Misuse of one-tailed Test in Research

Abstract

Experimental research is an area where investigators have to make a decision regarding the use of one-tailed or two-tailed tests in hypothesis testing. Statistical hypothesis starts with the negation of the alternative hypothesis. Type I and type II errors are involved in the testing of hypothesis. The choice of one-tail or two-tail test is neither simply mechanical nor mathematical statistics but primarily a matter of experimental logic and human judgment. The investigator is more interested in reducing the Type I error, so, a two-tailed test must be preferred over a one-tailed test. It is true that the one-tailed test is more powerful but it should be used sparingly in the experimental research.

Experimental research is an area where it would be critical to control variables other than those which are under investigation. The procedure of control is much complicated in social sciences and may be beyond what an investigator can possibly achieve with full confidnece. However, experimental designs provide some methods to surpass this problem. Investigators in experimental research try to establish the effect of an intervention/treatment on a sample of subjects. In this context this paper discusses the testing of hypothesis and the use of one-tailed and two-test tests in educational research.

Experimental data very often require a comparison and evaluation of two or more means. For example, an investigator may be interested to explore the effect of a learning package or model of teaching on achievement in a subject for a particular group of students. The experiment requires an evaluation of the difference between two mean scores. The mean score of the achievement of the experimental group and the mean score of the achievement of the control group which was taught using the traditional/conventional method are to be compared. The difference may either be ascribed to the sampling error or may it be argued with confidence that the package/method affects achievement. A decision is required between the two alternatives. Statistical procedures which lead to decision of this kind are known as tests of significance (Ferguson, 1971).

Hypothesis Testing

In the study of the kind mentioned above, a treatment/intervention/method of instruction is applied to the experimental group and the control group is deprived of the treatment or is taught using the conventional or traditional instructional method. Presumably any significant difference between the two groups can be ascribed with confidence to the treatment. Let X_1 and X_2 be the estimates of the population μ_1 and $\mu_{2.}$ The trail hypothesis may be formulated that no difference exists between μ_1 and $\mu_{2.}$

Null hypothesis

Statistical hypothesis starts with the negation of the alternative hypothesis that there is a difference between the means μ_1 and $\mu_{2.}$. Thus, the negation of the alternative hypothesis is called the null hypothesis H_0. Using the sample information the null hypothesis is accepted or rejected. The null hypothesis can represented as. H_0: $\mu_1 = \mu_2$ or $\mu_2 - \mu_1 = 0$

In order to show the tenability of the alternative hypothesis, $H_{A,}$ one must find evidence against the null hypothesis, H_0, (Lombardi & Hurlbert, 2009). The null hypothesis asserts that no difference exists between the two population means. In other words, the investigator operates on the hypothesis that the treatment applied will have no effect. The investigator examines the empirical data to see the difference between the two means. If there is a difference, then the important question is what is the probability of obtaining a difference equal to or greater than the one observed in drawing samples at random from populations where the null hypothesis is assumed to be true? If the probability is small, the observed result being highly improbable on the basis of the null hypothesis, the investigator may be prepared to reject the null hypothesis (Ferguson, 1971). This means that the observed difference cannot reasonably be explained by sampling error and presumably may be attributed to the treatment applied. Thus, the result may be said to be significant. On the contrary, if the probability cannot be considered small and the observed result is not highly improbable, the sampling error may account for the difference observed. Hence, one cannot with confidence infer that the difference results from the treatment applied. In the testing of any statistical hypothesis, it is necessary to specify an alternative hypothesis. H_A: $\mu_1 \neq \mu_2$ or $\mu_2 - \mu_1 \neq 0$

Error involved in the testing of hypothesis

Two types of error may occur in reaching a decision about the null hypothesis H_0. - - (1) H_0 is rejected when H_0 is the true state, i.e., an alternative H_1 may be accepted when the null hypothesis H_0 is true. This is called the Type I error. The H_0 is accepted when H_0 is false, i.e., the null hypothesis H_0 may be accepted when the alternative hypothesis H_A is true. This sis called a type II error. The probability of Type I error is symbolized by α, the probability of Type II error is denoted by β. There is no error when a true H_0 is accepted and a false H_0 is rejected (Glass & Hopkins, 1984). This situation is represented below.

	True State of Nature	
	H_0 True	**H_0 False**
Accept H_0		Type II error

	Correct	β
Reject H_0	Type I error α	Correct

In the procedure of testing, it is not possible to reduce both errors. The investigator is interested in the rejection (falsification) of the null hypothesis, that is why Type I error is considered as important and fixed at a certain level. The next step is to decide what error, or probability, we are willing to tolerate for incorrectly rejecting H_0 (making Type I error). This probability of making a Type I error is called the significance level (Kleinbaum, Kupper & Muller, 1988). For fixed sample size, α and β are inversely related. If one tries to guard against making Type I error by choosing a small rejection region, then the non-rejection region (and hence β) will be large. Conversely, protecting against a Type II error necessitates using a large rejection region, leading to a large value for α. Increasing the sample size generally decreases β, however, α remains unaffected (Kleinbaum, Kupper & Muller, 1988). So, a test is obtained by minimizing Type II error keeping Type I error at a level. Level significance is 0.05 means that here is a maximum probability of 0.05 wrong rejections. Conventionally the term null hypothesis has been restricted to a hypothesis of no difference. However, it is not inappropriate to extend the meaning of the null hypothesis to include hypotheses of equal to or less than and equal to or greater than (Ferguson, 1971).

