Sterling, T. D. (1959): Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance – Or Vice Versa. – J. Am. Stat. Assoc., 54(285):30-34, doi:10.2307/2282137.

PUBLICATION DECISIONS AND THEIR POSSIBLE EFFECTS ON INFERENCES DRAWN FROM TESTS OF SIGNIFICANCE —OR VICE VERSA*

THEODORE D. STERLING
University of Cincinnati

There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs—an "error of the first kind"—and is published. Significant results published in these fields are seldom verified by independent replication. The possibility thus arises that the literature of such a field consists in substantial part of false conclusions resulting from errors of the first kind in statistical tests of significance.

IT HAS become commonplace to speak of a "level of significance" in reporting outcomes of experiments. This significance level refers to risks of rejecting the null hypothesis, H_0, erroneously, and seemingly, has no other direct relationship to experimental work. The experimenter who uses so called tests of significance to evaluate observed differences usually reports that he has tested H_0 by finding the probability of the experimental results on the assumption that H_0 is true, and he does (or does not) ascribe some effect to experimental treatments. What with the shortage of publication space and the desire for objectivity it often seems that the responsibility for rejecting a hypothesis rests squarely on a crucial value in a table of probabilities.

The risk of choosing the incorrect inference from experimental observation depends on a stated risk of rejecting H_0 if true and on the risk of failing to do so if H_0 is not true. Here is a dilemma which is dealt with in practice by two conventions. As Savage notes [7, p. 256] publications tend to report the results of the test as well as that level of significance for which the corresponding test of the relevant family would be on the borderline between acceptance and rejection (in view the of the author). The individual reader now makes his own test at a level of significance appropriate to him. How much uncertainty such a reader is willing to tolerate in rejecting a hypothesis that might be true will depend on his confidence in the methods of data collection, his views concerning the relevance of alternative hypotheses, or the weight he gives to evidence from other sources. In addition, scientific readers differ in fundamental strategies for games against nature and their tolerance for errors can hardly be expected to remain unchanged from one experimental problem to another. The type of reporting mentioned by Savage may well be most satisfactory for author and reader alike.

Some publications, notably of social science content, have adopted a somewhat more extreme convention. Here a borderline between acceptance and rejection of H_0 is taken as a relatively fixed point, usually at $\Pr(E \mid H_0) \leq .05$ or

* The author wishes to express his thanks to Sir Ronald Fisher whose discussion on related topics stimulated this research in the first place, and to Leo Katz, Oliver Lacey, Enders Robinson, and Paul Siegel for reading and criticizing earlier drafts of this manuscript.

This content downloaded from 154.70.132.206 on Sat, 11 May 2019 11:22:03 UTC
All use subject to https://about.jstor.org/terms

at that approximate region for which the probability, (Pr) of the outcome (E) of the experiment, calculated on the assumption that H_0 is true, is no larger than five in a hundred[1] [3] [6] [8]. General adherence to such a rigid strategy is interesting by itself but might have no further consequences on the decisions reached. However, when a fixed level of significance is used as a critical criterion for selecting reports for dissemination in professional journals it may result in embarrassing and unanticipated results.

TABLE 31

OUTCOMES OF TESTS OF SIGNIFICANCE FOR FOUR PSYCHOLOGY RESEARCH JOURNALS

Journals: All Issues From January To December	Total Number of Research Reports (1)	Number of Research Reports Using Tests of Significance (2)	Number of Research Reports that Reject H_0 with $Pr(E\|H_0) \leq .05$ (3)	Number of Research Reports that Fail to Reject H_0 (4)	Number of Research Reports That are Replication of Previously Published Experiments (5)
Experimental Psychology (1955)	124	106	105	1	0
Comparative and Physiological Psychology (1956)	118	94	91	3	0
Clinical Psychology (1955)	81	62	59	3	0
Social Psychology (1955)	39	32	31	1	0
Total	362	294	286	8	0

Table 31 shows that for psychological journals a policy exists under which the vast majority of published articles satisfy a minimum criterion of significance. The table summarizes the number of research articles in four publications. The journals were selected at random from four major areas of psychology. The table gives the distribution for the number of reports that used tests of significance to test H_0 and either rejected H_0 or failed to do so at $Pr\ (E|H_0) \leq .05$. In addition the table gives the number of experiments that were replications of previously published investigations. Column 1 gives the number of experimental research reports and column 2 gives the number of those reports that used tests of singificance to choose among possible alternative hypotheses. Column 3 shows how many of the reports of column 2 managed to reject H_0 and column 4 counts the number of reports that failed to reject H_0 (either for the major hypothesis tested or for the majority of hypotheses under investigation.)[2] Finally, column 5 gives the number of experiments representing a replication of work previously reported in the literature.

[1] The fact that some tables present only the .05 and .01 levels of significance encourages the use of these two levels of significance [8, p. 292].

