Kasetsart J. (Social Sci) 6 : 63-69 (1985)

Negative Results-Should They be Published?

Suranant Subhadrabandhu

ABSTRACT

A growing controversy exists among scientists about presenting the negative results. This creates concerned especially for students and junior researchers. In developing countries where many technological methods have been adopted or modified from the more advanced countries, therefore both negative and positive results are abtained. The problem of publishing these data should be clarified by the editorial authorities. In this paper the pros and cons of publishing negative results have been gathered and discussed.

> "...there's this desert prison, see, with an old prisoner, resigned to his life, and a young one just arrived. The young one talks constantly of escape, and, after a few months, he makes a break, He's gone a week, and then he's brought back by the guard. He's half dead, crazy from hunger and thirst. He describes how awful it was to the old prisoner---the endless stretches of sand, no oasis, no signs of life anywhere. The old prisoner listens for a while, then says, "Yep, I know, I tried to escape myself, twenty years ago'. The young prisoner says, 'You did? Why didn't you tell me, all these months I was planning my escape? Why didn't you let me know it was impossible?' And the old prisoner shrugs, and says, 'So who publishes negative results?'"[1]

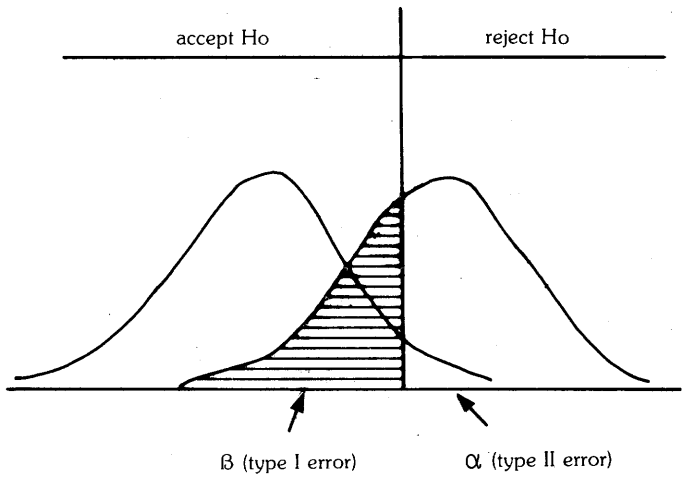
There is a growing controversy in the scientific communication concerning the editors' policies in publishing the scientific papers. Certain pressures exist in society which tend to place a burden on the admission of negative traits, happenings, trends, or results. Negative results play a major role in this argument, since most professional editors develop a policy of positive statistical test in determining which experimental results to be accepted for publication and which to reject.

Statistical tests are employed in many scientific experiments as a way of testing hypotheses. A statistical test examines a set of sample data and, on the basis of an expected distribution of the data, leads to a decision on whether to accept the hypothesis and accept an alternate one. The nature of the tests varies with the data and the hypothesis, but the same general philosophy of hypothesis testing is common to all test.[2] The data is chosen to represent the population from which it is extracted; the statistical tests can be used to show how well the data satisfies this representation. It is understood that significance does not imply absolute proof of the hypothesis; one can only assume this with a certain probability. Commonly, a researcher develops a null hypothesis (Ho), i.e. the hypothesis under test, and test against an alternate hypothesis (Hi). And a test of significance is a calculation by which the sample results are used to throw light on the truth or falsity of a null hypothesis. In common usuage positive results imply that the researcher can reject the null hypothesis 95 percent of the time (assuming $p = .05$); conversely, failures to reject the null hypothesis at the selected level of significance are referred to as negative results.

[1]Department of Horticulture, Faculty of Agriculture, Kasetsart University, Bangkok, 10900

In hypothesis testing, the researcher can make two kinds of errors: he can reject the hypothesis Ho when it is in fact true, or he can accept Ho when it is in fact false. If the hypothesis is true but the test rejects the hypothesis, an error known as a type I error is made. On the other hand, if the hypothesis is false but the test accepts the hypothesis, an error known as a type II error is made. The relative importance of these two kinds of errors depends upon what action is to be taken as a result of the test.

