the bets could be decided would often be missing. Procedures, circumstances, and events in an investigation may seem so unimportant to the investigator that he may not even remember them, and yet, if known, they might render the statistical analysis ridiculous.

The foregoing paragraph sounds like the usual statistician's criticism of investigators. On the contrary, it expresses a lament that a method of thinking, planning, and action which could lead to a better investigation is often perverted, so that it becomes a set of gadgets that do more harm than good. This was doubtless what impelled a biochemist to declare recently that one cannot be a biostatistician and a good biochemist, for he had observed a phenomenon that has become common in laboratory publications. The verdict of a small-sample significance test, whether it is "significant" or "not significant," appears very convincing to those investigators, editors, and manuscript reviewers who do not know how little it really tells them. Therefore, a significance-test devotee can achieve a much higher manuscript acceptance rate per unit of time than one who tests his first results by abundant repetition of his experiments. Significance testing thus becomes a substitute for thought, clean experimentation, and perseverance. Worse still, I have seen careful and critical experimenters, long resistant to statistical tests, become converted, and then draw from their tested data inferences which, in the days before their conversion, they would rightly have greeted with derision.

A cause of perversion. A clue to one cause of the perversion of statistics can, I believe, be found in the order of publication of Fisher's two principal books on the subject. *Statistical Methods* was concerned very largely with the presentation of significance tests—devices whereby an investi-

gator could discover how often random (pure chance) variation, as in card shuffling or in taking of samples from a box containing thoroughly mixed uniform disks, would produce differences of the magnitude that he met in his investigations. *The Design of Experiments* ten years later showed how to set up experiments in order to utilize this knowledge of the results of chance—how to make chance work for us, e.g., by card shuffling or the use of random numbers in assigning treatments to subjects, so that at the end of an experiment we could say: "The cause of the observed difference in outcome was either the randomization or the factors [fertilizers, drugs, or the like] that we introduced. The randomization has taken care of the effects of the innumerable, often hidden, factors that we did not eliminate (or balance in some systematic fashion) at the outset."

The order of appearance of Fisher's two books represented the development of knowledge and methodology by pioneers; but it gave statistical arithmetic a ten years' lead on statistical design, and even after *The Design* appeared, this lead was maintained or increased. Although *The Design* presented fundamental principles, applicable to a wide variety of experiments, its chief flavor was agricultural, because that was the area where the theory and methods had been worked out. To many medical laboratory workers whose predecessors had for generations been designing experiments which had led medicine from empiricism into science, there seemed little need for trying to apply to their work the principles enunciated by an agricultural statistician whose lines of thought and style of presentation made comprehension difficult even for statisticians.

Moreover, scientific methodology in the abstract is not interesting to many people, whereas techniques, which seem to have an immediate practical application, gain readier attention. Laboratory workers and clinical scientists came increasingly in contact, through statistical "cookbooks," with such techniques—statistical test recipes, both those that Fisher had invented and earlier recipes, which had originated with statisticians such as Karl Pearson. Thus there started the epidemic of statistical arithmetic. At first sight, investigators might appear to blame; but were they?

Assumption of random variation

In trying to answer that question it is interesting to look at the earliest and most familiar of the "small-sample" techniques, the *t* test, demonstrated by "Student" (the brewery chemist, W. S. Gosset) in 1908.[12] As an example of the application of the *t* test, "Student" used data of Cushny and Peebles[2] from research on the sleep-inducing properties of hyoscyamine and hyoscine; and Fisher used the same data as his principal example of the *t* test in *Statistical Methods*. Because "Student" apparently miscopied the names of the drugs in his paper we shall label the compounds "A" and "B," as did Fisher after the first few editions of his textbook. Each of 10 patients received drugs A and B on different occasions, and the numbers of hours of sleep after administration of each drug were recorded. The data to be tested, therefore, comprised ten B — A differences in hours of sleep, and "Student's" conclusion from the *t* test was that the odds were about 666 to 1 that B was a better soporific than A.