Variants of Null Hypothesis

Any of three different sets of hypotheses can be analysed with a *t*-test:

Set 1 H_0: $\mu_1 - \mu_2 = 0$ H_A: $\mu_1 \neq \mu_2$

Set 2 H_0: $\mu_1 - \mu_2 \leq 0$ H_A: $\mu_1 - \mu_2 > 0$

Set 3 H_0: $\mu_1 - \mu_2 \geq 0$ H_A: $\mu_1 - \mu_2 < 0$

Conventionally, a *t*-test to Set 1is applied if the investigator is interested in detecting any difference, positive or negative, between μ_1 and μ_2. Such a test is variously referred to as a two-tailed, two-sided, or non-directional test. If our concern is to determine only whether $\mu_1 < \mu_2$ or only whether $\mu_1 > \mu_2$ then we apply a *t*-test either to Set 2 or Set 3. Such a test is termed a one-tailed, one-sided, or directional test. The advantage of a one-tailed test is that, for fixed α, it has greater power than the two-tailed test for detecting a difference in the direction tested.

What is a two-tailed test?

In using a significance level of 0.05, a two-tailed test allots half of the alpha to testing the statistical significance in one direction and half of the alpha to testing statistical significance in the other direction. This means that .025 is in each tail of the distribution of the test statistic. When using a two-tailed test, regardless of the direction of the relationship hypothesized, the investigator is testing for the possibility of the relationship in both directions. For example, one may wish to compare the mean of a sample to a given value *x* using a t-test. The null hypothesis is that the mean is equal to *x*. A two-tailed test will test both if the mean is significantly greater than *x* and if the mean significantly less than *x*. The mean is considered significantly different from *x* if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.

What is a one-tailed test?

In using a significance level of .05, a one-tailed test allots the entire alpha to testing the statistical significance in the one direction of interest. This means that .05 is in one tail of the distribution of your test statistic. When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction. Consider the example of comparing the mean of a sample to a given value *x* using a t-test. The null hypothesis is that the mean is equal to *x*. A one-tailed test will test either if the mean is significantly greater than *x* or if the mean is significantly less than *x*, but not both. Then, depending on the chosen tail, the mean is significantly greater than or less than *x* if the test statistic is in the top 5% of its probability distribution or bottom 5% of its probability distribution, resulting in a p-value less than 0.05. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction. A discussion of when this is an appropriate option follows.

When is a one-tailed test appropriate?

The one-tailed test provides more power to detect an effect, so the investigator may be tempted to use a one-tailed test whenever s/he has a hypothesis about the direction of an effect. Before doing so, consider the consequences of missing an effect in the other direction. Imagine that a new drug has been developed and the investigator believes that it is an improvement over an existing drug. The investigator wants to maximize the ability to detect the improvement, so the investigator opts for a one-tailed test. In doing so, the investigator fails to test for the possibility that the new drug is less effective than the existing drug. The consequences in this example are extreme, but they illustrate a danger of inappropriate use of a one-tailed test.

Now the basic question is - - when is a one-tailed test appropriate? If you consider the consequences of missing an effect in the untested direction and conclude that they are negligible and in no way irresponsible or unethical, then you can proceed with a one-tailed test. For example, imagine again that you have developed a new drug. It is cheaper than the existing drug and, you believe, no less effective. In testing this drug, you are only interested in testing if it less effective than the existing drug. You do not care if it is significantly more effective. You only wish to show that it is not less effective. In this scenario, a one-tailed test would be appropriate.

When is a one-tailed test NOT appropriate?

Choosing a one-tailed test for the sole purpose of attaining significance is not appropriate. Choosing a one-tailed test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was. Using statistical tests inappropriately can lead to invalid results that are not replicable and highly questionable.

One-tail versus two-tail

The choice between one-tail and two-tailed test has been and continues to be a controversy among social scientists in general and educational researchers in particular (Pillemer, 1991). The choice of one- rather than two-tailed hypothesis testing strategy can influence research outcomes. Researchers are often confronted with the decision whether to use a one-tailed or a two-tailed test. In a one-tailed test one predicts that one group scores highly than the other, whereas in a two-tailed test one makes no such predictions. Ferguson (1971) is of the opinion that directional tests should be used more frequently. Marks (1953) commented that directional hypotheses have acceptance among statisticians whereas Eysenck (1960) remarked that one-tailed tests have no place in psychology. Moreover, the logic underlying one-tailed tests is incompatible with efforts to seek out and learn from unusual and unexpected variations in study outcomes. Unexpected results can generate new theories, even if the findings are not immediately interpretable within the existing theoretical context. Pillemer (1991) concluded that the popularity of one-tailed test is largely attributable to the continued emphasis on attaining statistical significance at the arbitrary 0.05 level, rather than the conceptual or methodological considerations.

