VOL. 66, No. 6 DECEMBER 1966

Psychological Bulletin

THE TEST OF SIGNIFICANCE IN PSYCHOLOGICAL RESEARCH

DAVID BAKAN

University of Chicago

The test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and a great deal of mischief has been associated with its use. The basic logic associated with the test of significance is reviewed. The null hypothesis is characteristically false under any circumstances. Publication practices foster the reporting of small effects in populations. Psychologists have "adjusted" by misinterpretation, taking the p value as a "measure," assuming that the test of significance provides automaticity of inference, and confusing the aggregate with the general. The difficulties are illuminated by bringing to bear the contributions from the decision-theory school on the Fisher approach. The Bayesian approach is suggested.

That which we might identify as the "crisis of psychology" is closely related to what Hogben (1958) has called the "crisis in statistical theory." The vast majority of investigations which pass for research in the field of psychology today entail the use of statistical tests of significance. Most characteristically, when a psychologist finds a problem he wishes to investigate he converts his intuitions and hypotheses into procedures which will yield a test of significance; and will characteristically allow the result of the test of significance to bear the essential responsibility for the conclusions which he will draw.

The major point of this paper is that the test of significance does not provide the information concerning psychological phenomena characteristically attributed to it; and that, furthermore, a great deal of mischief has been associated with its use. What will be said in this paper is hardly original. It is, in a certain sense, what "everybody knows." To say it "out loud" is, as it were, to assume the role of the child who pointed out that the emperor was really outfitted only in his underwear. Little of that which is contained in this paper is not already available in the literature, and the literature will be cited.

Lest what is being said in this paper be misunderstood, some clarification needs to be made at the outset. It is not a blanket criticism of statistics, mathematics, or, for that matter, even the test of significance when it can be appropriately used. The argument is rather that the test of significance has been carrying too much of the burden of scientific inference. Wise and ingenious investigators can find their way to reasonable conclusions from data because and in spite of their procedures. Too often, however, even wise and ingenious investigators, for varieties of reasons not the least of which are the editorial policies of our major psychological journals, which we will discuss below, tend to credit the test of significance with properties it does not have.

LOGIC OF THE TEST OF SIGNIFICANCE

The test of significance has as its aim obtaining information concerning a characteristic of a *population* which is itself not directly observable, whether for practical or more intrinsic reasons. What is observable is the *sample*. The work assigned to the test of significance is that of aiding in making inferences from the observed sample to the unobserved population.

The critical assumption involved in testing significance is that, if the experiment is conducted properly, *the characteristics of the population have a designably determinative*

influence on samples drawn from it, that, for example, the mean of a population has a determinative influence on the mean of a sample drawn from it. Thus if P, the population characteristic, has a determinative influence on S, the sample characteristic, then there is some license for making inferences from S to P.

If the determinative influence of P on S could be put in the form of simple logical *implication,* that P implies S, the problem would be quite simple. For, then we would have the simple situation: if P implies S, and if S is false, P is false. There are some limited instances in which this logic applies directly in sampling. For example, if the range of values in the population is between 3 and 9 (P), then the range of values in any sample must be between 3 and 9 (S). Should we find a value in a sample of, say, 10, it would mean that S is false; and we could assert that P is false.

It is clear from this, however, that, *strictly speaking,* one can only go from the denial of S to the denial of P; and not from the assertion of S to the assertion of P. It is within this context of simple logical implication that the Fisher school of statisticians have made important contributions—and it is extremely important to recognize this as the context.

In contrast, approaches based on the theorem of Bayes (Bakan, 1953, 1956; Edwards, Lindman, & Savage, 1963; Keynes, 1948; Savage, 1954; Schlaifer, 1959) would allow inferences to P from S even when S is not denied, as S adding something to the credibility of P when S is found to be the case. One of the most viable alternatives to the use of the test of significance involves the theorem of Bayes; and the paper by Edwards et al. (1963) is particularly directed to the attention of psychologists for use in psychological research.

The notion of the null hypothesis[1] promoted by Fisher constituted an advance *within this context* of simple logical implication. It allowed experimenters to set up a null hypothesis complementary to the hypothesis that the investigator was interested in, and provided him with a way of positively confirming his hypothesis. Thus, for example, the investigator might have the hypothesis that, say, normals differ from schizophrenics. He would then set up the *null hypothesis* that the means in the population of all normals and all schizophrenics were *equal.* Thus, the rejection of the null hypothesis constituted a way of *asserting* that the means of the populations of normals and schizophrenics *were different,* a completely reasonable device whereby to affirm a logical antecedent.

The model of simple logical implication for making inferences from S to P has another difficulty which the Fisher approach sought to overcome. This is that it is rarely meaningful to set up any simple "P implies S" model for parameters that we are interested in. In the case of the mean, for example, it is rather that P has a determinative influence on the *frequency* of any specific S. But one experiment does not provide many values of S to allow the study of their frequencies. It gives us *only one* value of S. The *sampling distribution* is conceived which specifies the relative frequencies of all possible values of S. Then, with the help of an adopted *level of significance,* we could, *in effect,* say that S was false; that is, any S which fell in a region whose relative theoretical frequency under the null hypothesis was, say, 5% would be *considered* false. If such an S actually occurred, we would be in a position to declare P to be false, still within the model of simple logical implication.

It is important to recognize that one of the essential features of the Fisher approach is what may be called the *once-ness* of the experiment; the inference model takes as critical that the experiment has been conducted *once.* If an S which has a low proba-

[1] There is some confusion in the literature concerning the meaning of the term null hypothesis. Fisher used the term to designate any exact hypothesis that we might be interested in disproving, and "null" was used in the sense of that which is to be nullified (cf., e.g., Berkson, 1942). It has, however, also been used to indicate a parameter of zero (cf., e.g., Lindquist, 1940, p. 15), that the difference between the population means is zero, or the correlation coefficient in the population is zero, the difference in proportions in the population is zero, etc. Since both meanings are usually intended in psychological research, it causes little difficulty.

