PSYCHOMETRIKA—VOL. 15, NO. 2
JUNE, 1950

BOOK REVIEWS

QUINN MCNEMAR. *Psychological Statistics.* New York: John Wiley and Sons, 1949. Pp. vi + 364. $4.00.

Compared with other texts in the field of statistics, which attempt to cover the same material at a relatively simple level, this book is refreshing. It represents an honest effort to convey a true, workable understanding of statistics rather than simply providing a number of computational methods which can be applied. While teaching statistics to students of psychology will continue, I fear, to be a difficult problem, a text such as this, which stresses understanding of statistical techniques, their meaning and limitations, and the theory behind them, should help in getting across much that will be valuable to the student.

As one would find in any book on statistics written at this level, and written by one person, there are inadequacies here and there. Most people, I suspect, will find the last chapter on sampling and statistical inference incomplete, inexact, and relatively useless. The book might have been a better book without this last chapter which, unfortunately, introduces several confusions.

Other inadequacies are mostly of the character of omissions in explanation. I shall select a few examples of these to illustrate the type of thing involved:

(1) On page 64, in discussing the sampling distribution of differences between means, the author states, "It will be noticed that we can have a correlational term in formula (26) only when it is possible to pair, on some basis other than chance, the scores which enter into the first mean with the scores which contribute to the second mean." The formula involved is the standard deviation of a distribution of differences between means. The author manages successfully to convey the impression that the correlation term exists theoretically only if one can compute it in practice. This is, of course, not true. There are many instances where correlation can exist between the means, but this correlation cannot be estimated from the sample. The student should know that, even though he cannot estimate the correlation and assumes it to be zero, he may be in error in so doing.

(2) On page 98, in summarizing his conclusions after a discussion of rank order correlation, the author states "Rho is not as consistent from sample to sample as r, does not possess the mathematical advantages inherent in r, and therefore has merit only when the observations involve ranks, i.e., are not measures." It is somewhat surprising that the author, who is clearly sensitive to problems of statistical inference and testing, and who is concerned with limitations and inadequacies involved in making the necessary assumptions about the tests employed, can ignore the important role of rho in being amenable to exact tests of significance which do not involve assumptions concerning the parent population. This use of rank order correlation has been developed and used. It is too bad for a book of this caliber to dismiss an excellent technique so summarily.

(3) On page 166, discussing limitations of the product moment correlation, the author says, "The product moment correlation . . . . . needs careful qualifying if either or both variables yield skewed distributions." And in the next paragraph he says, "There are no general rules to follow in the case of variables yielding skewed distributions. Frequently, one can use a logarithmic transformation of such a variable and thereby secure scores which are at least approximately normal; or one may deliberately normalize the distribution by converting the raw scores into T scores." It is never made clear by the author just how or why the product moment correlation needs careful qualifying if we have skewed distributions, and it is certainly not made clear how we are helped by such a dubious procedure as normalizing the distributions involved. It is common knowledge that the calculation of a product moment correlation involves no assumptions whatsoever concerning the normality of the variables involved and that interpretation of the correlation coefficient can be made unequivocally in terms of a best fitting straight line irrespective of the distribution functions. It is true, of course, that to the extent that we do not have a normal correlation surface, we are at somewhat of a loss in testing significance of correlations. This latter point, however, is certainly not helped by deliberately normalizing the distribution. It is not clear what the author has in mind or why he makes the recommendation.

It should be stressed, however, that such points as these might not even be points of criticism were the book otherwise not on such a good level of explanation. I would like to devote the rest of this review to a discussion of two basic questions which this book brings up about the relationship between statistics and psychology.

The basic place of statistics in relation to any science or any body of data is as a tool or as a servant. Statistics does not dictate to the science, but rather seeks to adapt itself and to develop so as to help the science. In psychology today we are in danger of losing this nice relationship with statistics. Statistics has in many ways been oversold to psychologists and is rapidly becoming a tyrant, inflexible and nonadaptive. There are two major aspects of this trend toward statistical tyranny, both of which show themselves well in McNemar's book. One of these concerns the question of what is and what is not significant statistically, and the other concerns the question of the value placed upon high-powered statistical techniques in themselves, rather than in relation to the data on which they are used. We shall discuss these two points in order.

(1) *The level of significance to insist upon*:

Fifteen years ago statistics textbooks that were used in connection with psychology tended to make the simple and rather dogmatic statement that a critical ratio of 3.0 was needed in order for a difference to be regarded as significant. Later when R. A. Fisher made his considerable impact upon statistics and statistical thinking, this was modified in most textbooks to permit the investigator to also consider P values of .05 and .01 as evidence for regarding a difference as significant.

