Commentary

The significance of "nonsignificance"

Although statistical textbooks have for a long time asserted that "not significant" merely implies "not proven," investigators still display confusion regarding the interpretation of the verdict. This appears to be due to the ambiguity of the term "significance," to inadequate exposition, and especially to the behavior of textbook writers who in the analysis of data act as if "not significant" means "nonexistent" or "unimportant."

Appropriate action after a verdict of "nonsignificance" depends on many circumstances and requires much thought.

"Significance" tests often could be, and in some instances should be, avoided; then "nonsignificance" would cease to be a serious problem.

Donald Mainland, M.B., D.Sc.* *New York, N. Y.*
New York University Medical Center

"My usual attitude is that if the statistical analysis has any merit, then the nonsignificant difference is automatically presumed to have *no significance whatever.*" I am not sure whether this remark by a clinical pharmacologist was made quite seriously or with tongue in cheek, but its purpose was clear—to apply a stimulus or perhaps an irritant, and evoke some further comments on "significance tests," which have been criticized already[4] as a cause of the "perversion of statistics." The response to the stimulus was twofold:

1. A search for an early manifestation of "nonsignificance" in biologic statistics.
2. A scrutiny of some of my own recent work in clinical drug trials to see if what I thought I believed about "nonsignificance" was what I really believed.

A backward glance

Perhaps most biologic and medical workers who use the terms "significant" and "not significant" associate them chiefly with R. A. Fisher (the late Sir Ronald Fisher), who did so much to develop and propagate "significance tests." Actually, however, the tests and terms antedated Fisher's work considerably. To find an example I dipped into a collection of the early statistical papers of Karl Pearson,[8] who provided much of the mathematical foundation of biologic statistics upon which Fisher and others

This paper was written as part of a project entitled "Promotion of Biometrical Methods in Medical Research," supported by grant GM-06100 from the National Institutes of Health, U. S. Public Health Service.

Received for publication June 20, 1963.

*Professor of medical statistics and member of the study group on rheumatic diseases. Address, 550 First Avenue, New York 16, N. Y.

have built. In a paper published in 1896, part of a series of *Mathematical Contributions to the Theory of Evolution,* data on statures and other body measurements are analyzed, and differences are stated to be "significant" or "not significant" by reference to the "probable error," which is approximately two thirds of the standard error. The dividing line between the two verdicts does not appear very rigid, but it became more so during the first 20 years of this century, i.e., before the publication of *Fisher's Statistical Methods for Research Workers* (1925).[1]

A "significant" difference was defined as a difference that was rarely found when random samples were taken from the same population, and "rare" came to be defined as outside the range mean $\pm 2\sigma$ in a Gaussian distribution. This range excludes slightly less than 5 per cent of the total distribution, and 5 per cent was adopted also as the cut off point in judging the "significance" of samples from non-Gaussian distributions. "Not significant" meant "not rare" in the same sense, e.g., within the $\pm 2\sigma$ range in a Gaussian distribution.

The statistical textbooks that I read when first starting to apply statistical techniques to research data (in 1928) made a clear distinction between a verdict of "not significant" and an inference that there was no "real" difference, i.e., an inference that the samples had actually come from the same population. It was pointed out that "not significant" meant simply "not proven." This distinction was emphasized also by the teachers and textbook writers who after 1925 spread Fisherian techniques among the various sciences. Fisher[2] himself expressed it in *The Design of Experiments* (1935, sect. 8) as follows: "The null [no-difference] hypothesis is never proved, but is possibly disproved or established, in the course of experimentation." At the same time the textbooks made it clear that the cause of a "not significant" verdict might be solely the smallness of the samples—that a very real difference might be found if larger samples were studied.

Sources of confusion. Since the lesson regarding "nonsignificance" has been clearly and repeatedly enunciated for about half a century, one may well wonder why the clinical pharmacologist's remark at the beginning of this paper should be necessary, and why some drug-trial investigators believe that a "nonsignificant" difference, if it does not quite mean "no difference," means an "insignificant" or "unimportant" difference. A facile explanation could be that the teachers have done their best but the pupils (the investigators) have failed to learn the lesson properly. I doubt if the teachers or the originators of the techniques and terminology can escape so lightly.

One cause of the confusion appears to be the incompleteness of expository statements. In presenting his now familiar table of the probabilities of chi square, Fisher,[1] referring to the testing of what he later called the "null hypothesis," wrote: "If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested" (*Statistical Methods,* sect 20). This statement, taken by itself, would surely suggest that a "nonsignificant" difference could be taken as a strong indication of the nonexistence of a difference. That is why, when that statement was previously quoted in this JOURNAL,[4] it was followed in the next paragraph by Fisher's other statement, published 10 years later, that the null hypothesis could never be proved or established.

