Psychological Review
1960, Vol. 67, No. 4, 269–271

THEORETICAL NOTES

THE CONCEPT OF STATISTICAL SIGNIFICANCE AND THE CONTROVERSY ABOUT ONE-TAILED TESTS

H. J. EYSENCK
University of London

Several controversial papers regarding the uses and abuses of the one-tailed test of significance have recently appeared (Burke 1953, 1954; Goldfried 1959; Hick 1953; Jones 1952, 1954; Marks 1951, 1953). As Goldfried (1959) points out, "the important question debated is not *if* it should be used, but rather *when* it should be used." It is suggested here that most of the disagreements emerging from this controversy stem from a misunderstanding of the term "significance," and it is further suggested that the same misunderstanding runs through many discussions of two-tailed tests as well. It will be suggested that in the sense in which Goldfried's statement is meant, it has been the wrong question which has been debated; neither one-tailed nor two-tailed tests should be used at all in the sense envisaged by most of the writers quoted.

The outcome of the statistical examination of experimental results is always stated in terms of the probability of disconfirmation of the null hypothesis; the set of values which these p values can take is continuous in the interval from 0 to 1. It is customary to take arbitrary p values, such as .05 and .01, and use them to dichotomize this continuum into a *significant* and an *insignificant* portion. This habit has no obvious advantage, if what is intended is merely a restatement of the probability values; these are already given in any case and are far more precise than a simple dichotomous statement. Indeed, gross absurdities result from taking these verbal statements too seriously; the difference between a C.R. of 1.90 and another of 2.00 is quite negligible, yet one falls on one side of the dichotomy, the other on the other side. This has led to such summary statements as: "almost significant," or "significant at the 10% level." If the verbal dichotomous scale is not satisfactory—as it clearly is not—the answer surely is to keep to the continuous p scale, rather than subdivide the verbal scale.

However, surplus meaning has accrued to the word "significant," and it has become a shibboleth which divides the successful from the unsuccessful research. It is frequently interpreted as almost guaranteeing reproducibility of results, while failure to reach significance is interpreted as disconfirmation. Hence the urgent desire to achieve respectability and significance by one-tailed tests, if need be, and the argument regarding when the cachet of "significance" can be bestowed upon a research result. Yet the argument, and the achievement or nonachievement of significance, do not alter the facts of the case, which are contained in the statement of the p value of the results. Anything beyond these facts depends upon interpretation, and is subjective; it does not alter the facts of the case in the slightest.

As an example of the necessity of interpretation, consider the a priori probability of the conclusion. Suppose that an experiment on ESP were carried out with all the precautions which human ingenuity can devise, so that even the most sceptical had to agree that no fault could be found with the experimental design. *Suppose* also that a p value of .05 were achieved. Would this be considered "significant," in the sense of guaranteeing reproducibility? Critics would point out quite rightly that where the a priori probability is very low, as in this case, much higher p values would be required to carry significance. Logicians are agreed that interpretation of experimental results must call on all available knowledge about the subject in question; a priori probability is a kind of summary statement of much of this knowledge. It cannot be overlooked in arriving at a conclusion regarding "significance" when the term carries the surplus meaning indicated.

That interpretation comes into the problem very much is clear when we look at such conditions as those suggested by Kimmel (1957) as criteria for the use of one-tailed tests. He suggests, for instance, that they may be used if results in the opposite direction would be psychologically meaningless or could not be deduced from any psychological theory. These are obviously not objective criteria, but depend on what the author (or reader) considers psychologically meaningless, or the kind of theory he may hold. Opinions will differ, and consequently some readers will agree to the use of the one-tailed test in a particular case, others will not. Thus to some readers the results will appear *significant,* to others *insignificant.*

The whole argument seems to be about *words,* not about *facts:* Is the word "significant" to be used in a given situation, or is it not? This would only matter if the word carried some objective meaning not contained in the probability figures; we have argued that it does carry surplus meaning, but that this is not of an objective kind. Consequently, nothing important is changed by omitting the term altogether in the report, leaving interpretation to the reader. After all, the only true proof of reproducibility is reproduction! Verbal assertions of "significance" have no more meaning than the *droit du pour* at the court of Louis XIV.