One of the assumptions underlying this conclusion, although not explicity stated, was that the experiment had been so conducted as to control, as if by card shuffling or disk sampling, the effects of all the factors, except the drug difference, that could cause a B — A difference—in other words, that the intersubject variation in the B — A differences was strictly random variation. In Fisher's presentation the difference was described as "clearly significant" since the *t* value obtained from the sample was 4.06 and with 9 degrees of freedom "only one value in a hundred will exceed 3.250 by chance." This statement does not, in so many words, attribute the difference in outcome to the difference in drugs, but the

inference is implied, and it becomes even more evident in Fisher's discussion of the signs of the ten B — A differences, of which 9 were positive, one difference being zero. He wrote: ". . . if the two drugs had been equally effective, positive and negative signs would occur with equal frequency." Here was the same implicit assumption that "Student" had made—that the experiment had been so conducted that nothing could have caused the B — A differences except chance and the drug difference. The nature of this assumption, and how to make it more than an assumption by randomization, became clear in *The Design of Experiments*, but the same wording of the statement about positive and negative signs in Cushny and Peebles' data has continued into the latest (1954) edition of *Statistical Methods*.

This comment is not hair-splitting semantics; nor is the example unique. The "prime object" of *Statistical Methods* was "to put into the hands of research workers, and especially biologists, the means of applying statistical tests accurately to numerical data accumulated in their own laboratories or available in the literature." The reader who used the book for laboratory reference was cautioned to "work through, in all numerical detail, one or more of the appropriate examples, so as to assure himself, not only that his data are appropriate for a parallel treatment, but that he has obtained some critical grasp of the meaning to be attached to the processes and results." This sounds like an excellent safeguard; but, in fact, this thorough study of the examples leads the investigator no nearer to the "meaning to be attached to the results" than did "Student's" or Fisher's treatment of Cushny and Peebles' data. It leaves to the opinion (or faith or desire) of the investigator the decision that effects found in his data are due only to the factors he is testing plus chance (strictly random error). For my part, I would be much more ready to accept, without any test at all, an experimenter's opinion that an observed difference between two sets of measurements would rarely (or commonly) occur if he wrote the measurements on cards and shuffled them a thousand times, than I would be to accept his opinion that his method of assigning treatments had been "equivalent to" randomization.

"Random variation" seems to be often considered synonymous with "whatever variation the research worker has not taken care of by a systematic investigational scheme such as stratification, or by subsequent arithmetic, such as covariance adjustment." Such ideas have probably arisen from our rather loose concepts of "chance" as comprising a multitude of causal factors that we cannot, or do not, identify. The same line of thinking probably accounts for the impression which some investigators still retain, that a significance test is a magic device for determining whether there was, or was not, some hidden bias in their investigation—as if a small value of P means a small probability of bias. This perhaps explains the description of the function of a statistician that a research worker suggested recently, as follows: "Most of us are not Claude Bernards; therefore we need a statistician to tell us whether our conclusions are probably right or probably wrong."

The *Statistical Methods* pattern of presenting techniques has been imitated by most of its simplified successors, and it is still often imitated by statistical writers in medical journals who present recently developed techniques. Therefore, the laboratory workers who in 1960 affirm that all they require from a medical statistician are mathematical formulas and computing services can point for their defense to a tradition sanctioned by the founders of experimental statistics and confirmed by their followers.

Sometimes I wonder whether the spread of perverted statistics would have been arrested if in 1935 *Statistical Methods* had been rewritten, with the fundamental ideas from *The Design of Experiments* attached to every numerical example, or if a fraction of the time spent by medical investigators and their helpers during the past

quarter of a century on statistical arithmetic had been devoted, instead, to the study and application of the main ideas contained in the first half-dozen chapters of *The Design of Experiments.*

Experimental proof and significance testing

Some of us who have been acquainted with *The Design* throughout its career dip into its first few chapters again from time to time and discover statements that have more meaning for us now than they had on the first (or sometimes the tenth) reading. In sect. 7, for example, is the following passage: "In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure . . . we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result." If instead of "a statistically significant result" we used a phrase such as "a result in the same direction and of similar magnitude," Fisher's statement would express a fundamental part of the philosophy of experimentation, and would be remarkably like one of the strongest criticisms that experimenters aim at statistics.

