EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
VOL. XX, NO. 4, 1960

THE PLACE OF STATISTICS IN PSYCHOLOGY

JUM NUNNALLY
University of Illinois

MOST psychologists probably will agree that the emphasis on statistical methods in psychology is a healthy sign. Although we sometimes substitute statistical elegance for good ideas and over-embellish small studies with elaborate analyses, we are probably on a firmer basis than we were in the prestatistical days. However, it will be argued that there are some serious misemphases in our use of statistical methods, which are retarding the growth of psychology.

The purpose of this article is to criticize the use of statistical "hypothesis-testing" models and some related concepts. It will be argued that the hypothesis-testing models have little to do with the actual testing of hypotheses and that the use of the models has encouraged some unhealthy attitudes toward research. Some alternative approaches will be suggested.

Few, if any, of the criticisms which will be made were originated by the author, and, taken separately, each is probably a well-smitten "straw man." However, it is hoped that when the criticisms are brought together they will argue persuasively for a change in viewpoint about statistical logic in psychology.

What is Wrong

Most will agree that science is mainly concerned with finding functional relations. A particular functional relationship may be studied either because it is interesting in its own right or because it helps clarify a theory. The functional relations most often sought in psychology are correlations between psychological variables, and differences in central tendency in differently treated groups of sub-

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

jects. Saying it in a simpler manner, psychological results are usually reported as correlation coefficients (or some extension thereof, such as factor analysis) and differences between means (or some elaboration, such as a complex analysis of variance treatment).

Hypothesis Testing. After an experiment is completed, and the correlations or differences between means have been obtained, the results must be interpreted. The experimenter is aware of sampling error and realizes that if the experiment is run on different groups of subjects the obtained relations will probably not be the same. How then should he take into account the chance element in the obtained relationship? In order to interpret the results, the experimenter would, as most of us have, rely on the statistical models for hypothesis testing. It will be argued that the hypothesis-testing models are inappropriate for nearly all psychological studies.

Statistical hypothesis testing is a decision theory: you have one or more alternative courses of action, and the theory leads to the choice of one or several of these over the others. Although the theory is very useful in some practical circumstances (such as in "quality control"), it is misnamed. It has very little to do with hypothesis testing in the way that hypotheses are tested in the work-a-day world of scientific activity.

The most misused and misconceived hypothesis-testing model employed in psychology is referred to as the "null-hypothesis" model. Stating it crudely, one null hypothesis would be that two treatments do not produce different mean effects in the long run. Using the obtained means and sample estimates of "population" variances, probability statements can be made about the acceptance or rejection of the null hypothesis. Similar null hypotheses are applied to correlations, complex experimental designs, factor-analytic results, and most all experimental results.

Although from a mathematical point of view the null-hypothesis models are internally neat, they share a crippling flaw: in the real world the null hypothesis is almost never true, and it is usually nonsensical to perform an experiment with the *sole* aim of rejecting the null hypothesis. This is a personal point of view, and it cannot be proved directly. However, it is supported both by common sense and by practical experience. The common-sense argument is that different psychological treatments will almost always (in the long run) produce differences in mean effects, even though the differences

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

may be very small. Also, just as nature abhors a vacuum, it probably abhors zero correlations between variables.

Experience shows that when large numbers of subjects are used in studies, nearly all comparisons of means are "significantly" different and all correlations are "significantly" different from zero. The author once had occasion to use 700 subjects in a study of public opinion. After a factor analysis of the results, the factors were correlated with individual-difference variables such as amount of education, age, income, sex, and others. In looking at the results I was happy to find so many "significant" correlations (under the null-hypothesis model)—indeed, nearly all correlations were significant, including ones that made little sense. Of course, with an N of 700 correlations as large as .08 are "beyond the .05 level." Many of the "significant" correlations were of no theoretical or practical importance.

The point of view taken here is that if the null hypothesis is not rejected, it usually is because the N is too small. If enough data is 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.