However, it was not necessary to jump 10 years in order to find a qualifying statement necessary for the passage quoted from section 20 of *Statistical Methods,* and necessary also for all the examples of "significance testing" throughout that book. In the introductory chapter of the same book (sect. 4) is the following statement: "In conformity with the purpose of the examples the reader should remember that they do not pretend to be discussions of general scientific questions, which would require the examination of much more extended data, and of other evidence, but

are solely concerned with the critical examination of the particular batch of data presented."

If in the interpretation of a verdict we supplemented the previous quotation in the light of this last statement, we could say: "If P is between .1 and .9 there is certainly no reason to suspect the null hypothesis, so far as we can judge from this particular batch of data taken alone." Fisher appears to have been overoptimistic in his belief that readers would supply the necessary limiting and qualifying clauses here or elsewhere in his publications.

The terminology itself is, I think, largely to blame for the persistent confusion. Our thinking today might have been much clearer if from the beginning the statisticians had used a more specific, more explanatory, and less pretentious term such as "random frequency test" instead of the ambiguous and grandiloquent "significance test."

The more recent terms, "Type I error" and "Type II error," when once the concepts are grasped, can help to remove the confusion.

I believe also that there would have been less misunderstanding if the inventors and expositors of tests had emphasized the conditional nature of the verdicts. "The difference was not significant at the 5 per cent level." "P was greater than 0.05." Such statements seem to give us positive information about our observed samples—the actual chance of their occurrence in the real world. What the tests really say should be expressed by such phrases as the following: "If pairs of random samples were taken from the same population, more than 5 per cent of them would show differences as large as (or larger than) the difference in the observed samples." "If the method of random assignment of treatments that was used in this experiment were repeated many times, even if the treatments did not differ in effect there would occur differences as large as this (and larger) in more than 5 per cent of the randomizations."

Perhaps the fallacy of equating "no significant difference" with "no real difference" could have been driven home by an expression that is reminiscent of the formal logic by which our forefathers were supposed to learn how to think. Letting D represent the difference in size between, say, the means of two samples, we can express the fallacy thus:

"If samples are random samples from the same population, differences of size D are common.

"Our sample difference is of size D.

"Therefore our samples came from the same population."

Even if we used the phrase "probably came" we would be no nearer to a valid inference.

The presence in statistical textbooks of some of the foregoing expressions might have helped to clarify readers' ideas, but I doubt if their absence is the most fundamental cause of the present confusion. In textbooks as elsewhere we tend to be more influenced by deeds than by words—in this case by what a textbook writer does after he has found a "nonsignificant" difference. What we find, not uncommonly, is an action *as if* "not significant" were equivalent to "nonexistent" or "unimportant." A writer may have clearly stated that "not significant" means merely "not proven," and then, when he comes to analyze a threefold comparison and finds that the outcomes after treatments A and B do not differ "significantly," he pools the A and B groups for comparison with the control (untreated) group.

If we do this kind of thing we are making an assumption that A and B do not differ in their effects, or at least are sufficiently alike for our purposes; but what is our risk if we act on that assumption and it is untrue? The 5 per cent error in a "significance" test tells us nothing about that risk. Further, why should we not make an equally plausible assumption that the effects of A and B differ, and by various amounts?[7] If we knew that they differed by these amounts, how would this affect our actions? It is regrettable that text-

books have been so silent on these matters.

Action after a verdict of "not significant." Although many people are becoming conscious of the limitations of "significance" tests, there is still too great a use of them in a mechanical fashion to obtain a definite "yes or no" answer, with a clear indication of what to do next. Whatever cut off point ("level of significance") we choose, and whatever verdict a test gives us, the use that we make of the verdict must depend on the purpose of our research, upon knowledge outside the particular data that we have tested, and upon various attendant circumstances.

Probably the nearest that we can come to a general rule regarding action after a "not significant" verdict applies only to what may be called "casual" or "incidental" observations. Outside any research, or merely incidental to it, we frequently see figures that suggest an association of variables, or the effect of some agent or other. The figures may appear rather impressive, but we may have no other reason to believe they are meaningful. If we can then show that random processes ("chance") would frequently cause such figures to occur, we can often save ourselves from following false clues. Our situation is quite different when we have obtained the "not significant" verdict as the result of a specific piece of research; it then requires much thought. A few of the kinds of things we must think about will be exemplified here.

Traumatic shock in dogs. This previously used[3] example is repeated here because it taught me a valuable lesson. During World War II there was presented, to a medical research committee, a report about dogs that had been experimentally injured by a certain standardized technique. After the trauma, 15 of the animals had been kept at an environmental temperature of 95° F., whereas 10 had been "cooled to an equivalent degree." Of the warmed animals, 11 (73 per cent) had died; of the cooled animals, only 4 (40 per cent) had died. The investigators seemed to be almost sure that the difference in temperature had caused the difference in the mortality rate. I quickly applied a chi square test (with Yates' correction) and found 1.56, which is far below the magic 3.841, the cut off point for $P = 0.05$. I felt superior, but said nothing and subsequently learned that the experiment, repeated on a much larger number of dogs, had shown convincing evidence of the benefit of cooling.