The quotation must seem rather strange to those who think of Fisher as the creator and deifier of significance tests; but it acually reveals rather well the attitude that he has displayed in personal communications and in various written statements. Those workers for whom $P = 0.05$ has become a rigid stop-and-go sign might profitably reread his remarks on first presenting his table of P values for chi square (*Statistical Methods,* sect. 20): "In preparing this table, we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy."

Fisher's view could, I think, be fairly expressed as follows: Significance tests are useful tools to *help* an investigator to decide what to do next—not to *tell* him what to do. His data may show "no reason to suspect the hypothesis tested," but "the null [no-difference] hypothesis is never proved or established, but is possibly disproved, in the course of experimentation" (*Design of Experiments,* sect. 8); and the investigator may have good reasons for not dropping the search. He can then increase the sensitiveness of his experiment quantitatively by taking a larger sample (*Design,* sect. 11) or qualitatively by reorganizing the structure of his experiment or by refining its technique (*Design,* sect. 12), or, of course, by all three methods.

Confidence limits

Even more useful than a significance test as a help for the investigator in deciding what to do next, and in revealing to him how little his data tell him, is an answer to the question: If, despite there being no proof of a difference (e.g., between the effects of two drugs) a real difference exists, how large or how small may it be? Although not expressed so simply, this was the kind of question that was answered in Fisher's discussion of the probable limits of a mean difference (*Design,* sect. 62), and it is unfortunate that this matter, so important to research workers, became obscured by arguments among statisticians regarding differences in the reasoning processes underlying "fiducial" limits and "confidence" limits.

Nine suggestions for developing a more rational attitude to statistics

As time goes on, more investigators will learn to distinguish the counterfeit, useless, or dangerous from the genuinely scientific

and useful elements in statistical procedures; but gadgets and bad habits spread more easily than does critical thinking, and unless we make positive efforts improvement will be slow. As an attempted contribution to such efforts, nine suggestions have been prepared.

Some statisticians may dislike the criticism implied in some of the remarks, but I believe that many of them will agree that it is better to encourage, and try to answer, penetrating questions, than to wait until investigators discover for themselves the spurious elements in statistical practice and by their criticism cast suspicion on all statistics. Hogben's[8] recent attack on Fisherian and other statistics is so wholesale and so difficult for many research workers to follow, that statisticans probably need have little fear of its consequences. But I doubt whether the same equanimity could be maintained if some research workers, well acquainted with the real meaning of statistical tests, estimates, assumptions, and predictions, were to go in detail through a number of investigations in which statistical help had been obtained, and were then to describe, in terms that other research workers could understand, the details of the investigations, the statistical analyses, and the conclusions that had been drawn.

The nine suggestions that may help us to develop a more rational attitude to statistics are as follows:

I. Avoidance of preconceptions and prejudice. To obtain a fresh look at statistics, we should try first to get rid of our preconceived ideas about it and our prejudices either against it or in its favor.

II. Knowing what we are doing. If we are going to use statistical ideas and the resulting techniques in the design, conduct, and analysis of our investigations, and yet retain the title "investigator," it seems essential that we know what we are doing, and why we are doing it. Statistical thinking is, essentially, thinking about variation (i.e., differences between things, events, or phenomena that bear the same label) and about the methods of dealing with variation; and if we are content to let someone else do this thinking and tell us what to do, we are, it seems to me, accepting the role of technician.

III. Experiments and surveys. A useful way of starting to clarify our thoughts about analysis and inference is to distinguish between an experiment, in the strict sense, and a survey. In an experiment, in the strict sense, we assign the factors under test *at will* to the individuals that comprise our experimental material (patients, animals, or bacterial culture tubes); therefore, we can assign by a method (randomization) which leads to the "either chance or the factors under test" type of inference.