The arguments above apply most straightforwardly to "two-tail tests," which are used in most experiments. A somewhat better argument can be made for using the null hypothesis in the one-tail test. However, even in that case, if rejection of the null hypothesis is not obtained for the specified direction, the hypothesis can be reversed and rejection will usually occur.

Perhaps my intuitions are wrong—perhaps there are many cases in which different treatments produce the same effects and many cases in which correlations are exactly zero. Even so, the emphasis on the null-hypothesis models is unfortunate. As is well recognized, the mere rejection of a null hypothesis provides only meager information. For example, to say that a correlation is "significantly" different from zero provides almost no information about the relationship. Some would argue that finding "significance" is only the first step, but how many psychologists ever go beyond this first step?

Psychologists are usually not interested in finding tiny relationships. However, once this is admitted, it forces either a modification or an abandonment of the null-hypothesis model.

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

An alternative to the null hypothesis is the "fixed-increment" hypothesis. In this model, the experimenter must state in advance how much of a difference is an important difference. The model could be used, for example, to test the differential effect of two methods of teaching psychology, in which an achievement test is used to measure the amount of learning. Suppose that the regular method of instruction obtains a mean achievement test score of 45. In the alternative method of instruction, laboratory sessions are used in addition to lectures. The experimenter states that he will consider the alternative method of instruction better if, in the long run, it produces a mean achievement test score which is at least ten points greater than the regular method of instruction. Suppose that the alternative method actually produces a mean achievement test score of 65. The probability can then be determined as to whether the range of scores from 55 upwards covers the "true" value (the parameter).

The difficulty with the "fixed-increment" hypothesis-testing model is that there are very few experiments in which the increment can be stated in advance. In the example above, if the desired statistical confidence could not be found for a ten point increment, the experimenter would probably try a nine point increment, then an eight point increment, and so on. Then the experimenter is no longer operating with a hypothesis-testing model. He has switched to a *confidence-interval* model, which will be discussed later in the article.

The Small N Fallacy. Closely related to the null hypothesis is the notion that only enough subjects need be used in psychological experiments to obtain "significant" results. This often encourages experimenters to be content with very imprecise estimates of effects. In those situations where the dispersions of responses are small, only a small number of subjects is required. However, such situations are seldom encountered in psychology. The question, "When is the N large enough?" will be discussed later in the article.

Even if the object in experimental studies were to test the null hypothesis, the statistical test is often compromised by the small N. The tests depend on assumptions like homogeneity of variance, and the small N study is not sufficient to say how well the assumptions hold. The small N experiment, coupled with the null hypothesis, is

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

usually an illogical effort to leap beyond the confines of limited data to document lawful relations in human behavior.

The Sampling Fallacy. In psychological experiments we speak of the group of subjects as a "sample" and use statistical sampling theory to assess the results. Of course, we are seldom interested only in the particular group of subjects, and it is reasonable to question the generality of the results in wider collections of people. However, we should not take the sampling notion too seriously, because in many studies no sampling is done. In many studies we are content to use any humans available. College freshmen are preferred, but in a pinch we will use our wives, secretaries, janitors, and anyone else who will participate. We should then be a bit cautious in applying a statistical sampling theory, which holds only when individuals are randomly or systematically drawn from a defined population.

The Crucial Experiment. Related to the misconceptions above are some misconceptions about crucial experiments. Before the points are argued, a distinction should be made between crucial designs and crucial sets of data. A crucial design is an agreed-on experimental procedure for testing a theoretical statement. Even if the design is accepted as crucial, a particular set of data obtained with the design may not be accepted as crucial.

Although crucial designs have played important parts in some areas of science, few of them are, as yet, available in psychology. In psychology it is more often the case that experimenters propose different designs for testing the same theoretical statement. Experimental designs that apparently differ in small ways often produce different relationships. However, this is not a serious bother. Antithetical results should lead to more comprehensive theory.

