Journal of Health Science & Education

An open access journal

eSciRes

JHSE-1-123

Research Article

# Education Pertaining to Understanding Common Inferential Procedures in Health Research

**Trafimow D**[*]

*Department of Psychology, New Mexico State University, USA*

## Abstract

There is a tendency of consumers of the health science literatures to engage in dichotomous thinking: the finding is significant or not significant. But there are important problems with null hypothesis significance testing that fail to justify drawing conclusions based on the p-values generated to conduct significance tests. Nor do confidence intervals solve these problems. The main goal of the present article is to provide clear descriptions of the problems so that consumers of health science literatures can better evaluate the contribution each article makes. Because of the problems with inferential statistical procedures, consumers of the health science literatures are urged to investigate, much more carefully, the descriptive findings.

**Keywords:** Null hypothesis significance testing; Confidence intervals; Modus tollens; Inverse inference

## Introduction

The obvious way for workers in the health areas to become educated in relevant research is to read the health science literatures. However, an obstacle for the statistically unsophisticated is that empirical articles are littered with inferential statistics, such as t, F, X2, and so on that are used to arrive at values for p. In turn, p is crucial for null hypothesis significance testing (NHST). If p is below a threshold, generally set at 0.05 ($p<0.05$), the finding is deemed "significant," with the strong implication that it should be published, believed, and taken seriously. In contrast, if p is above the threshold ($p>0.05$), the finding is unlikely to be published. Or if published, the reader is encouraged not to take the finding very seriously. An example of this is when researchers wish to collapse across groups, to gain statistical power for an upcoming analysis, and justify such collapsing based on a lack of a significant difference between the two groups on the dependent variable of interest. Unfortunately, this sort of dichotomous reasoning—significant versus not significant—which most readers of health science literatures use, is problematic [1,2]. To aid in health science education, the present goal is to expose the problems so that health sciences workers and researchers are better able to evaluate findings reported in health science literatures.

## The Logical Problem

Consider what p is. Under the assumptions of a statistical model, p is the probability of obtaining the finding (or a more extreme finding), given the null hypothesis (usually of no difference between the experimental and control conditions). And that is all it is. Thus, p is not the probability of replication, the probability of obtaining the finding by chance, the importance of the finding, and so on ([3] for a list of what p is not). Suppose a researcher obtains a small value for p (less than 0.05). The general tendency is to conclude that the null hypothesis is unlikely to be true. From the point of view of the logic of modus tollens, this may seem to make sense. To see this, imagine a syllogism such as the following.

1. If the null hypothesis is true, the finding cannot happen {Major Premise}
2. The finding happens {Minor Premise}
3. Therefore, the null hypothesis is not true {Conclusion}

The foregoing syllogism is logically valid and follows the ancient Greek form known as modus tollens. However, it is not sound because the major premise is generally untrue. That is, even if there is no difference between two populations, it nevertheless is possible to obtain the researcher's finding.

Let us rewrite the syllogism, taking probability into account.

1. If the null hypothesis is true, the finding is unlikely to happen {Major Premise}
2. The finding happens {Minor Premise}
3. Therefore, the null hypothesis is not likely to be true {Conclusion}

Although the new syllogism may seem logically valid, it is not, as modus tollens does not work with probability [4,5]. Unfortunately, there is no way to calculate the probability of the null hypothesis being true, given the finding, unless one knows the base rate probability of the null hypothesis, and the probability of the finding given that the null hypothesis is not true [5,6]. Both are problematic. How would one know the base probability of a hypothesis being true, prior to research? And how would one know the probability of the finding given that the null hypothesis is not true, when there is an infinitude of ways in which the null hypothesis can be false? The bottom line is that a low value for p fails to provide much information about the probability of the null hypothesis, given the finding. The commission of this "modus tollens error" is a major problem with using statistical significance to come to conclusions about hypotheses.

An alternative to the modus tollens error, to see the logical problem with NHST, is to consider the "inverse inference error." Consider the following bullet-listed questions.

- What is the probability that someone is president of the USA, if that person is an American citizen?