Most clinical researches and many laboratory investigations are not experiments in the strict sense, because we do not (often cannot) assign the factors under test (e.g., diseases) at will. Such investigations are best called "surveys," however small the sample size and however complex the procedures to which we subject our material. We may speak, loosely, of "Nature's experiments" in the assigning of diseases or other features, and there is always some randomness (several factors acting independently of each other) in such phenomena; but Nature does not try to randomize, and we know, from card shuffling, disk sampling, and the like, what prolonged efforts are necessary to remove trends and clusters. Therefore, although we may, for practical purposes, choose to accept the results of a survey as a demonstration of a causal relationship, we should remember that a survey, or even a million surveys on the same topic, can demonstrate only an *association* between the factors under test and the phenomena they seem to cause.

IV. The basis of statistical arithmetic. Whenever we use statistical arithmetic we should insist on seeing as clearly as possible the reasoning and knowledge on which it is based. This does not mean trying to understand the mathematical proof of a formula, because mathematical proof is no proof that the formula is safe in the real world.

To illustrate, let us suppose that we have conducted a clinical trial, after randomization (e.g., by card shuffling) of drugs A and B to 20 patients each, and that we emerge with 16 "successes" in the A group and 10 in the B group. The question that we wish our significance test to answer is this: If we performed a large number of card shuffling trials, always with 40 cards, 26 marked "S" (success) and 14 marked "F" (failure), and in each trial dealt them into two piles, 20 A's and 20 B's, in what percentage of trials would we find as strong apparent evidence of an A-B difference as we did in our clinical trial, i.e., 16 (or more) S's in one pile and 10 (or fewer) in the other pile? We have, of course, already decided that if this percentage is less than 5 per cent (or less than 1 per cent or some other figure), we are going to accept it as evidence that the randomization in the clinical trial did not account sufficiently for the observed difference in outcome.

Testing fourfold frequency tables. There are four ways in which we could find the required percentage of random arrangements:

A. We could actually perform the card shuffling trials, say a thousand or more. This would be time consuming; but in more complex problems, where we have no mathematical short cuts, this Monte Carlo ("gambling") method is much used.

B. We could write out all the possible arrangements of 26 S's and 14 F's in two groups of 20, and find the percentage of these arrangements that met our requirements—16 or more S's in one group, 10 or fewer in the other group. At this point we should watch our reasoning. We would be implying that, if we performed card shuffling trials, each of the possible arrangements, which we had written out, would occur with equal frequency—more exactly that, as we continued the shuffling trials, we would find that the percentage frequencies of each of the possible arrangements would approach equality. The reason for this belief is based on some centuries of experience of games of chance.

C. The writing out of all possible arrangements is, of course, not practical, because their number, even with these small samples, would run into many millions; but fortunately we can obtain the information that we need by a mathematical method, which is based on the discovery, made about the end of the seventeenth century, that the binomial expansion represents what is found in certain kinds of games of chance.

In the fifth (1934) edition (sect. 21.02) of *Statistical Methods,* Fisher showed how to use the binomial expansion for problems like our two-sample success-failure data. His "exact test for 2×2 tables" enables us to find, not the millions of possible individual arrangements, which we do not need to know about, but the percentage frequencies of the various classes of these arrangements, such as 16 S's in one sample of 20 with 10 S's in the other sample, 17 and 9, 18 and 8, etc. The easiest way for an experimenter to obtain insight into the method is to take an imaginary group of, say, 8 subjects containing 5 S's and 3 F's, and write out all the possible ways (70) in which two samples of 4 subjects can be formed. It will be found that 14.286 per cent of these arrangements contain 4 S's in one sample and 1 S in the other, while 85.714 per cent contain 3 S's in one sample and 2 S's in the other—exactly the proportions found by applying the "exact" method, either from Fisher's description or from more detailed arithmetic instructions.*