A more serious concern is whether particular sets of experimental data can be regarded as crucial. Even when different psychologists employ the same design they often obtain different relationships. Such inconsistencies are often explained by "sampling error," but this is not a complete explanation. Even when the *N's* are large, it is sometimes reported that Jones finds a positive correlation, Smith a negative correlation, and Brown a nil correlation. The results of psychological studies are sometimes particular to the ex-

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

perimenter and the time and place of the experiment. This is why most psychologists would place more faith in the results of two studies, each with 50 subjects, performed by different investigators in different places, than in the results obtained by one investigator for 100 subjects. Then we must be concerned not only with the sampling of people but with the sampling of experimental environments as well. The need to "sample" experimental environments is much greater in some types of studies than in others. For example, the need probably would be greater in group dynamic studies than in studies of depth perception.

What Should Be Done

Estimation. Hypotheses are really tested by a process of *estimation* rather than with statistical hypothesis-testing models. That is, the experimenter wants to determine what the mean differences are, how large the correlation is, what form the curve takes, and what kinds of factors occur in test scores. If, in the long run, substantial differences are found between effects or if substantial correlations are found, the experimenter can then speak of the theoretical and practical implications.

To illustrate our dependence on estimation, analysis of variance should be considered primarily an estimation device. The variances and ratios of variances obtained from the analysis are unbiased estimates of different effects and their interactions. The proper questions to ask are, "How large are the separate variances?" and "How much of the total variance is explained by particular classifications?" Only as a minor question should we ask whether or not the separate sources of variance are such as to reject the null hypothesis. Of course, if the results fail to reject the null hypothesis, they should not be interpreted further; but if the hypothesis is rejected, this should be considered only the beginning of the analysis.

Once it is realized that the basis for testing psychological hypotheses is that of estimation, other issues are clarified. For example, the Gordian-knot can be cut on the controversial issue of "proving" the null hypothesis. If, in the long run, it is found that the means of two differently treated groups differ inconsequentially, there is nothing wrong with believing the results as they stand.

Confidence Intervals. It is not always necessary to use a large N, and there are ways of telling when enough data has been gath-

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

ered to have faith in statistical estimates. Most of the statistics which are used (means, variances, correlations, and others) have known distributions, and, from these, confidence intervals can be derived for particular estimates. For example, if the estimate of a correlation is .50, a confidence interval can be set for the inclusion of the "true" value. It might be found in this way that the probability is .99 that the "true" value[1] is at least as high as .30. This would supply a great deal more information than to reject the null hypothesis only.

The statistical hypothesis-testing models differ in a subtle, but important, way from the confidence methods. The former make decisions for the experimenter on an all-or-none basis. The latter tell the experimenter how much faith he can place in his estimates, and they indicate how much the N needs to be increased to raise the precision of estimates by particular amounts.

The null-hypothesis model occurs as a special case of the confidence models. If, for example, in a correlational study the confidence interval covers zero, then, in effect, the null hypothesis is not rejected. When this occurs it usually means that not enough data has been gathered to answer the questions at issue.

Discriminatory Power. In conjunction with making estimates and using confidence methods with those estimates, methods are needed for demonstrating the strength of relationships. In correlational studies, this need is served by the correlations themselves. In measuring differences in central tendency for differently treated groups, no strength-of-relationship measure is generally used.

One measure that is sometimes used is obtained by converting mean differences for two groups into a point-biserial correlation. This is easily done by giving the members of one group a "group score" of 1 and the members of the other group a "group score" of 2 (any other two numbers would serve the purpose). The dichotomous "group scores" are then correlated with the dependent variable. When the N is large, it is an eye-opener to learn what small correlations correspond to "highly significant" differences.

There is a general strength-of-relationship measure that can be applied to all comparisons of mean differences. The statistic is Epsilon, which was derived by Kelley (1935) and extended by

[1] Technically, it would be more correct to say that the probability is .99 that the range from .30 to 1.00 covers the parameter.