bility under the null hypothesis actually occurs, it is taken that the null hypothesis is false. As Fisher (1947, p. 14) put it, why should the theoretically rare event under the null hypothesis actually occur to "us"? If it does occur, we take it that the null hypothesis is false. Basic is the idea that "the theoretically unusual does not happen to me."[2] It should be noted that the referent for all probability considerations is neither in the population itself nor the subjective confidence of the investigator. It is rather in a hypothetical population of experiments all conducted in the same manner, but *only one of which is actually conducted.* Thus, of course, the probability of falsely rejecting the null hypothesis if it were true is exactly that value which has been taken as the level of significance. Replication of the experiment vitiates the validity of the inference model, unless the replication itself is taken into account in the model and the probabilities of the model modified accordingly (as is done in various designs which entail replication, where, however, the total experiment, including the replications, is again considered as *one* experiment). According to Fisher (1947), "it is an essential characteristic of experimentation that it is carried out with limited resources [p. 18]." In the Fisher approach, the "limited resources" is not only a making of the best out of a limited situation, but is rather an integral feature of the inference model itself. Lest he be done a complete injustice, it should be pointed out that he did say, "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results [1947, p. 14]." However, although Fisher "himself" believes this, it is *not* built into the inference model.[3]

[2] I playfully once conducted the following "experiment": Suppose, I said, that every coin has associated with it a "spirit"; and suppose, furthermore, that if the spirit is implored properly, the coin will veer head or tail as one requests of the spirit. I thus invoked the spirit to make the coin fall head. I threw it once, it came up head. I did it again, it came up head again. I did this six times, and got six heads. Under the null hypothesis the probability of occurrence of six heads is $(\frac{1}{2})^6 = .016$, significant at the 2% level of significance. I have never repeated the experiment. But, then, the logic of the inference model does not really demand that I do! It may be objected that the coin, or my tossing, or even my observation was biased. But I submit that such things were in all likelihood not as involved in the result as corresponding things in most psychological research.

Difficulties of the Null Hypothesis

As already indicated, research workers in the field of psychology place a heavy burden on the test of significance. Let us consider some of the difficulties associated with the null hypothesis.

1. *The a priori reasons for believing that the null hypothesis is generally false anyway.* One of the common experiences of research workers is the very high frequency with which significant results are obtained with large samples. Some years ago, the author had occasion to run a number of tests of significance on a battery of tests collected on about 60,000 subjects from all over the United States. Every test came out significant. Dividing the cards by such arbitrary criteria as east versus west of the Mississippi River, Maine versus the rest of the country, North versus South, etc., all produced significant differences in means. In some instances, the differences in the sample means were quite small, but nonetheless, the p values were all very low. Nunnally (1960) has reported a similar experience involving correlation coefficients on 700 subjects. Joseph Berkson (1938) made the observation almost 30 years ago in connection with chi-square:

> I believe that an observant statistician who has had any considerable experience with applying the chi-square test repeatedly will agree with my statement that, as a matter of observation, when the numbers in the data are quite large, the P's tend to come out small. Having observed this, and on reflection, I make the following dogmatic statement, referring for illustration to the normal curve: "If

[3] Possibly not even this criterion is sound. It may be that a number of statistically significant results which are *borderline* "speak for the null hypothesis rather than against it [Edwards et al., 1963, p. 235]." If the null hypothesis were really false, then with an increase in the number of instances in which it can be rejected, there should be some substantial proportion of more dramatic rejections rather than borderline rejections.

the normal curve is fitted to a body of data representing any real observations whatever of quantities in the physical world, then if the number of observations is extremely large—for instance, on an order of 200,000—the chi-square *P* will be small beyond any usual limit of significance."

This dogmatic statement is made on the basis of an extrapolation of the observation referred to and can also be defended as a prediction from *a priori* considerations. For we may assume that it is practically certain that any series of real observations does not actually follow a normal curve *with absolute exactitude* in all respects, and no matter how small the discrepancy between the normal curve and the true curve of observations, the chi-square *P* will be small if the sample has a sufficiently large number of observations in it.

If this be so, then we have something here that is apt to trouble the conscience of a reflective statistician using the chi-square test. For I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the *P* that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all [pp. 526–527].

As one group of authors has put it, "in typical applications . . . the null hypothesis . . . is known by all concerned to be false from the outset [Edwards et al., 1963, p. 214]." The fact of the matter is that *there is really no good reason to expect the null hypothesis to be true in any population.* Why should the mean, say, of all scores east of the Mississippi be *identical* to all scores west of the Mississippi? Why should any correlation coefficient be *exactly* .00 in the population? Why should we expect the ratio of males to females be *exactly* 50:50 in any population? Or why should different drugs have *exactly* the same effect on any population parameter (Smith, 1960)? *A glance at any set of statistics on total populations will quickly confirm the rarity of the null hypothesis in nature.*

The reason why the null hypothesis is characteristically rejected with large samples was made patent by the theoretical work of Neyman and Pearson (1933). The probability of rejecting the null hypothesis is a function of five factors: whether the test is one- or two-tailed, the level of significance, the standard deviation, the amount of deviation from the null hypothesis, *and the number of observations.* The choice of a one- or two-tailed test is the investigator's; the level of significance is also based on the choice of the investigator; the standard deviation is a given of the situation, and is characteristically reasonably well estimated; the deviation from the null hypothesis is what is unknown; and the choice of the number of cases is in psychological work is characteristically arbitrary or expeditious. Should there be any deviation from the null hypothesis in the population, *no matter how small*—and we have little doubt but that such a deviation usually exists—a sufficiently large number of observations will lead to the rejection of the null hypothesis. As Nunnally (1960) put it,

if the null hypothesis is not rejected, it is usually because the *N* is too small. If enough data are gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data [p. 643].

2. *Type I error and publication practices.* The Type I error is the error of rejecting the null hypothesis when it is indeed true, and its probability is the level of significance. Later in this paper we will discuss the distinction between *sharp* and *loose* null hypotheses. The sharp null hypothesis, which we have been discussing, is an exact value for the null hypothesis as, for example, the difference between population means being precisely zero. A loose null hypothesis is one in which it is conceived of as being *around* null. Sharp null hypotheses, as we have indicated, rarely exist in nature. Assuming that loose null hypotheses are not rare, and that their testing may make sense under some circumstances, let us consider the role of the publication practices of our journals in their connection.