D. The "exact" method is rather laborious, and therefore the chi-square test for 2×2 tables is commonly used instead. This was devised, before the "exact" test, for comparing two samples taken at random from the same "infinite" population; that is, the conditions are not quite the same as the random assignment of a specified number of S's and F's in two finite samples. Therefore, in order that we may trust the chi square test as a substitute for the exact test,

*For example, Mainland,[10] p. 274.

we require empirical evidence—numerous comparisons of the two tests. Such comparisons[9,10] have shown that chi square with "Yates' correction" is, with rare exceptions, a safe test for significance at the 5 per cent and 1 per cent levels if we demand a chi square value greater than 4 as an indication of P less than 0.05 and a value greater than 7 as an indication of P less than 0.01. The rare exceptions are easily detected by applying certain precautions.[10] Even the arithmetic of chi square, and the residual doubts regarding its safety, can be avoided for pairs of equal samples containing up to 500 individuals in each sample, and for unequal samples up to size 20, by merely taking our data to published tables.[11]

Testing measurement data. As we have seen, it is not very difficult to appreciate what is going on in a 2×2 table frequency comparison. In other statistical arithmetic, especially tests of, and estimations from, measurement data, we run into assumptions. Worse still, we often do not run into them; we are either not told about them, or we are told about them so cryptically that we do not appreciate their implications.

One of the assumptions made in many statistical tests and estimates is the Gaussian frequency-curve assumption; e.g., tables for use in the *t* and F tests are derived from mathematically exact Gaussian distributions.

An assumption that is often involved in the interpretation of tests is the "homogeneity" of variation within the different groups that are compared or combined—that is, the assumption that the intragroup variation does not differ from group to group more than in random samples from the same population.

In making estimates after regression analysis many investigators hardly seem to be aware that they are assuming a straight-line relationship—an assumption which, if they really thought about the phenomena under study, they might seriously question.

Strictly speaking, the assumptions are that our measurements do not depart from the prescribed conditions (the Gaussian curve, equality of variation, linearity of regression, and so on) enough to vitiate our tests and estimates. Therefore, it might be supposed that if we apply tests to our data (e.g., a test for skewness or for difference in intrasample variation) and obtain a nonsignificant result, the assumptions will be safe; but this is by no means true.

Although the Gaussian assumption is perhaps not as dangerous as others, at least when it is involved in the comparison of mean values, we can use it as an example, because its defenses are multiple.

THE GAUSSIAN ASSUMPTION. First, we may recall the remark that has been attributed, in various forms, to several different authors: "Everybody believes in the Gaussian law—the experimenters because they think it can be proved by mathematics, the mathematicians because they think it has been established by observation." We shall look at three defenses of the Gaussian assumption:

A. It has been shown that if the variation between measurements is caused by factors that are independent of each other, even if the factors are few in number (four or five), the frequency distribution resembles a Gaussian distribution.

B. It has been shown mathematically that, whatever the shape of the frequency distribution in the parent population of measurements, when random samples of the population are taken, the frequency distribution of the *means* of those samples becomes more and more nearly Gaussian in shape when the sample sizes are increased.

C. Textbooks and statisticians often assure us that explorations of data have shown that we shall "seldom be led astray" if, under certain (rather loosely defined) conditions we use tests or estimates derived from the Gaussian distribution.

These explorations sound analogous to the exploration that has given us confidence in the 2×2 chi square; but the situation is very different. In testing chi square, randomization trials or display of all possible arrangements was not necessary, because

Fisher's "exact" test gave the required information; but when we have a set of measurements and wish to make all possible arrangements of them in two or more samples of a specified size, there is no mathematical short cut to the exact results. Except with extremely small samples, this permutation labor is too heavy; therefore randomization is employed instead. By the use of random numbers the cards bearing the measurements are arranged in two samples (or in more than two samples when the F test is being studied), and the t (or F) test is applied. This is repeated, say 1,000 times, to find whether the various t (or F) values occur approximately with the frequencies required by the Gaussian distribution theory.