- What is the probability that someone is an American citizen, if that person is president of the USA?

Obviously, the answers to the two bullet-listed questions are very different; the low probability for the first bullet-listed question fails to support a low probability for the second bullet-listed question. If one were to conclude a low probability for the second bullet-listed question based on a low probability for the first, this would be an example of the inverse inference error; that is, making the inference that a low probability for event A, given event B, supports a low probability for event B given event A. Moving to significance testing, a low probability for the finding, given the null hypothesis, is not sufficient to justify the inverse inference of a low probability for the null hypothesis, given the finding.

## The Type I Error Way around the Logical Problem

As Trafimow et al. described, there is a way of circumventing the foregoing logical issues (the modus tollens and inverse inference errors) [7]. It depends on the notion of Type I error, which is the error of wrongly rejecting the null hypothesis. The idea is to set a criterion level, and fail to reject the null hypothesis if the p-value comes in below the criterion, but to fail to reject it otherwise. This procedure ensures that one will only make the error of wrongly rejecting the null hypothesis—a Type I error—at an agreed-upon rate. In most contemporary journals in the health sciences, that rate is set at 0.05. Thus, by insisting that p comes in below 0.05 to reject the null hypothesis, researchers will wrongly reject the null hypothesis only 5% of the time.

Although this solution answers the logical problem described in the foregoing section, it creates new problems ([7] for a lengthy list). The most obvious problem is that there are other errors that can be committed, in addition to Type I errors. For example, one can fail to reject the null hypothesis when it is false. This is termed a Type II error. And there may be times when it is worse to commit a Type II error than to commit a Type I error. For example, suppose a researcher studies a mortal side-effect of a drug, commits a Type II error (fails to reject the null hypothesis), and thereby wrongly concludes that there is no mortal side-effect when there is one. This Type II error might be far worse than committing the Type I error (rejecting a true null hypothesis), wrongly concluding that that there is a mortal side effect. Which wrong conclusion is worse, failing to detect the mortal side-effect that is there or detecting a mortal side-effect that is not there? Likely, what is worse may depend on other factors such as the seriousness of the disorder, the success rate of the treatment, and so on. The larger point is that it sometimes is worse to commit a Type I error and it sometimes is worse to commit a Type II error, and setting a universal threshold level impedes researchers from adjusting the probabilities of Type I and Type II errors appropriately for the health context at hand.

A way out might be to allow researchers to set their own rates. But variable thresholds would fail to control the probability of a Type I error across the health sciences, which was the way of circumventing the logical problems of the modus tollens error and the inverse inference error described earlier. One cannot have it both ways. There is no way to both control the Type I error rate across health science research and at the same time provide researchers with the flexibility to adjust Type I and Type II error rates to fit the contextual factors at play.

## Having Overestimates of Effect Sizes in Health Science Literatures

There is a problem even direr than those already mentioned. To see the problem, consider a preliminary problem that, in some sense, significance is not very relevant because the effect size is much more important. To see this quickly, imagine a new medicine to cure dandruff that works for 87% of cases rather than the control medicine that only works 86% of the time, for an increase of 1%. Under typical sample sizes, this result likely would not be significant, but suppose that the researcher had a gigantic sample size, in which case the result could be significant. But significance does not indicate importance. The fact of the matter is that the effect size is so low, that even if it were statistically significant due to a gigantic sample size, it is not very meaningful. Thus, statistically sophisticated authorities have emphasized, and continue to emphasize, effect sizes rather than statistical significance [8,9]. This is a reason why top journals tend to require significance to demonstrate that the effect is there (though, as we have seen, significance does not accomplish this); but also require effect sizes to demonstrate importance.

What often is underappreciated is that p depends on two things: size of the effect and size of the study. One can obtain statistical significance with a small study, providing the effect size is sufficiently large; or with a small effect size, providing the study is sufficiently large (as in the dandruff example). We already have seen how the latter can be problematic. However, most published research in health sciences is based on small studies where the obtained effect sizes were sufficiently large to result in significance. Is this problematic too?