It is the practice of editors of our psychological journals, receiving many more papers than they can possibly publish, to use the magnitude of the *p* values reported as one criterion for acceptance or rejection of a study. For example, consider the following statement made by Arthur W. Melton (1962) on completing 12 years as editor of the *Journal of Experimental Psychology,* certainly one of the most prestigious and scientifically meticulous psychological journals. In enumerating the criteria by which articles were evaluated, he said:

> The next step in the assessment of an article involved a judgment with respect to the confidence to be placed in the findings—confidence that the results of the experiment would be repeatable under the conditions described. In editing the *Journal* there has been a strong reluctance to accept and publish results related to the principal concern of the research when those results were significant at the .05 level, whether by one- or two-tailed test. This has not implied a slavish worship of the .01 level, as some critics may have implied. Rather, it reflects a belief that it is the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounces [pp. 553–554].

His clearly expressed opinion that nonsignificant results should not take up the space of the journals is shared by most editors of psychological journals. It is important to point out that I am not advocating a change in policy in this connection. In the total research enterprise where so much of the load for making inferences concerning the nature of phenomena is carried by the test of significance, the editors can do little else. The point is rather that the situation in regard to publication makes manifest the difficulties in connection with the overemphasis on the test of significance as a principal basis for making inferences.

McNemar (1960) has rightly pointed out that not only do journal editors reject papers in which the results are not significant, but that papers in which significance has not been obtained are not submitted, that investigators select out their significant findings for inclusion in their reports, and that theory-oriented research workers tend to discard data which do not work to confirm their theories. The result of all of this is that "published results are more likely to involve false rejection of null hypotheses than indicated by the stated levels of significance [p. 300]," that is, published results which are significant may well have Type I errors in them far in excess of, say, the 5% which we may allow ourselves.

The suspicion that the Type I error may well be plaguing our literature is given confirmation in an analysis of articles published in the *Journal of Abnormal and Social Psychology* for one complete year (Cohen, 1962). Analyzing 70 studies in which significant results were obtained with respect to the power of the statistical tests used, Cohen found that power, the probability of rejecting the null hypothesis when the null hypothesis was false, was characteristically meager. Theoretically, with such tests, one should not often expect significant results even when the null hypothesis was false. Yet, there they were! Even if deviations from null existed in the relevant populations, the investigations were characteristically not powerful enough to have detected them. This strongly suggests that there is something additional associated with these rejections of the null hypotheses in question. It strongly points to the possibility that the manner in which studies get published is associated with the findings; that *the very publication practices themselves are part and parcel of the probabilistic processes on which we base our conclusions concerning the nature of psychological phenomena.* Our total research enterprise is, at least in part, a kind of scientific roulette, in which the "lucky," or constant player, "wins," that is, gets his paper or papers published. And certainly, going from 5% to 1% does not eliminate the possibility that it is "the way the ball bounces," to use Melton's phrase. It changes the odds in this roulette, but it does not make it less a game of roulette.

The damage to the scientific enterprise is compounded by the fact that the publication of "significant" results tends to stop further investigation. If the publication of papers containing Type I errors tended to foster further investigation so that the psychological phenomena with which we are concerned would be further probed by others, it would not be too bad. But it does not. Quite the contrary. As Lindquist (1940, p. 17) has correctly pointed out, the danger to science of the Type I error is much more serious than the Type II error—for when a Type I error is committed, it has the effect of stopping investigation. A highly significant result appears definitive, as Melton's comments indicate. In the 12 years that he edited the *Journal of Experimental Psychology,* he sought to select papers which were worthy of being placed in the "archives," as he put it. Even the strict repetition of an experiment and not getting significance in the same way does not speak against the result already re-

ported in the literature. For failing to get significance, speaking strictly within the inference model, only means that that experiment is inconclusive; whereas the study already reported in the literature, with a low *p* value, is regarded as conclusive. Thus we tend to place in the archives studies with a relatively high number of Type I errors, or, at any rate, studies which reflect small deviations from null in the respective populations; and we act in such a fashion as to reduce the likelihood of their correction.

Psychologist's "Adjustment" by Misinterpretation

The psychological literature is filled with misinterpretations of the nature of the test of significance. One may be tempted to attribute this to such things as lack of proper education, the simple fact that humans may err, and the prevailing tendency to take a cookbook approach in which the mathematical and philosophical framework out of which the tests of significance emerge are ignored; that, in other words, these misinterpretations are somehow the result of simple intellectual inadequacy on the part of psychologists. However, such an explanation is hardly tenable. Graduate schools are adamant with respect to statistical education. Any number of psychologists have taken out substantial amounts of time to equip themselves mathematically and philosophically. Psychologists as a group do a great deal of mutual criticism. Editorial reviews prior to publication are carried out with eminent conscientiousness. There is even a substantial literature devoted to various kinds of "misuse" of statistical procedures, to which not a little attention has been paid.

It is rather that the test of significance is profoundly interwoven with other strands of the psychological research enterprise in such a way that it constitutes a critical part of the total cultural-scientific tapestry. To pull out the strand of the test of significance would seem to make the whole tapestry fall apart. In the face of the intrinsic difficulties that the test of significance provides, we rather attempt to make an "adjustment" by attributing to the test of significance characteristics which it does not have, and overlook characteristics that it does have. The difficulty is that the test of significance can, especially when not considered too carefully, do *some* work; for, after all, the results of the test of significance *are* related to the phenomena in which we are interested. One may well ask whether we do not have here, perhaps, an instance of the phenomenon that learning under partial reinforcement is very highly resistant to extinction. Some of these misinterpretations are as follows:

1. *Taking the* p *value as a "measure" of significance.* A common misinterpretation of the test of significance is to regard it as a "measure" of significance. It is interpreted as the answer to the question "How significant is it?" A *p* value of .05 is thought of as less significant than a *p* value of .01, and so on. The characteristic practice on the part of psychologists is to compute, say, a *t*, and then "look up" the significance in the table, taking the *p* value as a *function of* t, and thereby a "measure" of significance. Indeed, since the *p* value is inversely related to the magnitude of, say, the difference between means *in the sample,* it can function as a kind of "standard score" measure for a variety of different experiments. Mathematically, the *t* is actually very similar to a "standard score," entailing a deviation in the numerator, and a function of the variation in the denominator; and the *p* value is a "function" of *t*. If this use were explicit, it would perhaps not be too bad. But it must be remembered that this is using the *p* value as a *statistic descriptive of the sample alone,* and does not automatically give an inference to the population. There is even the practice of using tests of significance in studies of total populations, in which the observations cannot by any stretch of the imagination be thought of as having been randomly selected from any designable population.[4] Using the *p* value in this way, in which the statistical inference model is even hinted at, is completely indefensible; for the single function of the statistical inference model is making inferences to populations from samples.