Since chi square $= 1.56$, $P = 0.2$ approximately. That is, if we accepted this value as indicative of something more than a random process, we would be adopting a standard that, if used consistently, would cause us to follow something like 20 per cent of false clues. The investigators accepted that risk, although apparently unaware of it, because they made no mention of testing their initial results. Disregard of such risks (i.e., acceptance of an unusually low standard of "significance") is legitimate provided that the reason is well defined. The actual reason here is unknown, but it might be one or more of the following:

1. The possession of some other knowledge of the physiologic effects of cooling that would agree with, or explain, the observed difference.
2. The ease with which a larger experiment could be performed.
3. The fact that if cooling did reduce mortality, it might be very important in treating shock in humans.
4. The large size of the difference, $73 - 40 = 33$ per cent. This apparently impressed the investigators, because they did not mention the sample sizes until they were requested. Such frequencies are often impressive to laboratory workers who are used to small percentage errors in measurement data.

In the subsequent experiments the investigators apparently continued until they had satisfied themselves that the difference was "real"; but nowadays, after obtaining a "nonsignificant" difference, if we wish to pursue the search we can set a definite

goal for ourselves. For example, after the study of the original 25 dogs we could say: "We wish to set up an experiment that will be unlikely to give us a 'nonsignificant' difference if the true (population) difference in mortality between heated and cooled dogs is as much as 25 percentage points." Having defined "unlikely" and "significant" in numerical terms, we could refer to a table[6] for an estimate of the number of animals required for the experiment. (Incidentally, it appears desirable to mention that if we included the original 25 dogs in the larger experiment, our final significance test would be a farce, for we would have broken the rules of the game. Our final samples would not be random, but in part selected by the inclusion of the first 25 dogs. Indeed, the larger samples might not have come into existence if the first 25 dogs had not shown such a large heated-versus-cooled difference in mortality.)

Clinical trials in rheumatoid arthritis. These drug trials, which are conducted by the Committee on Cooperating Clinics of the American Rheumatism Association,[5, 7] have presented many problems relevant to this discussion. Among them are sex differences and interclinic differences.

Sex differences. In drug trials on patients with rheumatoid arthritis it is not customary to "stratify" patients by sex because rheumatologists have learned that there is little, if any, sex difference in the behavior of the disease or in response to therapies that have been studied so far. Even if a sex difference in the behavior of the disease had been indisputably demonstrated, it would at present have little or no practical importance in the application of the results of a drug trial. If the trial had shown that drug A was preferable to drug B, it would make no difference whether the patients in the trial had been all females or all males, or a mixture of the two. Physicians who acted on the results of the trial would prescribe drug A to both sexes.

In conducting a drug trial in rheumatoid arthritis, therefore, we disregard sex, knowing that the randomization will control any bias due to this factor, in conjunction with other potential bias-causing factors, known and unknown. That is, we pool the sexes even if we feel almost certain that with larger numbers or more sensitive observational methods we would find a "statistically significant" sex difference in the frequency (or degree) of response to the A and B treatments.

Having analyzed data from the total group of patients we can, of course, make an A versus B comparison separately in each sex. Let us suppose that women showed a "significant" A-B difference and men did not. This supposition is quite realistic since there are usually more women than men in rheumatoid arthritis drug trials, apparently because in the general population the number of rheumatoid arthritic women who seek (orthodox) medical aid is greater than the number of men in the same category (a cautious statement, to avoid implications of sex differences in morbidity). We might indeed find that the A-B difference in men, though not "significant," was in the reverse direction from the difference in women, but even then, because of the medical background of knowledge described above, we would not conclude that the A-B difference in women was absent in men.

This does not imply that we should disregard an apparent sex difference simply because we are not going to make immediate use of it. If the same apparent difference occurred in one trial after another it could be the starting point for special researches into the sex difference in the behavior of the disease.

Let us now suppose that the drug under test was one that, from its biologic source, its chemical structure or its effects on animals might be expected to affect the sexes differently, or to a different degree. The proper procedure in the trial would be "stratification"—division of the original group by sex and random assignment of the new drug and the standard drug (or placebo) within each sex group separately.

Interclinic differences. In multiclinic trials, the smallness of each clinic's sample often makes it impossible to detect a "significant" interclinic variation in response to the same drug or in the interdrug differences. The analysts of the data from the American Rheumatism Association trials do not look for such variation. They take it for granted that it exists and could be detected if much larger samples were obtainable. One reason is that interobserver differences in the standards and techniques of assessment must be assumed, although detailed instructions are issued and great efforts at uniformity are made, including the exchange of visits by observers from the different clinics. Another reason is that even when patients meet the specifications laid down for admission to a trial, the groups available at different clinics may well differ in some relevant but unrecorded characteristic. Some of the implications of the presumed interclinic differences may now be noted.