It is, but for a complex reason. The problem is the interaction of having a threshold for publication and regression to the mean. To see the interaction, it is necessary to understand how regression to the mean works. There are two classes of reasons for extremely low or extremely high scores on anything. These are systematic and random, and typically both apply. Keeping this in mind, suppose that an experiment results in a significant value for p (less than 0.05). Why did the low value occur? It could have occurred because the manipulation really works but it also could have occurred because the researcher was lucky in that study. When the null hypothesis is true, p-values are distributed uniformly between 0 and 1, with a mean of 0.5. Thus, under the null hypothesis, even when a researcher obtains $p<0.05$, the best prediction for the p-value to be obtained in a replication experiment would be 0.5, rather than 0.05. The larger point to be made is that p-values have a sampling distribution, just like any other statistic. Thus, it is likely that at least some of the reason for an extremely low p-value is luck (the researcher happened to sample an atypically low value from the set of p-values that could have been sampled).

Here is where the phenomenon of regression to the mean interacts with setting a threshold level for significance. Because it is necessary to beat the threshold level of $p<0.05$ to publish, it should be obvious that if replications were performed, many of the replications would fail to result in statistically significant findings. Of course, researchers often perform questionable research practices [10-13], but that is not the present point. The present point is that even if researchers were perfectly ethical in their statistical practices, the phenomenon of regression to the mean guarantees that replication p-values likely would be less extreme—they would regress to the mean. An empirical demonstration of this regression to the mean was obtained by the researchers in the Open Science Collaboration [14]. They found that well over 60% of the replication studies resulted in failed replications, using statistical significance as the criterion.

But matters get worse. Consider the point covered earlier, that under typical sample sizes, significance is largely determined by the effect size. But note that it is the sample effect size that determines the p-value, and not the population effect size (though the population effect size hopefully is reflected in the sample effect size). Why does this matter? Well, the statistical phenomenon of regression to the mean is crucial for effect sizes as well as p-values. Using a p-value threshold implies not only the publication of findings that pass the threshold, but also the publication of effect sizes sufficiently large to allow p-values to pass the threshold. In other words, we have a literature of published effect sizes that necessarily are overestimates of the expected effect sizes that would be obtained if each study were replicated many times with all of them being published. Again, the Open Science Collaboration [14] obtained empirical evidence of regression to the mean applied to effect sizes; the average effect size in the original cohort of studies was 0.403 whereas it was only 0.197 in the replication cohort. Regression to the mean strikes again!

The statistical fact that published effect sizes necessarily overestimate true ones has been the topic of much discussion lately, with some of the discussion even involving a suggestion that manuscripts be pre-approved so that publication would not depend on p-values passing a threshold [15-20]. Although a discussion of this issue is beyond present scope, the mere fact that people are having the discussion points to an increased recognition that NHST forces many published effect sizes to be dramatic overestimates of true ones.

## Confidence Intervals

Many recognize the foregoing problems with significance testing, and have advocated that p-values and NHST be dropped in favor of confidence intervals ([21] for a review). The idea of a confidence interval is that the researcher uses the combination of the sample size, and the standard deviation, to compute a standard error. And it is the standard error that determines the size of the confidence interval. Note that the computation of p-values also depends on the standard error. Given that the mathematics are very similar, what makes confidence intervals better than significance tests?

Well, that is not clear. For one thing, although sophisticated confidence interval aficionados do not advocate this, the typical use of confidence intervals in published literatures is as an alternative way to perform NHST. If the obtained sample statistic is within the confidence interval, the researcher fails to reject the null hypothesis, whereas if the obtained sample statistic is outside the confidence interval, the researcher does reject the null hypothesis. Used in this way, confidence intervals suffer from all the problems we described earlier in the context of NHST.

