Reports from the Research Laboratories

of the

Department of Psychiatry

University of Minnesota

"Statistical Significance in Psychiatric Research"

by

DAVID T. LYKKEN

Report Number PR-66-9

STATISTICAL SIGNIFICANCE IN PSYCHIATRIC RESEARCH

by

David T. Lykken

December 30, 1966

PR-66-9

ABSTRACT

Most theories at issue in the areas of personality, clinical and social psychology predict no more than the direction of a correlation, group difference or treatment effect. Since the null hypothesis is never strictly true, such predictions have about a 50-50 chance of being confirmed by experiment when the theory in question is false, the statistical significance of the result being a function of the sample size. Therefore, while failure to confirm a single theoretical prediction may have great evidential weight, confirmation of a single directional prediction adds negligibly to the prior probability of a theory. Analysis of the notion of replication shows that not even the empirical "facts" alleged to be demonstrated by such experiments are credible in proportion to the statistical significance of the findings. It is argued that, before hurrying into print, most theories should be tested by multiple corroboration and most empirical generalizations by constructive replication. Statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for claiming that a theory has been usefully corroborated, that a meaningful empirical fact has been established, or that an experimental report ought to be published.

STATISTICAL SIGNIFICANCE IN PSYCHIATRIC RESEARCH

David T. Lykken

University of Minnesota

In a recent article in an American Psychological Association journal, Sapolsky (1964) developed the following substantive theory. Some psychiatric patients entertain an unconscious belief in the "cloacal theory of birth" which involves the notions of oral impregnation and anal parturition. Such patients should be inclined to manifest eating disorders: compulsive eating in the case of those who wish to get pregnant and anorexia for those who do not. In the Rorschach situation, such patients should also be inclined to see cloacal animals, such as frogs. This reasoning leads Sapolsky to predict that Rorschach frog-responders should show a higher incidence of eating disorders than patients not giving frog responses. A test of this hypothesis in a psychiatric hospital showed that 19 of 31 frog-responders had eating disorders indicated in their charts, compared to only 5 of the 31 control patients, which gives a highly significant Chi-square.

Since I regarded the prior probability of Sapolsky's theory (that frog responders unconsciously believe in impregnation per os) to be nugatory and its likelihood unenhanced by the experimental findings, I undertook to check my own reaction against that of 20 colleagues, most of them clinicians, by means of a formal questionnaire. The 20 estimates of the prior probability of Sapolsky's theory, which these psychologists made before being informed of his experimental results, ranged from 10^{-6} to 0.13 with a median value of 0.01, which can be interpreted to mean, roughly, "I don't believe it." Since the prior probability of many important scientific theories is considered to be vanishingly small when they are first propounded, this result provides no basis for alarm. However, after being given a fair summary of Sapolsky's experimental findings, which "corroborate" the theory by confirming

the operational hypothesis derived from it with high statistical significance, these same psychologists attached posterior probabilities to the theory which ranged from 10^{-5} to 0.14, with the median unchanged at 0.01; I interpret this consensus to mean, roughly, "I still don't believe it." This finding, I submit, is alarming because it signifies a sharp difference of opinion between, e.g., the consulting editors of the journal and a substantial segment of its readership, a difference on the very fundamental question of what constitutes good (i.e., publishable) clinical research.

The thesis of the present paper is that Sapolsky and the editors were in fact following, with reasonable consistency, our traditional rules for evaluating psychological research but that, as the Sapolsky paper exemplifies, these rules suffer from at least two important defects. One of the rules to be examined here asserts roughly the following: "When a prediction or hypothesis derived from a theory is confirmed by experiment, a non-trivial increment in one's confidence in that theory should result, especially when one's prior confidence is low." Clearly, my 20 colleagues were violating this rule here since their confidence in the frog responder cloacal birth theory was not, on the average, increased by the contemplation of Sapolsky's highly significant Chi-square. Judging from their comments, they found it too hard to accept that a belief in oral impregnation could lead to frog responding merely because the frog has a cloacus. (One must, after all, admit that few patients know what a cloacus is or that a frog has one and that those few who do know will also be inclined to know that the frog's eggs are both fertilized and hatched externally so neither oral impregnation nor anal birth are in any way involved. Hence, neither the average patient nor the biologically sophisticated patient should logically be expected to employ the frog as a symbol for an unconscious belief in oral conception.) My colleagues, on the contrary, found it relatively easy to believe that the