The practice of "looking up" the *p* value

[4] It was decided not to cite any specific studies to exemplify points such as this one. The reader will undoubtedly be able to supply them for himself.

for the t, which has even been advocated in some of our statistical handbooks (e.g., Lacey, 1953, p. 117; Underwood, Duncan, Taylor, & Cotton, 1954, p. 129), rather than looking up the t for a given p value, violates the inference model. The inference model is based on the presumption that one *initially* adopts a level of significance as the specification of that probability which is too slow to occur to "us," as Fisher has put it, in this one instance, and under the null hypothesis. A purist might speak of the "delicate problem . . . of fudging with a posteriori alpha values [levels of significance. Kaiser, 1960, p. 165]," as though the levels of significance were initially decided upon, but rarely do psychological research workers or editors take the level of significance as other than a "measure."

But taken as a "measure," it is only a measure of the sample. Psychologists often erroneously believe that the p value is "the probability that the results are due to chance," as Wilson (1961, p. 230) has pointed out; that a p value of .05 means that the chances are .95 that the scientific hypothesis is correct, as Bolles (1962) has pointed out; that it is a measure of the power to "predict" the behavior of a population (Underwood et al., 1954, p. 107); and that it is a measure of the "confidence that the results of the experiment would be repeatable under the conditions described," as Melton put it. Unfortunately, none of these interpretations are within the inference model of the test of significance. Some of our statistical handbooks have "allowed" misinterpretation. For example, in discussing the erroneous rhetoric associated with talking of the "probability" of a population parameter (in the inference model there is no probability associated with something which is either true or false), Lindquist (1940) said, "For most practical purposes, the end result is the same as if the 'level of confidence' type of interpretation is employed [p. 14]." Ferguson (1959) wrote, "The .05 and .01 probability levels are descriptive of our degree of confidence [p. 133]." There is little question but that sizable differences, correlations, etc., in *samples*, especially samples of reasonable size, speak more strongly of sizable differences, correlations, etc., in the population; and there is little question but that if there is real and strong effect in the population, it will continue to manifest itself in further sampling. However, these are inferences which *we* may make. They are outside the inference model associated with the test of significance. The p value within the inference model is only the value which we take to be as how improbable an event could be under the null hypothesis, which we judge will not take place to "us," in this one experiment. *It is not a "measure" of the goodness of the other inferences which we might make.* It is an a priori condition that we set up whereby we decide whether or not we will reject the null hypothesis, not a measure of significance.

There is a study in the literature (Rosenthal & Gaito, 1963) which points up sharply the lack of understanding on the part of psychologists of the meaning of the test of significance. The subjects were 9 members of the psychology department faculty, all holding doctoral degrees, and 10 graduate students, at the University of North Dakota; and there is little reason to believe that this group of psychologists was more or less sophisticated than any other. They were asked to rate their degree of belief or confidence in results of hypothetical studies for a variety of p values, and for n's of 10 and 100. That there should be a relationship between the average rated confidence or belief and p value, as they found, is to be expected. What is shocking is that these psychologists indicated substantially greater confidence or belief in results associated with the larger sample size for the same p values! According to the theory, especially as this has been amplified by Neyman and Pearson (1933), the probability of rejecting the null hypothesis for any given deviation from null and p value *increases* as a function of the number of observations. The rejection of the null hypothesis when the number of cases is small speaks for a more dramatic effect in the population; and if the p value is the same, the probability of committing a Type I error remains the same. Thus one can be more confident with a small n than a large n. The question is, how could a group of psychologists be so wrong? I believe that this wrongness is based on the commonly held

belief that the p value is a "measure" of degree of confidence. Thus, the reasoning behind such a wrong set of answers by these psychologists may well have been something like this: the p value is a measure of confidence; but a larger number of cases also increases confidence; therefore, for any given p value, the degree of confidence should be higher for the larger n. The wrong conclusion arises from the erroneous character of the first premise, and from the failure to recognize that the p value is a function of sample size for any given deviation from null in the population. The author knows of instances in which editors of very reputable psychological journals have rejected papers in which the p values and n's were small on the grounds that there were not enough observations, clearly demonstrating that the same mode of thought is operating in them. Indeed, rejecting the null hypothesis with a small n is indicative of a strong deviation from null in the population, the mathematics of the test of significance having already taken into account the smallness of the sample. Increasing the n increases the probability of rejecting the null hypothesis; and in these studies rejected for small sample size, that task has already been accomplished. These editors are, of course, in some sense the ultimate "teachers" of the profession; and they have been teaching something which is patently wrong!

2. *Automaticity of inference.* What may be considered to be a dream, fantasy, or ideal in the culture of psychology is that of achieving complete automaticity of inference. The making of inductive generalizations is always somewhat risky. In Fisher's *The Design of Experiments* (1947, p. 4), he made the claim that the methods of induction could be made rigorous, exemplified by the procedures which he was setting forth. This is indeed quite correct in the sense indicated earlier. In a later paper, he made explicit what was strongly hinted at in his earlier writing, that the methods which he proposed constituted a relatively *complete* specification of the process of induction:

> That such a process induction existed and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytic account can now be given, about as satisfying and complete, at least, as that given traditionally of the deductive processes [Fisher, 1955, p. 74].

Psychologists certainly took the procedures associated with the t test, F test, and so on, in this manner. *Instead* of having to engage in inference themselves, they had but to "run the tests" for the purpose of making inferences, since, as it appeared, the statistical tests were analytic analogues of inductive inference. The "operationist" orientation among psychologists, which recognized the contingency of knowledge on the knowledge-getting operations and advocated their specification, could, it would seem, "operationalize" the inferential processes simply by reporting the details of the statistical analysis! It thus removed the burden of responsibility, the chance of being wrong, the necessity for making inductive inferences, from the shoulders of the investigator and placed them on the tests of significance. The contingency of the conclusion upon the experimenter's decision of the level of significance was managed in two ways. The first, by resting on a kind of social agreement that 5% was good, and 1% better. The second in the manner which has already been discussed, by not making a decision of the level of significance, but only reporting the p value as a "result" and a presumably objective "measure" of degree of confidence. But that the probability of getting significance is also contingent upon the number of observations has been handled largely by ignoring it.