Random assignment of drugs is performed for each clinic separately. Therefore, the appropriate analysis is a comparison of treatment groups within each clinic, followed by a combination of the results of these separate analyses. The final verdict is, of course, heavily influenced by the clinics which have provided the largest numbers of patients, and the question arises: "How can we safely argue from this collection of patients to the general population of rheumatoid arthritics that would meet the specifications of the protocol?" The question would be just as cogent if we had obtained the same number of patients from each clinic. Indeed, such a question is cogent in all medical research, whether conducted on patients or on laboratory animals. We cannot generalize to an outside population of patients or animals with quantitative assurance such as we can specify when we have taken a random sample from a barrel of well-shuffled disks.

This problem has been discussed at greater length elsewhere.[3] In medical research, as in many other affairs, we have to trust what may be called "the uniformity of Nature"—the frequent occurrence of similar events when conditions are similar. We have confidence in that principle because man has depended on it since prehistoric times. We cannot affix a probability "P" to our confidence or to our predictions, but we can help ourselves to avoid overconfidence by the use of our "barrel of disks" experience.

After a clinical trial, whether the A-B difference in outcome is "significant" or "not significant," we can say: "If these samples were strictly random samples from their respective (A and B) populations, what would they tell us about the true (population) difference in the effects of A and B?" We can classify as "unlikely" all population values that would make our observed samples "rare," according to our definition of "rarity"—5 per cent, 1 per cent, or whatever we choose. Such estimates are often called "confidence limits" but, as one of my students remarked, they might be more appropriately called "no-confidence limits." Perhaps a still better title would be "minimal estimates of ignorance." This sounds discouraging, but it forces us to recognize that no single study can give us very precise knowledge of the value of a therapy.

Observational error. Clinical investigators have in recent years paid increasing attention to observational error (variation and bias), e.g., in the questioning of patients, in auscultation of the heart, and in instrumental measurements such as blood pressure and hand-grip determined by mercury manometer. Observers have made careful, tedious, and time-consuming studies, and it is regrettable to see how in the analysis of their data some of them have been misled by textbook statistical arithmetic. For example, two observers make readings of a certain variable on the same group of patients, and each makes duplicate readings using the same pair of instruments. Then they apply a "*t* test" in order to find the "significance" of the inter-

observer and interinstrument differences. This procedure can perhaps best be described as trying to prove the existence of something that is already known to exist.

Even an instrument with a Bureau of Standards certificate can be shown to differ from a duplicate instrument likewise certified, if we test the two instruments by one of higher precision; and any two observers can be assumed to differ much more than two such instruments. Whether the *t* test gives a verdict of "significant" or of "not significant," it is not answering the right question, namely: "What is the importance of the interobserver and interinstrument differences in the investigation where these observers and these instruments are going to be used?" By "importance" we mean the amounts that they contribute to the total variation in the experiment and the risk of bias if their effects are not controlled, either systematically or by randomization. Incidentally, although special studies of observational error are useful to reveal orders of magnitude of effects, the setting of such studies may be artificial. Only under actual working conditions can reliable estimates be made.

The possible demise of "significance" tests. The last example showed that sometimes "significance" tests are not only unnecessary but inappropriate. In the discussion of interclinic differences it was seen that we make estimates of population values whether our experiment has produced a "significant" or a "nonsignificant" difference; and it may well be asked why we need to perform the significance test at all. In fact there are signs that mechanical "significance" testing, although far from moribund, is not so vigorous as it was a few years ago; and with its death the problem of "nonsignificance" would no longer plague us. It is to be hoped that it would not be replaced by some other misunderstood mechanical trick.

References

1. Fisher, R. A.: Statistical methods for research workers, Edinburgh and London, 1925, Oliver and Boyd.
2. Fisher, R. A.: The design of experiments, Edinburgh and London, 1935, Oliver and Boyd.
3. Mainland, D.: Elementary medical statistics, ed. 2, Philadelphia, 1963, W. B. Saunders Company.
4. Mainland, D.: The use and misuse of statistics in medical publications, CLIN. PHARMACOL. & THERAP. 1: 411-422, 1960.
5. Mainland, D.: Experiences in the development of multiclinic trials, J. New Drugs 1:197-205, 1961.
6. 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.
7. Mainland, D., and Sutcliffe, M. I.: Hydroxychloroquine sulfate in rheumatoid arthritis, a six month, double-blind trial, Bull. Rheumat. Dis. 13:287-290, 1962.
8. Pearson, K.: Early statistical papers, Cambridge, 1948, Cambridge University Press.