observed association between frog responding and eating problems might be due to some other cause entirely (e.g., both symptoms are immature or regressive in character; the frog, with its disproportionately large mouth and voice may well constitute a common orality totem and hence be associated with problems in the oral sphere; "squemish" people might tend both to see frogs and to have eating problems; and so on.)

Assuming that this first rule is wrong in this instance, perhaps it could be amended so as to allow one to make exceptions in cases resembling this illustration. For example, one could add the codicil, "This rule may be ignored whenever one considers the theory in question to be overly improbable or whenever one can think of alternative explanations for the experimental results." But surely such an amendment would not do. ESP, for example, could never become scientifically respectable if the first exception were allowed, and one consequence of the second would be that the importance attached to one's findings would always be inversely related to the ingenuity of one's readers. The burden of the present argument is that this rule is wrong, not only in a few exceptional instances, but as it is routinely applied to the majority of experimental reports in the psychological literature.

Corroborating Theories by Experimental Confirmation of Theoretical Predictions[1]

Most psychological experiments are of three kinds: (1) studies of the effect of some treatment on some output variables; which can be regarded as a special case of (2), studies of the difference between two or more groups of individuals with respect to some variable; which in turn are a special

1. Much of the argument in this section is based upon ideas developed in certain unpublished memoranda by P. E. Meehl, circulated privately in 1963.

case of (3), the study of the relationship or correlation between two or more variables within some specified population. Using the bivariate correlation design as paradigmatic, then, one notes first that the strict null hypothesis must always be assumed to be false. Unless one of the variables is wholly unreliable (so that the values obtained are strictly random), it would be foolish to suppose that the correlation between any two variables is identically equal to 0.0000... (or that the effect of some treatment or the difference between two groups is exactly zero). The molar dependent variables employed in psychological research are extremely complicated in the sense that the measured value of such a variable tends to be affected by the interaction of a vast number of factors, both in the present situation and in the history of the subject organism. It is exceedingly unlikely that any two such variables will not share at least some of these factors and equally unlikely that their effects will exactly cancel one another out.

It might be argued that the more complex the variables the smaller their average correlation ought to be since a larger pool of common factors allows more chance for mutual cancellation of effects in obedience to the Law of Large Numbers. However, one knows of a number of unusually potent and pervasive factors which operate to unbalance such convenient symmetries and to produce correlations large enough to rival the effects of whatever causal factors the experimentor may have had in mind. Thus, we know: (1) that "good" psychological and physical variables tend to be positively correlated; (2) that experimentors, without deliberate intention, can somehow subtly bias their findings in the expected direction (Rosenthal, 1963); (3) that the effects of common method are often as strong or stronger than those produced by the actual variables of interest (for example, in a large and careful study of the factorial structure of adjustment to stress among

officer candidates, Holtzman & Bitterman (1956) found that their 101 original variables contained five main common factors representing, respectively, their rating scales, their perceptual-motor tests, the McKinney Reporting Test, their GSR variables, and the MMPI); (4) that transitory state variables such as the S's anxiety level, fatigue, or his desire to please, may broadly affect all measures obtained in a single experimental session.

This average shared variance of "unrelated" variables can be thought of as a kind of ambient noise level characteristic of the domain. It would be interesting to obtain empirical estimates of this quantity in our field to serve as a kind of Plimsoll Mark against which to compare obtained relationships predicted by some theory under test. If, as I think, it is not unreasonable to suppose that "unrelated" molar psychological variables share on the average about 4 to 5 percent of common variance, then the expected correlation between any such variables would be about .20 in absolute value and the expected difference between any two groups on some such variable would be nearly 0.5 standard deviation units. (Note that these estimates assume zero measurement error. One can better explain the near-zero correlations often observed in psychological research in terms of unreliability of measures than by assuming that the true scores are in fact unrelated.)