The other great advantage in the validation of chi square is the simplicity of the data. We are concerned simply with the numbers of X's and not-X's, whatever X may stand for; and a pattern of agreement between chi square and the "exact" test quickly emerges—for example, the agreement improves as the X's and not-X's become more nearly equal in number. By contrast, possible varieties of samples of measurements are innumerable. Therefore, we may laboriously validate t or F for one set of measurements but feel no safety in applying the results to a set of measurements of a different kind, even if sample sizes are the same as in the set we have tested.

Because of the heavy labor of empirical testing, very few extensive explorations have been made; and the same is true of the validation of most other assumptions. Electronic computers reduce the labor greatly, but it would take a long time to explore the infinite variety of medical measurement data. Actually, research workers' confidence in the tests and estimates seems to depend largely on the fact that statisticians or textbooks have shown them how to perform the arithmetic; and statisticians' confidence in the techniques often seems rather analogous to some physicians' faith in uncontrolled clinical experience when evaluating drugs.

GAUSSIAN ASSUMPTIONS AND CLINICAL "NORMALITY." Perhaps a medical research worker feels justified in accepting at second hand the statisticians' faith when he feels unable to judge for himself; but this hardly justifies him in exceeding the statisticians in faith, especially in matters where he can obtain some direct evidence. For example, the Gaussian assumption perhaps often does little harm when we are comparing mean values; but this does not justify us in trusting it for individual measurements. So many frequency distributions of anatomic, physiologic, and biochemical readings are obviously non-Gaussian that it is surprising to find clinicians accepting the Gaussian curve—the "normal" (i.e., standard) curve in the mathematical sense—as a standard biologic phenomenon when they are assigning upper and lower "normal" limits in the clinical sense, and accepting the standard deviation, a mathematical convenience, as if it were a biologic standard.

In choosing standards of clinical "normality" it is much more reasonable to make no assumptions regarding the shape of the frequency distribution and to use, instead of multiples of the standard deviation, the easily comprehended and arithmetically simple method of percentiles.[6]

Questions regarding assumptions. Second-hand faith is not a sound basis for a rational attitude to statistics. Therefore, before doing any statistical arithmetic we ought to ask three questions:

A. Instead of doing this arithmetic, how could we find the information that we desire by randomization trials? In general, for significance tests the trials would be of the card shuffling type, for confidence limit estimates they would be of the disk sampling type; but for each particular test or estimate we ought to be as specific as possible regarding the procedure.

B. What are all the assumptions that underlie the arithmetic?

C. What is the risk that the arithmetic, based on these assumptions and applied to our particular data, will give us an answer

(significance verdict, confidence limit, or other estimate) that will lead us to a different course of action than would the answer obtained by prolonged randomization trials?

For the reasons indicated above, the reply to the third question is often vague; and I am beginning to wonder if we can ever feel confident that the risk of wrong action is negligible unless a method which avoids the assumptions would give us the same verdict as an assumption-cluttered method. That is why rank-order tests in which measurements, in ascending order, are replaced by ranks (1,2,3,etc.), have a strong appeal, apart from their intelligibility and their arithmetic simplicity. They are not quite as sensitive in detecting real differences as are *t* and F tests; but often a slight increase in sample size will compensate for this defect. Theoretical statisticians ought to be encouraged to invent a wide variety of rank-order tests, and significance tables to use with them.

V. Obeying the rules of the game. If we wish to play the game of statistical arithmetic we ought to obey the rules of the game. For example, let us suppose that we perform an experiment on a certain number of animals, test the result (e.g., by *t* or chi square) and find that it has not quite reached the chosen level of significance. If we repeat the experiment on some more animals, pool the data with the previous set, test again and accept the result if it is "significant," we are fooling ourselves and others. The probability tables of *t* and chi square say that, if we take *random* samples from the same population we shall meet such and such values (of *t* or chi square) in a certain percentage of trials. But the samples to which we applied our final test were not random; they were determined in part by what we found after the first experiment. Therefore, when there is no real (population) difference we shall find more than 5 per cent of differences "significant at the 5 per cent level." If we wish to do step-by-step testing, we must adopt another design, usually a "sequential" design, and that, also, has rules that we must obey.