It has always been said that the particular P value upon which an investigator insists is an arbitrary matter. This, however, has not prevented dogma in relation to the selection of a P value. When the notion was introduced into statistics of two types of errors, namely, accepting the null hypothesis when false and rejecting the null hypothesis when true, the reasoning behind the

selection of what P value to insist upon became more sophisticated but the recommendations of writers still remained fairly dogmatic. In most cases the reasoning went something like this: There are two types of possible errors and the investigator must choose his P level according to the relative importance he places upon these two kinds of errors, but nevertheless one should insist upon a high level of confidence for rejecting the null hypothesis. This type of sophisticated *non sequitur* is very well illustrated in McNemar's book. After a very excellent discussion of the two types of errors, McNemar concludes on page 67:

> At the other extreme, a few are willing to accept as significant a difference which is 1.5 times its standard error. Since $P = .13$ for a CR of 1.5, it is readily seen that such persons would all too frequently have their publics believing that chance differences are real. A less lax level, which has had general acceptance by some workers, is represented by a P of .05, or a CR of nearly 2.0. This is also a rather low level of significance for announcing something as "fact." Those writers who advocate the .05 level for research workers in psychology, sociology, and education cite R. A. Fisher, the world's leading statistician, as their authority, *but* they fail to point out that Fisher's applications are to experimental situations wherein there is far better control of sampling than is ordinarily the case in the social sciences.

In short, after discussing the two types of errors, McNemar seems content to base his conclusions about P values on only one of those types of errors. He seems to regard only the rejection of the null hypothesis as "announcing something as fact" and does not seem to feel that accepting the null hypothesis is also announcing something as fact. If there is something in between rejecting the null hypothesis and accepting the null hypothesis, McNemar fails to point out what this third alternative is. He also fails to support the somewhat specious reasoning that a lower level of significance is appropriate to experimental situations.

McNemar also states on page 69, "If some reader must have a criterion regarding what is or is not significant, the author suggests that he compromise by taking the level indicated by a P of .01." Again in the midst of an excellent discussion concerning the kinds of errors of statistical inference that are involved, the conclusion is drawn that one must have a very high level of confidence for rejecting the null hypothesis. This, morever, does not express all of McNemar's opinions on the matter. On page 233, in discussing small sample methods, we find the following interesting statement: "It seems to us that those who publish statistical results based on a small number of cases should, unless they are positively sure that the basic assumptions underlying it have been met (and this assurance can seldom be attained), adopt a more stringent level of significance, says a P of .001 before drawing definite conclusions, and a P of .01 as the borderline of significance, i.e., as suggestive."

Let us examine whether such recommendations and such opinions are valid conclusions from the arguments which are and can be made, or whether these opinions and conclusions stem rather from prejudices. The tabulation below presents an example of how frequently the error of accepting the null hypothesis, with a specified true difference existing, would occur with different numbers of cases in the samples and specified P values for rejecting the null hypothesis. The example presupposes a true difference of $+5$ and a standard deviation of differences equal to 10.

Proportion of Time Null Hypothesis Would Be Accepted Where True Difference between Means is 5.0 and $\sigma = 10$

Number of Cases	$\sigma_{M_1-M_2}$	P Value Required for Rejecting Null Hypothesis .003	.01	.05	.13
9	3.33	.95	.84	.69	.50
25	2.00	.69	.50	.31	.16
49	1.43	.24	.08	.02	.002
100	1.00	.05	.01	.003	.0002

It can be seen that for small numbers of cases, if a true difference exists, there would be a tremendous number of errors committed by insisting upon a high level of confidence for rejecting the null hypothesis. Thus, in this example, with an n of 25, if we insisted upon the 5% level of confidence, 31% of the time we would accept the null hypothesis erroneously. Only with a relatively large number of cases can one set a high level of confidence for rejecting the null hypothesis and also be reasonably sure that he will not be committing a large number of errors of the other type.

If, then, in some instance both types of errors were equally important, our choice of level of significance for rejecting the null hypothesis would tend to be very different from what is customarily employed in psychology today, and radically different from what is recommended by McNemar. In fact, we would come to one conclusion exactly contradictory to McNemar. He suggests that the smaller the number of cases the higher the level of confidence we should insist on before rejecting the null hypothesis. From the example in our tabulation we would, however, tend to conclude that this would greatly magnify the number of errors of the other type which we would make. It is probably more reasonable to insist upon less confidence in rejecting the null hypothesis when the number of cases is small. The criteria for rejecting the null hypothesis can become stricter and stricter as our number of cases increase. In an *a priori* sense, one should not expect as much confidence from small samples as from large ones.

It is important to realize that a conclusion of no difference is frequently as important as concluding that there is a difference. Both of these types of conclusions can have equal implications for action. It is consequently important to choose a level of confidence which tends to minimize both kinds of errors rather than maximize one kind of error while minimizing another. Statistics must be looked at as a help in understanding and interpreting our data rather than as a determiner of specified standards of a statistical nature which should be met.

(2) *The self-perpetuating character of statistical techniques*:

It is unfortunate for psychology and for psychologists that very few of the commonly used statistical techniques have been developed in connection with the practical problems of analyzing psychological data. The result is that excellent statistical techniques whose assumptions are geared to conditions of data in other fields are now used in psychology. The attempt is made to force the design of psychological experiments into the required pattern which would then make them amenable to treatment by these statistical techniques. The question is rarely raised of the efficiency of these designs and of these statistical techniques for the problems involved in psychology. McNemar, in spite of his excellent treatment of the conventional statistical techniques, is a good example of

an author paying much attention to the requirements of the statistics and relatively little attention to the requirements of the data which the psychologist is interested in gathering. This omission is made even more glaring because the book is entitled *Psychological Statistics,* carrying the implication that it is for the psychologist.