Suppose now that an investigator predicts that two variables are positively correlated. Since we expect the null hypothesis to be false, we expect his prediction to be confirmed by experiment with a probability of very nearly 0.5; by using a large enough sample, moreover, he can achieve any desired level of statistical significance for this result. If the ambient noise level for his domain is represented by correlations averaging, say, 0.20 in absolute value, then his chances of finding a statistically significant confirmation of his prediction with a reasonable sample size will be

quite high (e.g., about 1 in 4 for N = 100) even if there is no truth whatever to the theory on which the prediction was based. Since most theoretical predictions in psychology, especially in the areas of clinical and personality research, specify no more than the direction of a correlation, difference or treatment effect, we must accept the harsh conclusion that a single experimental finding of this usual kind (confirming a directional prediction), no matter how great its statistical significance, will seldom represent a large enough increment of corroboration for the theory from which it was derived to merit very serious scientific attention. (In the natural sciences, this problem is far less severe for two reasons: (1) theories are powerful enough to generate point predictions or at least predictions of some narrow range within which the dependent variable is expected to lie, and (2) in these sciences, the degree of experimental control and the relative simplicity of the variables studied are such that the ambient noise level represented by unexplained and unexpected correlations, differences and treatment effects is often vanishingly small.)

The Significance of Large Correlations

It might be argued that, even where only a weak directional prediction is made, obtaining a result which is not only statistically significant but large in absolute value should constitute a stronger corroboration of the theory. For example, although Sapolsky predicted only that frog responding and eating disorders would be positively related, the tetrachoric correlation between these variables in his sample was nearly +.70, surely much larger than the average relationship expected between random pairs of molar variables on the premise that "everything is related to everything else." Does not such a large effect therefore provide stronger corroboration for the theory in question?

One difficulty with this reasonable sounding doctrine is that, in the complex sort of research considered here, really large effects, differences or relationships are not usually to be expected and, when found, may even argue against the theory being tested. To illustrate this, let us take Sapolsky's theory seriously and, by making reasonable guesses concerning the unknown base rates involved, attempt to estimate the actual size of the relationship between frog responding and eating disorder which the theory should lead us to expect. Sapolsky found that 16 percent of his control sample showed eating disorders; let us take this value as the base rate for this symptom among patients who do not hold the cloacal theory of birth. Perhaps we can assume that all patients who do hold this theory will give frog responses but surely not all of these will show eating disorders (any more than will all patients who believe in vaginal conception be inclined to show coital or urinary disturbances!); it seems a reasonable assumption that no more than 50 percent of the believers in oral conception will therefore manifest eating problems. Similarly, we can hardly suppose that the frog response always implies an unconscious belief in the cloacal theory; surely this response can come to be emitted now and then for other reasons. Even with the greatest sympathy for Sapolsky's point of view, we could hardly expect more than, say, 50 percent of frog responders to believe in oral impregnation. Therefore, we might reasonably predict that 16 of 100 non-responders would show eating disorders in a test of this theory, 50 of 100 frog-responders would hold the cloacal theory and half of these show eating disorders, while 16 percent or 8 of the remaining 50 frog-responders will show eating problems too, giving a total of 33 eating disorders among the 100 frog-responders. Such a finding would produce a significant Chi-square but the actual degree of relationship as indexed by the tetrachoric coefficient

would only be about +.35. In other words, if one considers the supplementary assumptions which would be required to make a theory compatible with the actual results obtained, it becomes apparent that the finding of a really strong association may actually embarrass the theory rather than support it.

Multiple Corroboration

In the social, clinical and personality areas especially, we must expect that the size of the correlations, differences or effects which might reasonably be predicted from our theories will typically not be very large relative to the ambient noise level of correlations and effects due solely to the "all-of-a-pieceness of things." The conclusion seems inescapable that the only really satisfactory solution to the problem of corroborating such theories is that of multiple corroboration, the derivation and testing of a number of separate, quasi-independent predictions. Since the prior probability of such a multiple corroboration is about equal to $(0.5)^n$, where n is the number of independent predictions experimentally confirmed, a theory of any useful degree of predictive richness should in principle allow for sufficient empirical confirmation through multiple corroboration to compel the respect of the most critical reader or editor.