VI. Disagreeable doubts. We ought not to dodge disagreeable doubts. For example, at the end of an experiment we say that the alternative causes of the result are "either chance (our randomization) or the factors under test"; but how do we know that our randomization was truly random? Random numbers, which we commonly use nowadays, are safer than any single card shuffling or disk sampling, for they have been extensively tested by comparing them with what would occur if card shuffling could be perfect; but for randomizing in a particular experiment we may have picked an area where there is some clustering or systematic sequence of digits. Again, in spite of every precaution in a double blind trial a leak of information may have occurred. All that we can say after any single experiment is that we believe the risks are trivial compared with the allowance that we make for purely random variation; and we must remember that, in Fisher's words, "an isolated record" is not an experimental proof.

VII. Limitations of confidence limits. When we meet, or make, confidence limit estimates we should know their limitations. If a consignment of screws or tables or raisins is sampled by a random method and in the sample Y per cent of the items are defective, we can find lower and upper confidence limits, X and Z per cent, and make a statement such as: "The proportion of defective items in this consignment may lie anywhere between X and Z per cent, but there is a probability of at least 95 per cent that it does not lie outside those limits."

If after a laboratory or clinical experiment the same kind of statement is made regarding percentage frequencies or mean differences or other measurements, we should change the wording after "but" into a form such as "*if* our sample *were* a strictly random sample of its population there *would be* a probability of at least 95 per cent that. . . ." In dealing with meas-

urements we should often add other "if" clauses, such as "if the frequency distribution of the population were strictly Gaussian."

In whatever detail we describe our laboratory animals or patients, we cannot safely consider them as strictly random samples of the populations so described. The great value of confidence limits, therefore, is that they reveal how little we know by showing us how little we would know, even if our samples were strictly random. The only way to learn something about the safety of our numerical or other findings is by more extensive exploration, i.e., repetition of the experiment under other conditions, in other places, and at other times, and by probing more deeply, to discover underlying mechanisms.

VIII. Is the technique really useful? This is a very effective question in our efforts to evaluate statistical methods—to break away from habit and tradition, and to avoid being mesmerized by methods of experiment design and analysis invented by some statistical mathematicians "who do most valuable work in the theoretical development of the subject, but who have no serious interest in the applications of statistical methods to happenings in the real world . . . [and] seem to imagine that they can design and analyze experiments on pigs one day and on pig iron the next without knowing anything about the personal pecularities of either animal."*

"Is it useful?" implies several questions such as "Does this technique (of design or analysis) tell us what we actually wish to know?" and "Is it reliable enough for our purpose?" If we cannot answer such questions, surely we ought not to employ the techniques.

IX. Avoiding the blind use of techniques. The thorough application of statistical thinking to the design, performance, and analysis of an investigation is an art in which even experienced practical statisticians make mistakes. *A fortiori,* even a senior research worker, if he has previously done nothing with statistical techniques except apply tests, is likely to go astray, even in apparently simple projects, unless he is willing to be guided in planning and conducting his investigation by someone (whether labeled "statistician" or not) who is truly able to guide him.

To be a safe guide a statistician must, as Bradford Hill has said, immerse himself in the particular project "up to his neck." Some statisticians, it is true, are willing, or are forced by financial or other pressures, to dip no more than a finger or two into projects in which they will be held responsible for analyses and inferences. Probably this is one reason why many investigators, administrators, and research-sponsoring agencies appear to think either that one or two fingers will suffice or that a statistician has an unlimited number of necks. Actually, there are not enough suitable persons to guide, in two or three projects per investigator, more than a small fraction of the investigators who are willing to be guided. Therefore, it might seem that we must continue for years to witness such pathetic events as the arrival at a statistician's office of a junior research worker or graduate student to have a test done because his research advisor fears that an editor, or the editor's statistical referee, or a thesis-review committee will frown upon a report that is lacking in tests. A more likely, and more fearful, alternative is that more research advisors will learn to do the tests themselves and show their disciples how to do them.