The solution is to separate quite clearly and decisively the *objective statement of the probability of disproof* of the *null hypothesis* (by means of a two-tailed test), and the *subjective evaluation and interpretation of the results.* The reader would be able to accept the first statement as a statement of fact and would then be able to judge for himself the arguments presented by the author regarding the *meaning* of these facts. These arguments might be based on results of previous experiments, predictions made on the basis of more or less widely accepted theories, number of cases involved, a priori lack of accepability of the conclusions, and other similar grounds; an explicit statement of the arguments would enable the reader to decide for himself the acceptability of the conclusions in a manner precluded by the simple statement of one-tailed probability. *A statement of one-tail probability is not a statement of fact, but of opinion, and should not be offered instead of, but only in addition to, the factual two-tailed probability;* if it is offered at all, it should be accompanied by a full statement of the arguments in favor of its facilitating a more meaningful interpretation of the data. In the writer's opinion, it would be better to drop such statements of one-tailed probability altogether and rely entirely on appropriate argumentation to establish the meaning of the observed (two-tailed) probabilities.

Implicit in this recommendation is the corollary that the mechanical evaluation of experimental results in terms of "significant" and "not significant" be dropped outright. Interpretation is implicit in the statement of one-tailed probabilities, but it is also implicit in the statement of two-tailed probabilities if these are *automatically* interpreted as being significant or not significant, with all the surplus meaning carried by these terms. The experimenter should give his (two-tailed) p values and then proceed to argue regarding the acceptability of the conclusions on the basis already indicated. There have appeared in the literature solemn discussions about the possible causes for discrepancies between two experiments, one of which gave significant, the other insignificant results; yet the respective t values were almost identical, one lying just on the one side, the other just on the other side, of the arbitrary 5% line. Such arguments are unrealistic and would be avoided if p values were compared, rather than verbal statements. Two experiments giving p values of .048

and .056 are in excellent agreement, although one is significant, while the other is not.

To summarize the main point of this note briefly, we would say that verbal statements regarding "significance" are at best supererogatory restatements in an inconvenient dichotomous form of results already properly stated in terms of a continuous system of *p* values; at worst they carry unjustified surplus meaning of an entirely subjective kind under the guise of an objective and mathematically meaningful statement. Subjective judgments of reproducibility cannot reasonably be based on the mechanical application of a rule of thumb whose only usefulness lies in the elementary instruction of undergraduates lacking in mathematical background; if they are to be made at all they demand complex consideration of a priori probabilities. It is suggested that the accurate and factual statement of probabilities (two-tailed) should be mandatory and that all subjective considerations, arguments, and judgments should be clearly separated from such factual statements. It is implied that judgments of "significance" belong with the subjective side, and it is also implied that the calculation of *p* values on the basis of one-tailed tests has no place in psychology.

REFERENCES

BURKE, C. J. A brief note on one-tailed tests. *Psychol. Bull.*, 1953, **50**, 384-387.

BURKE, C. J. Further remarks on one-tailed tests. *Psychol. Bull.*, 1954, **51**, 587-590.

GOLDFRIED, M. R. One tailed tests and "unexpected" results. *Psychol. Rev.*, 1959, **66**, 79-80.

HICK, W. E. A note on one-tailed and two-tailed tests. *Psychol. Rev.*, 1952, **59**, 316-318.

JONES, L. V. Tests of hypotheses: One-sided and two-sided alternatives. *Psychol. Bull.*, 1952, **49**, 43-46.

JONES, L. V. A rejoinder on one-tailed tests. *Psychol. Bull.*, 1954, **51**, 585-586.

KIMMEL, H. D. Three criteria for the use of one-tailed tests. *Psychol. Bull.*, 1957, **54**, 351-353.

MARKS, M. R. Two kinds of experiment distinguished in terms of statistical operations. *Psychol. Bull.*, 1951, **58**, 179-184.

MARKS, M. R. One- and two-tailed tests. *Psychol. Rev.*, 1953, **60**, 203-208.

(Received August 28, 1959)