The Relation of Experimental Findings to Empirical Facts

We turn now to the examination of a second popular rule for the evaluation of psychological research, which states roughly that, "When no obvious errors of sampling or experimental method are apparent, one's confidence in the general proposition being tested (e.g., Variables A and B are positively correlated in Population C) should be proportional to the degree of statistical significance obtained." We are following this rule when we say,

"Theory aside, Sapolsky has at least demonstrated an empirical fact, vis., that frog responders have more eating disturbances than patients in general." This conclusion means, of course, that in the light of Sapolsky's highly significant findings we should be willing to give very generous odds that any other competent investigator (at another hospital, administering the Rorschach in his own way and determining the presence of eating problems in whatever manner seems reasonable and convenient for him) will also find a substantial positive relationship between these two variables.

Let us be more specific. The 99 percent confidence interval for the tetrachoric coefficient of +.70 obtained from Sapolsky's four-fold table ranges from about +.40 to +1.00 and a similar interval centered on +.40 ranges from +.10 to +.70. With 99 percent confidence that the population value is at least +.40, we should have .99 (99) = 98 percent confidence that a new sample from that population should produce a tetrachoric no smaller than +.10; that is, we should be willing to bet $98 against only $1 that the replication will yield a tetrachoric of at least +.10. The reader may decide for himself whether his faith in the "empirical fact" demonstrated by this experiment can meet the test of this gambler's challenge.

Three Kinds of Replication

If, as suggested above, "demonstrating an empirical fact" must involve a claim of confidence in the replicability of one's findings, then to clearly understand the relation of statistical significance to the probability of a 'successful' replication it will be helpful to distinguish between three rather different methods of replicating or cross-validating an experiment. Literal replication, of course, would involve exact duplication of the first investigator's sampling procedure, experimental conditions, measuring

techniques and methods of analysis; asking the original investigator to simply run an additional N Ss would perhaps be about as close as we could come to attaining literal replication and even this, in psychological research, might often not be close enough. In the case of operational replication, on the other hand, one strives to duplicate exactly just the sampling and experimental procedures given in the first author's report of his research. The purpose of operational replication is to test whether the investigator's 'experimental recipe' -- the conditions and procedures he considered salient enough to be listed in the Methods section of his report -- will in other hands produce the results that he obtained. For example, replication of the 'Clever Hans' experiment revealed that the apparent ability of that remarkable horse to add numbers had been due to an uncontrolled and unsuspected factor (the presence of the horse's trainer within his field of view). This factor, not being specified in the "methods recipe" for the result, was omitted in the replication which for that reason failed. Operational replication would be facilitated if investigators would accept more responsibility for specifying what they believe to be the minimum essential conditions and controls for producing their results. Psychologists tend to be inconsistently prolix in describing their experimental methods; thus, Sapolsky tabulates the age, sex and diagnosis for each of his 62 Ss. Does he mean to imply that the experiment won't work if these details are changed? -- surely not, but then why describe them?

In the quite different process of constructive replication, one deliberately avoids imitation of the first author's methods. To obtain an ideal constructive replication, one would provide a competent investigator with nothing more than a clear statement of the empirical 'fact' which the first author would claim to have established -- e.g., "psychiatric patients who give frog responses on the Rorschach have a greater tendency toward eating