[2] Some explanatory remarks concerning Table 31 are in order. Almost all of the 294 studies that used tests of significance were of a multivariable design. All evaluated observed differences against the assertion of H_0, however, H_0 was sometimes not rejected for all variables tested. The following rules were adopted in compiling Table 31:

a. The attempt was made to determine the major variable or prediction tested by the research design. Such was usually clear from the author's preliminary remarks; the multivariable design was most frequently used to control for conditions not covered by the experimental procedure. The level of significance for which H_0 was rejected for the major prediction was noted (if H_0 was rejected at all).

b. If the design tested two or more variables for which no unambiguous decision as to major importance could be made, the lowest level of significance for which at least half the variables rejected H_0 was noted. If H_0 was not rejected for at least half the variables, the article was placed in the class of studies for which H_0 was *not* rejected.

This content downloaded from 154.70.132.206 on Sat, 11 May 2019 11:22:03 UTC
All use subject to https://about.jstor.org/terms

Table 32 shows the same distributions as proportions of columns 1 and 2.

A glance at the tables is sufficient to show that most articles published during the year by the journals in question used tests of significance as aides in choosing among alternative experimental hypotheses and, at the same time, that nearly all managed to reject H_0 at the recommended level of certainty. It need not be assumed that the observed distributions are due to explicit edi-

TABLE 32

PER CENT OF ARTICLES USING TESTS OF SIGNIFICANCE AND PER CENT OF ARTICLES REJECTING H_0

Journals: All Issues From January to December	Per Cent of Articles Using Tests of All Articles Published (2/1)	Per Cent of Articles Rejecting H_0 of All Articles Using Tests (3/2)	Per Cent of Articles Not Rejecting H_0 of All Articles Using Tests (4/2)
Experimental Psychology (1955)	85.48	99.06	0.94
Comparative and Physiological Psychology (1956)	79.66	96.81	3.19
Clinical Psychology (1955)	76.54	95.16	4.84
Social Psychology (1955)	82.05	96.88	3.12
Total	81.22	97.28	2.72

torial rules. The single factor contributing most to the selection of articles in which H_0 is rejected may be implicit agreement among authors. The term "publication policy" will be used here largely as a matter of convenience. In fact, the distribution of articles in psychological journals in general appears to be similar to the ones shown in the table and it seems likely that the authors selection rather than editorial policy accounts for the observed profession-wide selection. Whatever the reasons, the tables indicate what gets printed with a high probability; namely, research reports that use tests of significance and at the same time reject H_0 for the effects of treatments in the design.[3] To state the above more concisely:

c. Where results from more than one research design were reported, an attempt was made to determine the one study deemed most crucial by the author and the level of significance for that study was recorded if it rejected H_0.

d. If all studies seemed of equal importance, the lowest level of significance for which at least half the reported studies rejected H_0 was recorded. If H_0 was not rejected for at least half the studies reported, the article was placed in the class of studies for which H_0 was *not* rejected. (This special provision in 2 and 4 was not really necessary since for no single article were less than half of the quoted results in the significant category.)

e. Two studies that obtained $\Pr(E|H_0) \leq .1$ were included because the authors had expressly pointed out that they rejected H_0 since the obtained significance level was close enough to the conventional .05 to suit their purposes.

Since the *Psychological Abstracts* essentially attempt to present an outline of the major points made in almost all research articles of interest to psychologists the procedure used here could be checked for reliability with that publication. Of 100 research articles selected at random from volumes covering 1952 to 1957, 94 reported positive results, 5 reported negative results, and one was a replication of a previous study. These proportions agreed by and large with the total proportions in Table 31. No comparison for use of tests of significance were made since that journal seldom reports results of statistical tests. However, the words "significantly different" were applied to most of the reported results.

[3] It is interesting that the *Journal of Experimental Psychology* appears to set the pace for the use of statistical tests as well as for the selection of articles that reject H_0. Some years ago the same journal was used [4] to show that χ^2 was consistently misused by psychologists. The authors noted at the time that analyses in this journal would be typical for psychological publications in general and that the expectation of finding sound statistical treatments would be better in that journal than in others.

This content downloaded from 154.70.132.206 on Sat, 11 May 2019 11:22:03 UTC
All use subject to https://about.jstor.org/terms

A_1 Experimental results will be printed with a greater probability if the relevant test of significance rejects H_0 for the major hypothesis with Pr $(E|H_0) \leq .05$ than if they fail to reject H_0 at that level.

A_2 The probability that an experimental design will be replicated becomes very small once such an experiment appears in print.

With respect to A_1, it is not known how many research results either reject H_0 or do not do so, or, are submitted or not submitted for publication. However, it does seem clear [2] that pressure exists which leads to the selection of a very small number of publications from a large number of submitted manuscripts. From a commonly admitted tendency to acknowledge only the most significant findings, and from perusal of statements concerning publication pressures [2], one could infer another reasonable assumption:

A_3 A great many more experiments are performed than appear in the pages of professional journals.