Such a depressing prophecy could, I think, be falsified if investigators would stick boldly by a belief to which we all pay lip service: the belief that a research worker can contribute valuable information if he will confine himself to efforts that are within his knowledge and skill, plan his work with much thought, perform it meticulously, record the observations in detail, criticize his results severely, and offer a modest—"it seems as if"—conclusion.

A report of such work often reveals ba-

*Finney, D. J.: Personal communication.

sically statistical thinking, and it is a pity that the author does not know how much he could be helped by more knowledge of the art; but this is no reason why he should sprinkle his report with sigmas, *t*'s, F's, P's and the like, when he has not planned and conducted his work so that these things will have a real meaning, and when he does not understand what they mean. Why should it not be permissible for him to state that he had not the knowledge, or adequate guidance, to design and conduct the research in such a way as make statistical tests and estimates meaningful?

Attitude to statisticians

If we developed a better-informed attitude to statistics we would develop a more realistic attitude to statisticians. For example, if we took our data to a statistician *after* an investigation we would choose one whose office would merit the title bestowed on the office of Professor Greenwood whose remarks were quoted at the beginning of this article—"The cold water department." We would vie with the statistician in the hunt for defects in our work, and would be suspicious of those statisticians who applied tests to such *post facto* data, unless they did so in order to reveal defects and limitations. If we reported any of their tests we would report, also, their precise interpretations. We would be suspicious of those statisticians who are willing to give us advice at the outset of our investigation and then, having had little or no contact with the work during its progress, perform analyses and give us unqualified positive answers at the end.

We would, of course, condemn those investigators who in their reports shelter themselves behind statisticians' skirts, or who, in order to win approval for a grant application, name a statistician as cooperator without first obtaining his permission and giving him time (usually several months) to cooperate in preparing the plan.

A hope for the future

The foregoing suggestions are presented in the hope that they will be supplemented by other writers, and will, by reducing the perversion of statistics, enable medical research workers to benefit from true statistical thinking.

References

1. Bronowski, J.: The Common Sense of Science, New York, 1959, Random House, Inc.
2. Cushny, A. R., and Peebles, A. R.: The Action of Optical Isomers. II. Hyoscines, J. Physiol. 32:501-510, 1905.
3. Fisher, R. A.: Statistical Methods for Research Workers, Edinburgh and London, 1925, Oliver and Boyd.
4. Fisher, R. A.: The Design of Experiments, Edinburgh and London, 1935, Oliver and Boyd.
5. Greenwood, M.: What Is Wrong With the Medical Curriculum? Lancet 1:1269-1270, 1932.
6. Herrera, L.: The Precision of Percentiles in Establishing Normal Limits in Medicine, J. Lab. & Clin. Med. 52:34-42, 1958.
7. Hogben, L.: Chance and Choice by Cardpack and Chessboard, New York, 1950, Chanticleer Press, vol. 1.
8. Hogben, L.: Statistical Theory. The Relationship of Probability, Credibility and Error, London, 1957, Allen and Unwin.
9. Mainland, D.: Statistical Methods in Medical Research. I. Qualitative Statistics (Enumeration Data), Canad. J. Research, Sect. E, 26:1-166, 1948.
10. Mainland, D.: Elementary Medical Statistics. The Principles of Quantitative Medicine, Philadelphia, 1952, W. B. Saunders Company.
11. Mainland, D., Herrera, L., and Sutcliffe, M. I.: Statistical Tables for Use With Binomial Samples—Contingency Tests, Confidence Limits, and Sample Size Estimates, New York, 1956, New York University Department of Medical Statistics.
12. "Student" (Gosset, W. S.): The Probable Error of a Mean, Biometrika 6:1-25, 1908.
13. Wilson, E. B., Jr.: An Introduction to Scientific Research, New York, 1952, McGraw-Hill Book Company, Inc.