A crisis was experienced among psychologists when the matter of the one- versus the two-tailed test came into prominence; for here the contingency of the result of a test of significance on a decision of the investigator was simply too conspicuous to be ignored. An investigator, say, was interested in the difference between two groups on some measure. He collected his data, found that Mean A was greater than Mean B in the sample, and ran the ordinary two-tailed t test; and, let us say, it was not significant. Then he bethought himself. The two-tailed test tested against *two* alternatives, that the population Mean A was greater than population Mean B and vice versa. But then, he really wanted to know whether Mean A was greater than

Mean B. Thus, he could run a one-tailed test. He did this and found, since the one-tailed test is more powerful, that his difference was now significant.

Now here there was a difficulty. The test of significance is not nearly so automatic an inference process as had been thought. It is manifestly contingent on the decision of the investigator as to whether to run a one- or a two-tailed test. And somehow, making the decision *after* the data were collected and the means computed, seemed like "cheating." How should this be handled? Should there be some central registry in which one registers one's decision to run a one- or two-tailed test before collecting the data? Should one, as one eminent psychologist once suggested to me, send oneself a letter so that the postmark would prove that one had pre-decided to run a one-tailed test? The literature on ways of handling this difficulty has grown quite a bit in the strain to somehow overcome this particular clear contingency of the results of a test of significance on the decision of the investigator. The author will not attempt here to review this literature, except to cite one very competent paper which points up the intrinsic difficulty associated with this problem, the *reductio ad absurdum* to which one comes. Kaiser (1960), early in his paper, distinguished between the *logic* associated with the test of significance and other forms of inference, a distinction which, incidentally, Fisher would hardly have allowed: "The arguments developed in this paper are based on logical considerations in statistical inference. (We do not, of course, suggest that statistical inference is the only basis for scientific inference) [p. 160]." But then, having taken the position that he is going to follow the logic of statistical inference relentlessly, he said (Kaiser's italics): *"we cannot logically make a directional statistical decision or statement when the null hypothesis is rejected on the basis of the direction of the difference in the observed sample means* [p. 161]." One really needs to strike oneself in the head! If Sample Mean A is greater than Sample Mean B, and there is reason to reject the null hypothesis, in what other direction can it reasonably be? What kind of logic is it that leads one to believe that it could be otherwise than that Population Mean A is greater than Population Mean B? We do not know whether Kaiser intended his paper as a *reductio ad absurdum*, but it certainly turned out that way.

The issue of the one- versus the two-tailed test genuinely challenges the presumptive "objectivity" characteristically attributed to the test of significance. On the one hand, it makes patent what was the case under any circumstances (at the least in the choice of level of significance, and the choice of the number of cases in the sample), that the conclusion is contingent upon the decision of the investigator. An astute investigator, who foresaw the results, and who therefore pre-decided to use a one-tailed test, will get one p value. The less astute but honorable investigator, who did not foresee the results, would feel obliged to use a two-tailed test, and would get another p value. On the other hand, if one decides to be relentlessly logical within the logic of statistical inference, one winds up with the kind of absurdity which we have cited above.

3. *The confusion of induction to the aggregate with induction to the general.* Consider a not atypical investigation of the following sort: A group of, say, 20 normals and a group of, say, 20 schizophrenics are given a test. The tests are scored, and a t test is run, and it is found that the means differ significantly at some level of significance, say 1%. What inference can be drawn? As we have already indicated, the investigator could have insured this result by choosing a sufficiently large number of cases. Suppose we overlook this objection, which we can to some extent, by saying that the difference between the means in the population must have been *large enough* to have manifested itself with only 40 cases. But still, what do we know from this? The *only* inference which this allows is that the mean of all normals is different from the mean of all schizophrenics in the populations from which the samples have presumably been drawn at random. (Rarely is the criterion of randomness satisfied. But let us overlook this objection too.)

The common rhetoric in which such results are discussed is in the form "Schizophrenics

differ from normals in such and such ways." The sense that both the reader and the writer have of this rhetoric is that it has been justified by the finding of significance. Yet clearly it does not mean *all* schizophrenics and *all* normals. All that the test of significance justifies is that *measures of central tendency of the aggregates* differ in the populations. The test of significance has *not* addressed itself to anything about the schizophrenia or normality which characterizes *each* member of the respective populations. Now it is certainly possible for an investigator to develop a hypothesis about the nature of schizophrenia *from which he may infer* that there should be differences between the means in the populations; and his finding of a significant difference in the means of his sample would add to the credibility of the former. However, that 1% which he obtained in his study bears only on the means of the populations, and is not a "measure" of the confidence that he may have in his hypothesis concerning the nature of schizophrenia. There are *two* inferences that he must make. One is that of the sample to the population, for which the test of significance is of some use. The other is from his inference concerning the population to his hypothesis concerning the nature of schizophrenia. The p value does not bear on this second inference. The psychological literature is filled with assertions which confound these two inferential processes.

Or consider another hardly atypical style of research. Say an experimenter divides 40 subjects at random into two groups of 20 subjects each. One group is assigned to one condition and the other to another condition, perhaps, say, massing and distribution of trials. The subjects are given a learning task, one group under massed conditions, the other under distributed conditions. The experimenter runs a t test on the learning measure and again, say, finds that the difference is significant at the 1% level of significance. He may then say in his report, being more careful than the psychologist who was studying the difference between normals and schizophrenics (being more "scientific" than his clinically-interested colleague), that "the mean in the population of learning under massed conditions is lower than the mean in the population of learning under distributed conditions," feeling that he can say this with a good deal of certainty because of his test of significance. But here too (like his clinical colleague) he has made *two* inferences, and not one, and the 1% bears on the one but not the other. The statistical inference model certainly allows him to make his statement for the population, but only for *that* learning task, and the p value is appropriate only to that. But the generalization to "massed conditions" and "distributed conditions" beyond that particular learning task is a second inference with respect to which the p value is not relevant. The psychological literature is plagued with any number of instances in which the rhetoric indicates that the p value does bear on this second inference.