With respect to A_2, the lack of replication of experimentation in psychology has been noted elsewhere [5]. Replications are sometimes reported at professional meetings. Since such papers are rarely used as references unless they have been published they may be ignored as sources for widespread professional or scientific information.

The three assumptions are admittedly substantive in nature and strong supporting evidence for them, beyond that given here, is hard to come by. They may be taken as a fair statement of the prevailing conditions in which the scientific community is not equally aware of all experimental results. As a consequence, experiments for which Pr $(E|H_0)$ is large may well have a high frequency of replication by individuals who do not know that this particular comparison had been made previously, and that previous tests of significance had failed to reject H_0 at acceptable levels of significance. Once a study does result in a level of significance that meets this criterion, not only will it be published, but the likelihood of its ever being repeated appears to become very small. A picture emerges for which the number of possible replications of a test between experimental variates is related inversely to the actual magnitude of the differences between their effects. The smaller this difference the larger may be the likelihood of repetition. This chain is terminated apparently by an observation for which the relevant statistical test can reject H_0 with reasonable certainty. For any set of observed differences that are randomly variable (and which experimental observations are not?) a difference of some substance should then appear in print—*irrespective of the actual state of nature.* What credence can then be given to inferences drawn from statistical tests of H_0 if the reader is not aware of all experimental outcomes of a kind? Perhaps even more pertinent is the question: Can the reader justify adopting the same level of significance as does the author of a published study?

Two points are worth noting with respect to the last two questions. Both refer to the expectations a reader may form when he picks up an article in one of the journals of Table 31 (or in a journal following like practices).

First the reader's best expectation is that the author will reject H_0. The probability that he will commit a Type II error (accepting the null hypothesis when it is false) if he adopts the author's conclusion is, in consequence, extremely small. In fact, from Table 31 it appears that this risk is scarcely more

This content downloaded from 154.70.132.206 on Sat, 11 May 2019 11:22:03 UTC
All use subject to https://about.jstor.org/terms

than zero. One may therefore conclude that any and all tests used by authors are of equally high power for the reader. This obviously was not true for the individual investigator who attempted to choose the most powerful test in the first place.

There is also another side to this problem. The reader's expectations are that H_0 will be rejected. What risks does he take in making a Type I error by rejecting H_0 with the author? The author intended to indicate the probability of such a risk by stating a level of significance. On the other hand, the reader has to consider the selection that may have taken place among a set of similar experiments for which the one that obtained large differences by chance had the better opportunity to come under his scrutiny. The problem simply is that a Type I error (rejecting the null hypothesis when it is true) has a fair opportunity to end up in print when the correct decision is the acceptance of H_0 for a particular set of experimental variables. Before the reader can make an intelligent decision he must have some information concerning the distribution of outcomes of similar experiments or at least the assurance that a similar experiment has never been performed. Since the latter information is unobtainable he is in a dilemma. One thing is clear however. The risk stated by the author cannot be accepted at its face value once the author's conclusions appear in print. It may be safe to conclude that pursuing statistical analyses under the conditions outlined here may have considerable less merit than psychologists like to ascribe to statistics in experimental design.

It would be unfair to close with the impression that the malpractices discussed here are the private domain of psychology. A few minutes of browsing through experimental journals in biology, chemistry, medicine, physiology, or sociology show that the same usages are widespread through other sciences. Some onus appears to be attached to reporting negative results. Certainly such results occur with lesser frequency in the literature than they may reasonably be expected to happen in the laboratory—even if it is assumed that all experimenters are outstandingly clever in selecting hypotheses. Perhaps the trend of our time is exemplified by the editors of a cancer journal who in a recent announcement took action to change the name of their yearly supplement from "Negative Data . . ." to ". . . Screening Data" [1, p. 619].[4]

BIBLIOGRAPHY

[1] American Association for Cancer Research, *Cancer Research 18*, University of Chicago Press, 1958.
[2] American Psychological Association, *Publication Manual* (1957 Revision), Washington, D. C., Psychological Association, 1957.
[3] Edwards, A. L., *Experimental Design in Psychological Research*, New York, Rinehart and Co., 1950.
[4] Lewis, D. and C. L. Burke, The Use and Misuse of the Chi-Square Test, *Psychological Bulletin* 46 (1949) 433–489.
[5] Lubin, A., Replicability as a publication criterion, *American Psychologist* 8 (1957) 519–520.
[6] McNemar, *Psychological Statistics* (Second Ed.), New York, John Wiley and Sons, 1955.
[7] Savage, L. J., *The Foundations of Statistics*, New York, John Wiley and Sons, 1954.
[8] Walker, H. M., *Elementary Statistical Methods*, New York, Henry Holt and Co., 1947.

[4] This was pointed out to me by Charles Stevens of the Kettering Laboratory, University of Cincinnati.

This content downloaded from 154.70.132.206 on Sat, 11 May 2019 11:22:03 UTC
All use subject to https://about.jstor.org/terms