About choosing one-tailed or two-tail test, Kimmel (1957) observed - - "… it is important to note that the argument is not one of mathematical statistics but primarily one of the experimental logic" (p. 351). Null hypothesis testing cannot be done mechanically without running the risk of obtaining nonsensical results; human judgment must be an integral, and controlling, aspect of the process (Nickerson, 2000, p. 290). In experimental studies, 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 (Eysenck, 1960, p. 270). The most common reason for conducting a one-tailed test is the researcher's a priori expectation about irectionality. The possibility of an experimental programme designed to enhance achievement may instead lower the achievement. How a researcher can rule out this possibility? However, when theoretical considerations predict that the difference will be in a given direction, one-tailed test is certainly appropriate.

It is a matter of fact that a given mean difference in the hypothesized direction is more significant under one-tailed hypothesis than under a two-tailed hypothesis. It is true that one-tailed tests have increased power than two-tailed tests as it makes assumptions about the population and the direction of outcome (Cohen, Manion & Morrison, 2011). Some researchers have observed that one-tailed tests should never be used because they introduce greater potential for Type I errors and create an uneven playing field when outcomes are compared across programs (Ringwalt, Paschall, Gorman, Derzon, & Kinlaw, 2011).

Significance: A misnomer?

The probability of Type I or α error is called the level of significance of a test. Ordinarily the investigator arbitrarily adopts the common convention of a significance of 0.05 or 0.01. That is, at 0.05 level of significance the chances are 5 in 100, or less, the difference could result when a treatment applied is having no effect. Nickerson (2000) has discussed in detail the misunderstandings and false beliefs about the p values.

Consider an example in which IQ of 2000 students (when n is very large, even a trivial difference may be large enough to be highly statistically significant) was measured and the mean value is 101, one would reject H_0: $\mu = 100$ at 0.01 level of statistical significance. Obviously a one-point difference in IQ has little or no practical significance even though it may be highly significant in a statistical sense. It is unfortunate that the term "significant" was ever chosen to denote the untenability of H_0. Perhaps the term "reliable" would have been a better term for describing sample results and would less often be confused with practical significance or importance (Glass & Hopkins, 1984).

Criteria for the use of one-tailed tests

Kimmel (1957) has suggested three criteria for the use of one-tailed tests which are very useful. They are - - (1) use a one-tailed test when a difference in the unpredicted direction, while possible, would be psychologically meaningless, (2) use a one-tailed test when results in the unpredicted direction will, under no condition, be used to determine a course of behavior different in any way from that determined by no difference at all and (3) use a one-tailed test when a directional hypothesis is deducible from a psychological theory but results in the opposite direction are not deducible from coexisting psychological theory. If the result in the opposite direction (contrary to the directional hypothesis) is explainable by a coexisting theory or construct then the experimenter is biased in using a one-tailed test. The third criterion mentioned above is very critical in seeking proof for the effectiveness of a method of instruction or any such intervention where there is no evidence in literature to make a directional hypothesis. In the case one-tailed test the investigator must provide arguments such as results of previous experiments, predictions made on the basis of widely accept theories, number of cases involved, meta-analysis of studies and the like. Without a concrete ground it will be a misuse (abuse) of the one-tailed test. 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 (Eysenck, 1960, p. 271). Lombardi & Hurlbert (2009) reported that 34% of experimental medical studies published during 1975-88 used one-tailed tests and that a sizable segment of the research community believes that one-tailed test is appropriate. The recommendation is that one-tailed tests should be rarely used for basic and applied research.

References

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). New York: Routledge.

Eysenck, H. J. (1960). The concept of statistical significance and the controversy about one-tailed tests. *Psychological Review*, 67(4), 269-271.

Ferguson, G. A. (1971). *Statistical analysis in psychology & education*. (3rd ed.). New York: McGraw-Hill Book Company.

Glass, G. V., & Hopkins, K. D. (1984). Statistical methods in education and psychology ((2nd ed.). New Jersey: Prentice-Hall.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, 54(4), 351-353.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods* (2nd ed.). Boston: PWS-KENT Publishing Company.

Lombardi, C. M., & Hurlbert, S. H. (2009). Misprescription and misuse of one-tailed test. *Austral Ecology*, 34, 447-468.

Marks, M. R. (1953). One- and two-tailed tests. *Psychological Review*, 60, 207-208.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241-301.

Pillemer, D. B. (1991). One- versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher*, 20(9), 13-17.

Ringwalt, C., Paschall, M. J., Gorman, D., Derzon, J., & Kinlaw, A. (2011). The use of one- versus two-tailed tests to evaluate prevention programs. *Evaluation & the Health Professions*, 34(2), 135-150.