As an illustration, suppose that an innocent man is being tried for a crime and that his sentence hinges on the result of a certain experiment. If a hypothesis corresponding to innocence was set up and was rejected by the experiment, then an innocent man would be convicted and a type I error would result. On the other hand, if the man were guilty but the experimental result accepted the hypothesis corresponding to innocence, then a guilty man would be freed and a type II error would result. This can be seen in the figure below.



β = Probability of setting the guilty man free

α = Probability of convicting an innocent man.

From the figure, we can determine the probability of each of the two kinds of error. The following notation are adopted.

α = P (type I error) = Probability of rejecting Ho when Ho is true

β = P (type II error) = Probability of accepting Ho when Ho is false

A logical procedure for selecting efficient tests of statistical hypothesis can be designed by means of these two types of error. The procedure consists in first specifying that all the tests under consideration shall have the same-size type I error and then selecting that test for which the type II error is a minimum. Since the size of the type I error is the probability that the sample will yield a value falling in the critical region, the size of this error can be regulated by changing the size of the critical region.

Many editors of the professional journals in social sciences, natural sciences etc. tend to develop a policy of publishing papers having significant test. It has become apparent that an experiment failing to reject the null hypothesis at $p = .05$, i.e. exhibiting negative results at this level, will unlikely be published. Sterling (1959) reported that of 294 published research of four journals of psychology using tests of significance only eight failed to reject the null hypothesis. Sterling drew three conclusions from his research as:

1. Printing of experimental results becomes much more difficult if an experiment fails to reject the null hypothesis.

2. The probability of replication of an experimental design becomes quite small once such an experiment is published.

3. A great number of experiments are conducted that never appear in print (rejection rates for some journals range from 50 to 80 percent.)

Sterling stated that these three trends represent the prevailing condition in the scientific community. That is, there is not an equal awareness of all the experimental results obtained by those involved in research. This is due to the fact that experiments for which the probability of not rejecting the null hypothesis is large are repeated often by scientists not aware of the fact that the experiment has already been conducted many times with a failure to reject the null hypothesis. It appears that once a study achieves significance at the $\alpha = .05$ level, it will be published but not repeated. Sterling points that the problem of all of this is that a type I error (rejecting Ho when it is true) has a reasonable chance of being published when the correct decision is the rejection of the null hypothesis.

One critics to the ideas of accepting positive results for publication by journal editors are based on this acceptance of type I error. It has been pointed out that achieving a significance level of .05 implies only that 19 of 20 times a certain experiment will produce a certain result. But as Sterling suggests, how is the reader to assure himself that the experiment of which he is reading is not that one of the twenty case? Thus, before a reader of such an article can make an intelligent decision, he must have an idea of the distribution of outcomes of similar experiments or assurance that a similar study has never been conducted. Since the latter information can not be obtained, the reader is left in an uncertained situation.

Also there are some who claim that the scientific journals are filled with this type I error. If this problem does exist, then there is need for replication of published works. However, Sterling has revealed that in a review of 362 research reports published not one was a replication of a previously published experiment. Therefore not only might we have type I error in the journals but the danger exists that erroneous conclusions drawn from the experiments by readers may never be corrected.

Commenting on Sterling research and conclusion, Tullock (1959) does not think the repitition of some given experiment will frequently occur. Tullock feels that fellow researchers reading of an experiment which they have previously conducted and suspected to contain type I error, would be very quick to respond thereby. They may at least write the editor of the journal in which it appears.[4] In his opinion, Tullock think that the type of duplication of an experimental design in which Sterling is so alarming of is rarely occurred. Whether Sterling's alarm or Tullock's confidence is more justified, both agree that there is considerable opportunity for errors and the drawing of erroneous conclusions under the present publishing criteria.