Alternatively, confidence interval aficionados have argued that they can be used for parameter estimation. But this is highly questionable. To see the problem, suppose a researcher constructs a 95% confidence interval. The obvious interpretation is that the population parameter has a 95% chance of being in the interval. But this interpretation is plain wrong. The correct interpretation is that, if the experiment were performed many times, and confidence intervals constructed after each iteration, 95% of the various confidence intervals would enclose the population parameter. The problem is that knowing the probability of constructing a confidence interval enclosing the population parameter is not equivalent to knowing the probability of the population parameter being in the interval that happens to have been constructed by dint of any one study. The former can be known but the latter cannot. This is a subtle inverse inference problem not dissimilar from the one we faced earlier, when discussing making an inference from the probability of the finding given a hypothesis, to the probability of hypothesis given the finding. In summary, confidence intervals do not provide the probability that the population parameter of interest is within the interval. There is no way to know this. More generally, if one wishes to use a confidence interval for parameter estimation, it is not clear what the constructed confidence interval justifies about the degree of confidence the researcher can have that the parameter of interest is within the constructed interval.

## Random and Independent Sampling

One requirement for both NHST and confidence intervals is random and independent sampling from a defined population [22]. In many cases, the population is not defined. In other cases, the population is defined, but vaguely. For example, the population might be people who suffer from anxiety. But what kind of anxiety? In what geographic location? What race or nationality? What income level? And even in the rare case where there is a reasonably precise definition of the population, there is almost never random sampling from it, where each person sampled is independent of each other person sampled. Rather, samples tend to be those people whom the researcher can get, and not a random sample where every member of the population of interest has an equal chance of being chosen to participate in the study. If the assumptions are not met, and they rarely are even reasonably approximated, then it follows that, irrespective of the other issues described earlier, the significance tests or confidence intervals the researchers use cannot be trusted.

## Conclusion

We have seen that it is logically invalid to use p-values to draw conclusions about the probabilities of hypotheses; it is desirable to avoid the modus tollens and inverse inference errors. A possible way out was to perform NHST with a threshold level, thereby controlling the Type I error rate. But this solution creates additional problems, such as inflexibility with respect to accounting for different contexts where Type I or Type II error may be more of a concern. There also is a problem with respect to the statistical phenomenon of regression to the mean. A substantial replication failure rate, for both p-values, and the effect sizes on which they are based, is a statistical inevitability due to regression to the mean. And a related statistical inevitability is the systematic overestimation of effect sizes in the published literatures. Although this is a statistical argument, it has been supported empirically by the Open Science Collaboration [14].

Nor do confidence intervals solve the problem. When used the usual way, as an alternative way to perform significance tests, confidence intervals suffer from the same problems. And when used for parameter estimation, there is no way to know the probability that the parameter of concern is within a computed confidence interval. One is forced to make an inverse inference error about this probability from the probability of constructing a confidence interval that will enclose the population parameter. Thus, conclusions about the interval estimates are blatantly unfounded in logic.

Finally, even if these problems were not so—and they are—NHST and confidence intervals depend on the assumption of random and independent sampling from a defined population. When are these assumptions met or even reasonably approximated?

So how is a worker in one of the health science areas to become educated in relevant research by reading articles in health science literatures? There is no simple answer, but the following should help. The first point is not to take any significance tests or confidence intervals too seriously. Simply ignoring t, F, X2, and so on, and the significance tests or confidence intervals associated with these statistics, should make articles in the health science literatures more readable [23]. The second point is to pay much more attention to descriptive statistics [24]. In those cases where there are meaningful measures (e.g., blood pressure), the reader should think carefully about whether the obtained difference (a) would be likely to hold up in a replication experiment and (b) warrants being considered important, even under the assumption that it would hold up. A third—but related—point is that much attention should be paid to the sample size [25]. Trafimow et al. [25,26] has demonstrated how insufficient sample sizes can importantly decrease the accuracy with which sample statistics estimate population parameters. Fourth, readers should not believe the implications of summary statements, such as “the medicine had a significant effect,” without first carefully considering the descriptive findings. The implication of the summary statement is that the researcher should believe that the medicine should be used whereas a careful perusal of the actual data might imply otherwise [27].