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

Peters and Van Voorhis (1940). The latter showed how Epsilon applies to analysis of variance methods and recommended its use in general. Their advice was not followed, and the suggestion here is that we reconsider Epsilon.

Epsilon is an unbiased estimate of the correlation ratio, Eta. It is unbiased because "degrees of freedom" are employed in the variance estimates. To show how Epsilon is applied, consider the one-classification analysis of variance results shown in Table 1.

TABLE 1

Hypothetical Results Illustrating the Use of Epsilon

Source	Sums of squares	df	Variance Est.
Experimental treatments (between column means)	510	4	127.50
Within columns	490	119	4.12
Total	1000	123	8.13

$$(\text{Epsilon})^2 = 1 - \frac{\text{Within var.}}{\text{Total var.}}$$

$$= 1 - \frac{4.12}{8.13}$$

$$= .49$$

$$\text{Epsilon} = .70$$

Epsilon is obtained by dividing the error variance (in the example in Table 1, the within columns variance) by the total variance, subtracting that from one, and taking the square-root of the result. The one classification in Table 1 explains 49 per cent of the total variance, which shows that the classification has high discriminatory power. Of course, in this case, the null hypothesis would have been rejected, but that is not nearly as important as it is to show that the classification produces strong differences.

Whereas Epsilon was applied in Table 1 to the simplest analysis of variance design, it applies equally well to complex designs. Each classification produces an Epsilon, which shows directly the discriminatory power of each (See Peters and Van Voorhis, 1940).

Epsilon is simply a general measure of correlation. If levels within a classification are ordered on a quantitative scale and regressions are linear, Epsilon reduces to the familiar r.

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

A Point of View. Statisticians are not to blame for the misconceptions in psychology about the use of statistical methods. They have warned us about the use of the hypothesis-testing models and the related concepts. In particular they have criticized the null-hypothesis model and have recommended alternative procedures similar to those recommended here (See Savage, 1957; Tukey, 1954; and Yates, 1951).

People are complicated, and it is hard to find principles of human behavior. Consequently, psychological research is often difficult and frustrating, and the frustration can lead to a "flight into statistics." With some, this takes the form of a preoccupation with statistics to the point of divorcement from the headaches of empirical study. With others, the hypothesis-testing models provide a quick and easy way of finding "significant differences" and an attendant sense of satisfaction.

The emphasis that has been placed on the null hypothesis and its companion concepts is probably due in part to the professional milieu of psychologists. The "reprint race" in our universities induces us to publish hastily-done, small studies and to be content with inexact estimates of relationships.

There is a definite place for small N studies in psychology. A chain of small studies, each elaborating and modifying the hypotheses and procedures, can eventually lead to a good understanding of a domain of behavior. However, if such small studies are taken out of context and considered (or published) separately, they usually are of little value, even if null hypotheses are successfully rejected.

Psychology had a proud beginning, and it would be a pity to see it settle for the meager efforts which are encouraged by the use of the hypothesis-testing models. The original purpose was to find lawful relations in human behavior. We should not feel proud when we see the psychologist smile and say "the correlation is significant beyond the .01 level." Perhaps that is the most that he can say, but he has no reason to smile.

REFERENCES

Kelley, T. L. "An Unbiased Correlation Ratio Measure." *Proceedings of the National Academy of Science,* Washington, XXI (1935), 554–559.

Peters, C. C. and Van Voorhis, W. R. *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill, 1940.

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014

Savage, R. J. "Nonparametric Statistics." *Journal of the American Statistical Association,* LII (1957), 332–333.

Tukey, J. W. "Unsolved Problems of Experimental Statistics." *Journal of the American Statistical Association,* XLIX (1954), 710.

Yates, F. "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics." *Journal of the American Statistical Association,* XLVI (1951), 32–33.

Downloaded from epm.sagepub.com at UNIV OF WESTERN ONTARIO on June 2, 2014