Part of the blame for this confusion can be ascribed to Fisher who, in *The Design of Experiments* (1947, p. 9), suggested that the mathematical methods which he proposed were exhaustive of scientific induction, and that the principles he was advancing were "common to all experimentation." What he failed to see and to say was that after an inference was made concerning a population parameter, *one still needed to engage in induction* to obtain meaningful scientific propositions.

To regard the methods of statistical inference as exhaustive of the inductive inferences called for in experimentation is completely confounding. When the test of significance has been run, the necessity for induction has hardly been completely satisfied. However, the research worker knows this, in some sense, and proceeds, as he should, to make further inductive inferences. He is, however, still ensnarled in his test of significance and the presumption that *it* is the whole of his inductive activity, and thus mistakenly takes a low p value for the measure of the validity of his *other* inductions.

The seriousness of this confusion may be seen by again referring back to the Rosenthal and Gaito (1963) study and the remark by Berkson which indicate that research workers believe that a large sample is better than a small sample. We need to refine the rhetoric somewhat. Induction consists in making in-

ferences from the particular to the general. It is certainly the case that as confirming particulars are added, the credibility of the general is increased. However, *the addition of observations to a sample is,* in the context of statistical inference, *not the addition of particulars* but the modification of what is *one particular* in the inference model, the sample aggregate. In the context of statistical inference, it is not necessarily true that "a large sample is better than a small sample." For, as has been already indicated, obtaining a significant result with a small sample suggests a larger deviation from null in the population, and may be considerably more meaningful. Thus more particulars are better than fewer particulars on the making of an inductive inference; but not necessarily a larger sample.

In the marriage of psychological research and statistical inference, psychology brought its own reasons for accepting this confusion, reasons which inhere in the history of psychology. Measurement psychology arises out of two radically different traditions, as has been pointed out by Guilford (1936, pp. 5 ff.) and Cronbach (1957), and the matter of putting them together raised certain difficulties. The one tradition seeks to find propositions concerning the nature of man in *general*—propositions of a general nature, with each *individual a particular* in which the general in manifest. This is the kind of psychology associated with the traditional experimental psychology of Fechner, Ebbinghaus, Wundt, and Titchener. It seeks to find the laws which characterize the "generalized, normal, human, adult mind [Boring, 1950, p. 413]." The research strategy associated with this kind of psychology is straightforwardly inductive. It seeks inductive generalizations which will apply to *every* member of a designated class. A single particular in which a generalization fails forces a rejection of the generalization, calling for either a redefinition of the class to which it applies or a modification of the generalization. The other tradition is the psychology of individual differences, which has its roots more in England and the United States than on the continent. We may recall that when the young American, James McKeen Cattell, who invented the term *mental test,* came to Wundt with his own problem of individual differences, it was regarded by Wundt as *ganz Amerikanisch* (Boring, 1950, p. 324).

The basic datum for an individual-differences approach is not anything that characterizes *each* of two subjects, but the *difference between them.* For this latter tradition, it is the *aggregate* which is of interest, and not the general. One of the most unfortunate characteristics of many studies in psychology, especially in experimental psychology, is that the data are treated as aggregates while the experimenter is trying to infer general propositions. There is hardly an issue of most of the major psychological journals reporting experimentation in which this confusion does not appear several times; and in which the test of significance, which has some value in connection with the study of aggregates, is not interpreted as a measure of the credibility of the general proposition in which the investigator is interested.

The distinction between the aggregate and the general may be illuminated by a small mathematical exercise. The methods of analysis of variance developed by Fisher and his school have become techniques of choice among psychologists. However, at root, the methods of analysis of variance do not deal with that which any two or more subjects may have in common, but consider only *differences between* scores. This is all that is analyzed by analysis of variance. The following identity illustrates this clearly, showing that the original total sum squares, of which everything else in any analysis of variance is simply the partitioning of, is based on the literal difference between each pair of scores (cf. Bakan, 1955). Except for n, it is the only information used from the data:

$$\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{2}\left[\frac{(X_1-X_2)}{1}\right]^2+\frac{2}{3}\left[\frac{(X_1-X_3)+(X_2-X_3)}{2}\right]^2+\cdots+\frac{n-1}{n}\left[\frac{(X_1-X_n)+\cdots+(X_{n-1}-X_n)}{n-1}\right]^2.$$

Thus, what took place historically in psychology is that instead of attempting to

synthesize the two traditional approaches to psychological phenomena, which is both possible and desirable, a syncretic combination took place of the methods appropriate to the study of aggregates with the aims of a psychology which sought for general propositions. One of the most overworked terms, which added not a little to the essential confusion, was the term "error," which was a kind of umbrella term for (at the least) variation among scores from different individuals, variation among measurements for the same individual, and variation among samples.

Let us add another historical note. In 1936, Guilford published his well-known *Psychometric Methods*. In this book, which became a kind of "bible" for many psychologists, he made a noble effort at a "Rapprochement of Psychophysical and Test Methods" (p. 9). He observed, quite properly, that mathematical developments in each of the two fields might be of value in the other, that "Both psychophysics and mental testing have rested upon the same fundamental statistical devices [p. 9]." There is no question of the truth of this. However, what he failed to emphasize sufficiently was that mathematics is so abstract that the same mathematics is applicable to rather different fields of investigation without there being any necessary further identity between them. (One would not, for example, argue that business and genetics are essentially the same because the same arithmetic is applicable to market research and in the investigation of the facts of heredity.) A critical point of contact between the two traditions was in connection with scaling in which Cattell's principle that "equally often noticed differences are equal unless always or never noticed [Guilford, 1936, p. 217]" was adopted as a fundamental assumption. The "equally often noticed differences" is, of course, based on aggregates. By means of this assumption, one could collapse the distinction between the two areas of investigation. Indeed, this is not really too bad if one is alert to the fact that *it is* an assumption, one which even has considerable pragmatic value. As a set of techniques whereby data could be analyzed, that is, as a set of techniques whereby one could *describe* one's findings, and then make inductions about the nature of the psychological phenomena, that which Guilford put together in his book was eminently valuable. However, around this time the work of Fisher and his school was coming to the attention of psychologists. It was attractive for several reasons. It offered advice for handling "small samples." It offered a number of eminently ingenious new ways of organizing and extracting information from data. It offered ways by which several variables could be analyzed simultaneously, away from the old notion that one had to keep everything constant and vary only one variable at a time. It showed how the effect of the "interaction" of variables could be assessed. But it also claimed to have mathematized induction! The Fisher approach was thus "bought," and psychologists got a theory of induction in the bargain, a theory which seemed to exhaust the inductive processes. Whereas the question of the "reliability" of statistics had been a matter of concern for some time before (although frequently very garbled), it had not carried the burden of induction to the degree that it did with the Fisher approach. With the "buying" of the Fisher approach the psychological research worker also bought, and then overused, the test of significance, employing it as the measure of the significance, in the largest sense of the word, of his research efforts.