There is a tendency for consumers of health science literatures to employ the dichotomous thinking of “significant” versus “not significant” to decide whether to take empirical findings seriously. Although this has the large advantage of simplifying matters, it is a bad idea. Hopefully, the present article shows some of the problems with such dichotomous thinking, and will aid consumers of health science literatures in coming to more valid conclusions based on careful investigations of the data.

## References

1. Hubbard R (2016) Corrupt research: The case for reconceptualizing empirical management and social science. Sage Publications, Los Angeles, California, USA.
2. Ziliak ST, McCloskey DN (2016) The cult of statistical significance: How the standard error costs us jobs, justice, and lives. The University of Michigan Press, Ann Arbor, Michigan, USA.
3. Trafimow D, Marks M (2016) Editorial. Basic and Applied Social Psychology 38(1): 1-2.
4. Cohen J (1994) The earth is round (p<0.05). American Psychologist 49: 997-1003.
5. Trafimow D (2003) Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes’s theorem. Psychological Review 110(3): 526-535.
6. Mayo D (1996) Error and the growth of experimental knowledge. The University of Chicago Press, Chicago, USA.
7. Trafimow D, Earp BD (2017) Null hypothesis significance testing and the use of P values to control the Type I error rate: The domain problem. New Ideas in Psychology 45: 19-27.
8. Cohen J (1988) Statistical power analysis for the behavioral sciences. 2nd Edition. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey, USA.
9. Rosenthal R, Rosnow RL (1991) Essentials of Behavioral Research: Methods and Data Analysis. 2nd Edition. McGraw-Hill, Inc., New York.
10. Bakker M, van Dijk A, Wicherts JM (2012) The rules of the game called psychological science. Perspect Psychol Sci 7(6): 543-554.
11. John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23(5): 524-532.
12. Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22(11): 1359-1366.

13. Woodside A (2016) The good practices manifesto: Overcoming bad practices pervasive in current research in business. J Business Research 69(2): 365-381.
14. Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251): aac4716.
15. Grice JW (2017) Comment on Locascio's results blind manuscript evaluation proposal. Basic and Applied Social Psychology 39(5): 254-255.
16. Hyman M (2017) Can 'results blind manuscript evaluation' assuage 'publication bias'? Basic and Applied Social Psychology 39(5): 247-251.
17. Kline R (2017) Comment on Locascio, results blind science publishing. Basic and Applied Social Psychology 39(5): 256-257.
18. Locascio J (2017) Results blind publishing. Basic and Applied Social Psychology 39(5): 239-246.
19. Locascio J (2017) Rejoinder to responses to "results blind publishing. Basic and Applied Social Psychology 39(5): 258-261.
20. Marks MJ (2017) Commentary on Locascio 2017. Basic and Applied Social Psychology 39(5): 252-253.
21. Cumming G (2012) Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge, New York, USA.
22. Berk RA, Freedman DA (2003) Statistical assumptions as empirical commitments. 2nd Edition, Aldine de Gruyter, pp: 235-254.
23. Trafimow D, Marks M (2015) Editorial. Basic and Applied Social Psychology 37(1): 1-2.
24. Valentine JC, Aloe AM, Lau TS (2015) Life after NHST: How to describe your data without "p-ing" everywhere. Basic and Applied Social Psychology 37(5): 260-273.
25. Trafimow D (2017) Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. Educational and Psychological Measurement 77(2): 831-854.
26. Trafimow D, MacDonald, JA (2017) Performing inferential statistics prior to data collection. Educational and Psychological Measurement 77(2): 204-219.
27. Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician 70(2): 129-133.

***Corresponding author:** David Trafimow, Department of Psychology, New Mexico State University, P. O. Box 30001 Las Cruces, 88001-8003, New Mexico, USA, Tel: 575-646-4023; Email: dtrafimo@nmsu.edu

**Received date:** December 22, 2017; **Accepted date:** January 08, 2018; **Published date:** January 22, 2018

**Citation:** Trafimow D (2018) Education Pertaining to Understanding Common Inferential Procedures in Health Research. *J Health Sci Educ* 2(1): 123.