On the other side of the argument, Bakan (1967) thinks that editorial policies should

be prejudiced against negative results. He states that under the pressure of receiving far more papers than they could realistically publish, journal editors should use the magnitude of the level in the appropriate significance test as a basis for acceptance or rejection of manuscripts. Bakan further feels that studies in which significance is not obtained are not even submitted for review. With researchers selecting only their significant studies for publication, as well as discarding data which does not support their hypotheses, the obvious consequence is that published results are more likely to contain type I errors (rejecting the null hypothesis when it is true) in them far in excess of the five percent which we allow under the assumptions of the test of significance at p = .05. Bakan's suspicious that type I errors are plaguing scientific literature is given somewhat of a confirmation in a review of articles published in the Journal of Abnormal and Social Psychology. Analysis of 70 reports in which significant results were obtained revealed that the studies were not using statistical tests powerful enough to have detected any difference between the means of the populations or groups involved. Bakan further says that the damage to scientific investigation is aggravated by the fact that printing of studies containing type I errors tends to halt further study in that particular area. If the publication of studies containing type I errors would stimulate others to probe the same area of study, the situation would not be bad at all. However, with the current predication of published type I errors stopping further investigation, the type I error thus presents a much more serious danger to science than does the type II error (accepting the null hypothesis when it is false)[5]

A former publisher of the Journal of Experimental Psychology, Aurthur W. Melton listed his criteria for publication of an article as follows:

1. Is the research a valid experiment?
2. Can the results be repeated under the conditions described?
3. Did the experiment achieve a level of significance of at least .05?[6]

Melton (1962) defends the use of the positive results criterion as being as logical as any and better than others suggested. Melton's reason for rejecting negative results is that they are most frequently submitted as a result of an experiment which has not been given a chance to reject the null hypothesis. Melton describes his journal as an archive of the science; as such, he believes it should not be confused with a newsletter which prints information and notices of a less substantial nature.

The controversy over what should or should not be published has intensified in recent years due to several factors. Firstly, the number of scientists performing experiments has increased many fold during this period; this results in an increasing number of experiments submitted for publication. Secondly, the enormous and irresistable pressure to publish which has been brought to bear upon the scientists; their professional reputation, job potential, salary level, etc. have been linked directly to the quantily and quality of their publications. Another factor contributing to the lack of publication space is that while the first two points mentioned have joined to produce a deluge of experiments for publication consideration, neither the number, size nor publication frequency of the professional journals have increased proportionally. This results in the high percentage of rejections mentioned by Sterling.

Walster and Cleary (1970), in referring to Sterling's critique, cite several consequences of the current editorial policy that preclude the use of journals to distribute and document knowledge. These consequences are:-

1. It becomes difficult, even impossible in some case for a reader to differentiate those articles that report true findings from those that report type I errors.

2. It leads researchers to analyze their data or conduct their studies in such a manner that they consciously search for statistical significance. This search is motivated by the knowledge that statistical significance is a prerequisite for publication.

3. A large number of studies are not published even though they contain important, but statistically insignificant findings. This results in the loss of valuable imformation for interested readers.

4. Journal readers are letting editors and reviewers decide for them the level of significance that is indicative of an important finding. It is well known that the power of a statistical test is a very fragile quantity: it can be affected by a variety of factors, such as sample size, reliability of the dependent measure, and the strength of the various treatments used in the experiment. Upon consideration of these factors, a knowledgeable reader of the literature will wish to determine for himself the level of statistical significance that indicates what to him is an important finding. If the studies whose results fall below a given arbitrary level of significance are not published, it is impossible for a reader to make rational decisions concerning the presence or absence of effects of interest to him.[7]

These 4 consequences of the biased editorial policy pointed out by Walster and Cleary aimed at proposing a change in publication policy. The central basis of their proposed alternative policy lies in the cardinal rule of experimental design. This rule argues that all decisions regarding the treatment of data should be design decisions. That is, data handling decisions must be made prior to inspection of the data in order to avoid personal bias in the analysis. Also this rule should be applied to publication decisions. Walster and Cleary propose that the extension of the design decision to publication involve submission of articles for review without the data and results. This would insure that the publication decision would be based upon factors such as adequacy of design and relevance to current issues of interest, and be free of bias from the editors and reviewers.