Sharp and Loose Null Hypotheses

Earlier, a distinction was made between sharp and loose null hypotheses. One of the major difficulties associated with the Fisher approach is the problem presented by sharp null hypotheses; for, as we have already seen, there is reason to believe that the existence of sharp null hypotheses is characteristically unlikely. There have been some efforts to correct for this difficulty by proposing the use of loose null hypotheses; in place of a single point, a region being considered null. Hodges and Lehmann (1954) have proposed a distinction between "statistical significance," which entails the sharp hypothesis, and "material significance," in which one tests the hypothesis of a deviation of a stated amount from the null point instead

of the null point itself. Edwards (1950, pp. 30–31) has suggested the notion of "practical significance" in which one takes into account the meaning, in some practical sense, of the magnitude of the deviation from null together with the number of observations which have been involved in getting statistical significance. Binder (1963) has equally argued that a subset of parameters be equated with the null hypothesis. Essentially what has been suggested is that the investigator make some kind of a decision concerning "How much, say, of a difference makes a difference?" The difficulty with this solution, which is certainly a sound one technically, is that in psychological research we do not often have very good grounds for answering this question. This is partly due to the inadequacies of psychological measurement, but mostly due to the fact that the answer to the question of "How much of a difference makes a difference?" is not forthcoming outside of some particular practical context. The question calls forth another question, "How much of a difference makes a difference *for what*?"

Decisions versus Assertions

This brings us to one of the major issues within the field of statistics itself. The problems of the research psychologist do not generally lie within practical contexts. He is rather interested in making assertions concerning psychological functions which have a reasonable amount of credibility associated with them. He is more concerned with "What is the case?" than with "What is wise to do?" (cf. Rozeboom, 1960).

It is here that the decision-theory approach of Neyman, Pearson, and Wald (Neyman, 1937, 1957; Neyman & Pearson, 1933; Wald, 1939, 1950, 1955) becomes relevant. The decision-theory school, still basing itself on some basic notions of the Fisher approach, deviated from it in several respects:

1. In Fisher's inference model, the two alternatives between which one chose on the basis of an experiment were *reject* and *inconclusive*. As he said in *The Design of Experiments* (1947), "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation [p. 16]." In the decision-theory approach, the two alternatives are rather *reject* and *accept*.

2. Whereas in the Fisher approach the interpretation of the test of significance critically depends on having one sample from a *hypothetical* population of experiments, the decision-theory approach conceives of, is applicable to, and is sensible with respect to numerous repetitions of the experiment.

3. The decision-theory approach added the notions of the Type II error (which can be made only if the null hypothesis is accepted) and power as significant features of their model.

4. The decision-theory model gave a significant place to the matter of what is concretely lost if an error is made in the practical context, on the presumption that accept entailed one concrete action, and reject another. It is in these actions and their consequences that there is a basis for deciding on a level of confidence. The Fisher approach has little to say about the consequences.

As it has turned out, the field of application par excellence for the decision-theory approach has been the sampling inspection of mass-produced items. In sampling inspection, the acceptable deviation from null can be specified; both accept and reject are appropriate categories; the alternative courses of action can be clearly specified; there is a definite measure of loss for each possible action; and the choice can be regarded as one of a series of such choices, so that one can minimize the overall loss (cf. Barnard, 1954). Where the aim is only the acquisition of knowledge without regard to a specific practical context, these conditions do not often prevail. Many psychologists who learned about analysis of variance from books such as those by Snedecor (1946) found the examples involving log weights, etc., somewhat annoying. The decision-theory school makes it clear that such practical contexts are not only "examples" given for pedagogical purposes, but actually are essential features of the methods themselves.

The contributions of the decision-theory school essentially revealed the intrinsic nature of the test of significance beyond that seen by Fisher and his colleagues. They demonstrated that the methods associated with the

test of significance constitute not an assertion, or an induction, or a conclusion calculus, but a decision- or risk-evaluation calculus. Fisher (1955) has reacted to the decision-theory approach in polemic style, suggesting that its advocates were like "Russians [who] are made familiar with the ideal that research in pure science can and should be geared to technological performance, in the comprehensive organized effort of a five-year plan for the nation." He also suggested an American "ideological" orientation: "In the U. S. also the great importance of organized technology has I think made it easy to confuse the process appropriate for drawing correct conclusions, with those aimed rather at, let us say, speeding production, or saving money [p. 70]."[5] But perhaps a more reasonable way of looking at this is to regard the decision-theory school to have explicated what was already implicit in the work of the Fisher school.

Conclusion

What then is our alternative, if the test of significance is really of such limited appropriateness as has been indicated? At the very least it would appear that we would be much better off if we were to attempt to *estimate* the magnitude of the parameters in the populations; and recognize that we then need to make other inferences concerning the psychological phenomena which may be manifesting themselves in these magnitudes. In terms of a statistical approach which is an alternative, the various methods associated with the theorem of Bayes which was referred to earlier may be appropriate; and the paper by Edwards et al. (1963) and the book by Schlaifer (1959) are good starting points. However, that which is expressed in the theorem of Bayes alludes to the more general process of inducing propositions concerning the nonmanifest (which is what the population is a special instance of) and ascertaining the way in which that which is manifest (which the sample is a special instance of) bears on it. This is what the scientific method has been about for centuries. However, if the reader who might be sympathetic to the considerations set forth in this paper quickly goes out and reads some of the material on the Bayesian approach with the hope that thereby he will find a *new basis for automatic inference*, this paper will have misfired, and he will be disappointed.