Thus, in the book we find little attention being given to such techniques as rank order correlation, tests of significance based on rank orders and serial orders, and other such devices which can be of great help in many kinds of psychological data. Nowhere in the book can one find an attempt to teach the student how to use elementary probability theory in connection with his data. There is no elaboration of such simple devices as, for example, six independent differences in the same direction being significant at the 3% level of confidence, or a predicted order of five quantities being significant at the 1% level in terms of the possible permutations that could have occurred. Knowledge and use of such simple devices, however, is more important in many fields of psychology than complicated techniques such as analysis of covariance, to which the author does devote considerable time.

There is altogether no attempt to search the statistical literature for techniques and devices which are suitable for the peculiar conditions of psychological data. On the contrary, the book is merely an excellent presentation of statistical techniques which are best applicable to other fields than psychology. The discussions of the use of chi-square and of analysis of variance, for example, are the standard kinds of presentation for use where the experimental variations do not form a continuum. So frequently in psychology, experimental groups do fall along a graded continuum and nowhere in McNemar's book will one find the statement that chi-square and analysis of variance are inefficient techniques when the experimental groups are degrees of variation along a continuum since these techniques ignore the ordering of the results. It will have to be from books other than this one that psychologists will be able to learn to use statistics as a flexible tool rather than submit to statistics as a rigid master.

University of Michigan. LEON FESTINGER

On Festinger's Review of *Psychological Statistics*

Naturally I applaud certain parts of Festinger's review and I find suggestions therein that are worth filing for a possible revision of the book. But I also find points of fact and of philosophy with which I disagree. Of course I can say nothing to the nonparticularized criticism that the last chapter is "incomplete, inexact, and relatively useless," and that it "introduces several confusions."

A word on Festinger's third illustration of inadequacies. His implication that I think normality of distribution is assumed for r is controverted by my statement on p. 113, "r does not assume normal distributions." He also evidently missed my reasons, stated on the same page, for injecting a word of caution when nonnormal distributions are involved. My suggestion that distributions be normalized by transformation is said to be a "dubious procedure"—perhaps as a strong advocate of rank correlation the reviewer would prefer to rectangularize the distributions. He states that the testing of the significance of correlation "is certainly not helped by deliberately normalizing the distribution." Regarding this generalization, we refer the reader to the article and bibliography of Mueller (*Psychol. Bull.*, 1949, 46, 198-223).

Readers will have noted that Festinger uses a large portion of the review to expound his own ideas on statistics in psychology. He seems to be worried about statistics being "oversold" to psychologists, "becoming a tyrant, inflexible and nonadaptive," a "rigid master." I am indicted for aiding and abetting this "trend toward statistical tyranny."

In partial support of the above apprehensions, one-half the review is devoted to alleged dogma concerning what is significant statistically. Anyone who reads pages 66-69 and 232-234 may judge for himself whether the exposition is "dogmatic" and "prejudiced" or whether Festinger's discussion, and his use of quotations out of context, misrepresent these pages. Furthermore, there are points in his argument which demand comment.

When he says that "accepting the null hypothesis is also announcing something as fact," we ask what kind of fact emerges except the trivial fact that the null hypothesis couldn't be rejected. When he says that it is "important to choose a level of confidence which tends to minimize both kinds of errors," we ask whether there is a method for doing this. A search of the literature indicates that this problem has not been solved. (An outstanding mathematical statistician informs me that simultaneous minimization of the two types of errors is simply impossible.)

When the reviewer supposedly shows by a cleverly devised example that my expressed lack of faith in small samples is ill-founded, he unwittingly provides an excellent illustration of my claim that small samples are not conducive to rejection of the null hypothesis. Moreover, he does not, and for reasons which will soon be obvious, carry his own argument to its logical conclusion. Suppose that instead of a true difference of $+5$ it is presumed that the true difference is $+1$. Then with an N of 25 and the .05 level for significance, one would erroneously accept the null hypothesis 93% of the time. Now to overcome this sad state of affairs, let us follow the reviewer's suggestion of using a less stringent level of significance. A little arithmetic shows that even to reduce the 93% to 50% would entail using a critical ratio of only .5, or a P level of about .62, for rejecting the null hypothesis!

Being ever aware that statistical inference is a part of the logic of scientific method and being always wary of the use of any "flexible" logic, I regard Festinger's concluding sentence as a compliment. Submitting to statistics as a "rigid master" involves nothing more than submitting to the rigors of scientific method.

Stanford University QUINN MCNEMAR

BOOKS RECEIVED

ARLEY, NIELS, AND BUCH, K. RANDER. *Introduction to the Theory of Probability and Statistics.* New York: John Wiley and Sons, 1950.

COHN, ROBERT. *Clinical Electroencephalography.* New York: McGraw-Hill Book Co., 1949.

GOODENOUGH, FLORENCE L. *Mental Testing.* New York: Rinehart and Co., 1949.