When the report of a research study is submitted for review it could include topics from the following lists of supporting information:

1. Theoretical relevance and/or justification.

2. Relevance to applied and/or topical issues.

3. Predicted outcomes and the implications for the theoretical and/or applied problems.

4. If the study is, or includes, a replication of previously published research, a discussion of the need for the proposed replication;

5. Detailed description of the procedure of the study, including the source of subjects, description of randomization scheme,transcript of instructions given subjects, description of independent and dependent variables;

6. Any previous research and/or data which indicates the extent to which the independent variables are validly being manipulated.

7. Any previous research and/or data which indicate the extent to which the dependent variables are reliable and/or valid;

8. Discussion of the proposed data analysis and its relevance to the predicted outcomes;

9. Pilot data that tend of support the predicted outcome, especially if the predictions are at variance with existing theories or published results.

The benefits resulted from the change proposed in the review policy of social science journals by Walster and Cleary are 4 folds.

First, an examination of the literature would enable one to distinguish which studies are reporting type I errors and which reporting important findings.

Second, the pressure to publish would be re-directed so that researchers would feel a need to design substantively important and method logically sound studies rather than merely to achieve statistical significance.

Third, the information contained in studies which do not achieve statistical significance would not be lost as is now.

Fourth, the individual reader would be allowed to decide for himself the real worth of the results of published research studies.

An additional benefit would be a shift in research planning and execution. A scientist considering a research topic would submit for journal for review a proposed study. If the proposed study was accepted for publication, a researcher could execute his study with the guarantee that it would be published regardless of statistical significance for his data. If the proposal was rejected the time and money invested in executing the study would not be lost. The overall result will be an increase in the soundness of experimental designs and the validity of experimental results.[8]

In conclusion, I think that the policy of publishing only positive results should be revised, since publication of only positive results of experiments has several very detrimental aspects and may not be an ideal situation to undertaken. However, this policy of positive result publication may be hard to change if the publication space is not increased. Although I am agree with Melton that the negative result experiments should not be published in comparison to the number of positive result articles waiting for the queue of available space in the journal. Furthermore Tullock's recommendation of allocating space in the journals for a summary of replications and experiments with negative results is an excellent one and is worth considering.

FOOTNOTES

Hudson, Jeffrey. 1968. *A Case of Need.* New York: The New American Library Inc. p. 168
Sakal, R.R. and F.J. Rohlf. 1969, pp. 155-158
Sterling, 1959, pp. 30-34
Tullock, 1959, pp. 593
Bakan, 1967, pp. 1-29
Melton, 1962, p. 556
Walster and Cleary, 1970, p. 17
Walster and Cleary, 1970, p. 18

LITTERATURE CIPED

Bakan, D. 1967. "The test of Significance in psychological research." *On Method.* San Francisco, Jossey-Bass Inc.,

Hays, W.L., and R.L. Winkler 1970. Statistics, Probability, Inference, and Decision. Volume I. Holt, Rinehart and Winston, Inc. Ch. 7 pp. 375-443

Hoel, P.G. 1947. Introduction to Mathematical Statistics:-John Wiley & Sons, Inc. N.Y. Ch. 11 pp. 201-213

Melton, A.W. 1962. Editorial:- Jour. Expt. Psychology 64:553-557

Sakal, R.R. and J.F. Rohlf. 1969. Biometry. San Francisco:-W.H. Freeman and Company pp. 155-160

Sterling, T.D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa:- Jour. Amer. Stat. Assoc. 54: 30-34

Tullock G. 1959. Publication decisions and tests of significance -a comment. Jour. Amer. Stat. Assoc. 54:593

Walster G.W. and A.T. Cleary 1970. A proposal for a new editorial policy in the social sciences. The Amer. Stat. April 1970: 16-19