That which we have indicated in this paper in connection with the test of significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research which may be, as the author has suggested elsewhere (Bakan, 1965), related to the presumption of the nonexistence of mind in the subjects of psychological research. Karl Pearson once indicated that higher statistics were only common sense reduced to numerical appreciation. However, that base in common sense must be maintained with vigilance. When we reach a point where our statistical procedures are substitutes instead of aids to thought, and we are led to absurdities, then we must return to the common sense basis. Tukey (1962) has very properly pointed out that statistical procedures may take our attention away from the data, which constitute the ultimate base for any inferences which we might make. Robert Schlaifer (1959, p. 654) has dubbed the error of the misapplication of statistical procedures the "error of the third kind," the most serious error which can be made. Berkson has suggested the use of "the interocular traumatic test, you know what the data mean when the conclusion hits you between the eyes [Edwards et al., 1963, p. 217]." We must overcome the myth that if our treatment of our subject matter is mathematical it is therefore precise and valid. Mathematics can serve to obscure as well as reveal.

Most importantly, we need to get on with the business of generating *psychological* hypotheses and proceed to do investigations and make inferences which bear on them; instead of, as so much of our literature would attest, testing the statistical null hypothesis in any number of contexts in which we have every reason to suppose that it is false in the first place.

[5] For a reply to Fisher, see Pearson (1955).

References

Bakan, D. Learning and the principle of inverse probability. *Psychological Review,* 1953, **60**, 360–370.

BAKAN, D. The general and the aggregate: A methodological distinction. *Perceptual and Motor Skills,* 1955, **5**, 211–212.

BAKAN, D. Clinical psychology and logic. *American Psychologist,* 1956, **11**, 655–662.

BAKAN, D. The mystery-mastery complex in contemporary psychology. *American Psychologist,* 1965, **20**, 186–191.

BARNARD, G. A. Sampling inspection and statistical decisions. *Journal of the Royal Statistical Society* (B), 1954, **16**, 151–165.

BERKSON, J. Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association,* 1938, **33**, 526–542.

BERKSON, J. Tests of significance considered as evidence. *Journal of the American Statistical Association,* 1942, **37**, 325–335.

BINDER, A. Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review,* 1963, **70**, 101–109.

BOLLES, R. C. The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports,* 1962, **11**, 639–645.

BORING, E. G. *A history of experimental psychology.* (2nd ed.) New York: Appleton-Century-Crofts, 1950.

COHEN, J. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 1962, **65**, 145–153.

CRONBACH, L. J. The two disciplines of scientific psychology. *American Psychologist,* 1957, **12**, 671–684.

EDWARDS, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.

EDWARDS, W., LINDMAN, H., & SAVAGE, L. J. Bayesian statistical inference for psychological research. *Psychological Review,* 1963, **70**, 193–242.

FERGUSON, L. *Statistical analysis in psychology and education.* New York: McGraw-Hill, 1959.

FISHER, R. A. *The design of experiments.* (4th ed.) Edinburgh: Oliver & Boyd, 1947.

FISHER, R. A. Statistical methods and scientific induction. *Journal of the Royal Statistical Society* (B), 1955, **17**, 69–78.

GUILFORD, J. P. *Psychometric methods.* New York: McGraw-Hill, 1936.

HODGES, J. L., & LEHMAN, E. L. Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society* (B), 1954, **16**, 261–268.

HOGBEN, L. *The relationship of probability, credibility and error: An examination of the contemporary crisis in statistical theory from a behaviourist viewpoint.* New York: Norton, 1958.

KAISER, H. F. Directional statistical decision. *Psychological Review,* 1960, **67**, 160–167.

KEYNES, J. M. *A treatise on probability.* London: Macmillan, 1948.

LACEY, O. L. *Statistical methods in experimentation.* New York: Macmillan, 1953.

LINDQUIST, E. F. *Statistical analysis in educational research.* Boston: Houghton Mifflin, 1940.

MCNEMAR, Q. At random: Sense and nonsense. *American Psychologist,* 1960, **15**, 295–300.

MELTON, A. W. Editorial. *Journal of Experimental Psychology,* 1962, **64**, 553–557.

NEYMAN, J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society* (A), 1937, **236**, 333–380.

NEYMAN, J. "Inductive behavior" as a basic concept of philosophy of science. *Review of the Mathematical Statistics Institute,* 1957, **25**, 7–22.

NEYMAN, J., & PEARSON, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society* (A), 1933, **231**, 289–337.

NUNNALLY, J. The place of statistics in psychology. *Education and Psychological Measurement,* 1960, **20**, 641–650.

PEARSON, E. S. Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society* (B), 1955, **17**, 204–207.

ROSENTHAL, R., & GAITO, J. The interpretation of levels of significance by psychological researchers. *Journal of Psychology,* 1963, **55**, 33–38.

ROZEBOOM, W. W. The fallacy of the null-hypothesis significance test. *Psychological Bulletin,* 1960, **57**, 416–428.

SAVAGE, L. J. *The foundations of statistics.* New York: Wiley, 1954.

SCHLAIFER, R. *Probability and statistics for business decisions.* New York: McGraw-Hill, 1959.

SMITH, C. A. B. Review of N. T. J. Bailey, *Statistical methods in biology. Applied Statistics,* 1960, **9**, 64–66.

SNEDECOR, G. W. *Statistical methods.* (4th ed.; orig. publ. 1937) Ames, Iowa: Iowa State College Press, 1946.

TUKEY, J. W. The future of data analysis. *Annals of Mathematical Statistics,* 1962, **33**, 1–67.

UNDERWOOD, B. J., DUNCAN, C. P., TAYLOR, J. A., & COTTON, J. W. *Elementary statistics.* New York: Appleton-Century-Crofts, 1954.

WALD, A. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics,* 1939, **10**, 299–326.

WALD, A. *Statistical decision functions.* New York: Wiley, 1950.

WALD, A. *Selected papers in statistics and probability.* New York: McGraw-Hill, 1955.

WILSON, K. V. Subjectivist statistics for the current crisis. *Contemporary Psychology,* 1961, **6**, 229–231.

(Received June 30, 1965)