disorders than do patients in general" -- and then let the replicator formulate his own methods of sampling, measurement and data analysis. One must keep in mind that the data, the specific results of a particular experiment, are only seldom of any real interest in themselves. The "empirical facts" which we value so highly consist usually of confirmed conceptual or constructive (not operational) hypotheses of the form "Construct A is positively related to Construct B in Population C." We are interested in the construct "tendency toward eating disorders," not in the datum "has reference made to over-eating in the nurse's notes for May 15th." An operational replication tests whether we can duplicate our findings using the same methods of measurement and sampling; a constructive replication goes further in the sense of testing the validity of these methods. Thus, if I cannot confirm Sapolsky's results for patients from my hospital, assessing eating disorders by means of informant interviews say, or actual measurements of food intake, then clearly Sapolsky has not demonstrated any "fact" about eating disorders among psychiatric patients in general. I could then revert to an operational replication, assessing eating problems from the psychiatric notes as Sapolsky did and selecting my sample to conform with the age, sex and diagnostic properties of his, although I might not regard this endeavor to be worth the effort since, under these circumstances, even a successful operational replication could not establish an empirical conclusion of any great generality or interest. Just as a reliable but invalid test can be said to measure something, but not what it claimed to measure, so an experiment which replicates operationally but not constructively could be said to have demonstrated something, but not the relation between meaningful constructs, generalizable to some broad reference population, which the author originally claimed to have established.

Relation of the Significance Test to the Probability of a 'Successful' Replication

The probability values resulting from significance testing can be directly used to measure one's confidence in expecting a 'successful' literal replication only. Thus, we can be "98 percent confident" of finding a tetrachoric greater than +.10 in repeating Sapolsky's experiment only if we reproduce all of the conditions of his experiment with absolute fidelity, something that he himself could not undertake to do at this point. Whether we are entitled to anything approaching such high confidence that we could obtain such a result from an operational replication depends entirely upon whether Sapolsky has accurately specified all of the conditions which were in fact determinative of his results. That he did not in this instance is suggested by the fact that, investigating the feasibility of replicating his experiment at the University of Minnesota Hospitals, I found that I should have to review several thousand case records in order to turn up a sample of 31 frog-responders like his. Although he does not indicate how many records he examined, one strongly suspects that the base rate of Rorschach frog-responding must have been higher at Sapolsky's hospital, either because of some difference in the patient population or, more probably, because an investigator's being interested in some class of responses will tend to subtly elicit such responses at a higher rate unless the testing procedure is very rigorously controlled. If the base rates for frog-responding are so different at the two hospitals, it seems doubtful that the response can have the same correlates or meaning in the two populations and therefore one would be reckless indeed to offer high odds on the outcome of even the most careful operational replication. The likelihood of a 'successful' con-

structive replication is, of course, still smaller since it depends on the additional assumptions that Sapolsky's samples were truly representative of psychiatric patients in general and that his method of assessing eating problems was truly valid, i.e., would correlate highly with a different, equally reasonable appearing method.

Conclusions

The moral of this story is that the finding of statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition (and perhaps not always even a necessary condition) for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence--or that an experimental report ought to be published. The value of any research can be determined, not from the statistical results, but only by skilled, subjective evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on. Ideally, all experiments would be replicated before publication but this goal is impractical. "Good" experiments will tend to replicate better than poor ones (and, when they do not, the failures will tend to be informative in themselves, which is not true for poor experiments) and should be published so that they may stimulate replication and extension by others. Editors must be bold enough to take responsibility for deciding which studies are good and which are not, without resorting to letting the p-value of the significance tests determine this decision. There is little real danger that anything of value will be lost through this approach since the unpublished investigator can always resort to constructive replication to induce editorial

acceptance of his empirical conclusions or to multiple corroboration to compel editorial respect for his theory. Since operational replication must really be done by an independent second investigator and since constructive replication has greater generality, its success strongly implying that an operational replication would have succeeded also, one should usually replicate one's own work constructively, using different sampling and measurement procedures within the purview of the same constructive hypothesis. Only unusually well done, provocative and important research should be published without such prior authentication, and operational replication of such research by others will become correspondingly more valuable and entitled to the respect now accorded capable replication in the other experimental sciences.

References

Holtzman, W. H. & Bitterman, M. E. A factorial study of adjustment to stress. J. abn. soc. Psychol., 1956, 52, 179-185.

Rosenthal, R. On the social psychology of the psychological experiment: the experimentor's hypothesis as unintended determinant of experimental results. American Scientist, 1963, 51, 268-283.

Sapolsky, A. An effort at studying Rorschach content symbolism: the frog response. J. consult. Psychol., 1964, 28, 469-472.