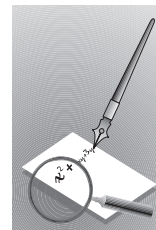


commentary and analysis



The Insignificance of Significance Testing

Abstract

Null hypothesis significance testing (NHST) is commonplace in atmospheric research, despite the many criticisms that have been leveled against its use in other fields of research. NHST is used, in papers in almost every issue of the Society's journals, to test correlations, means, and trends. Some of the major criticisms of NHST are summarized, and possible alternatives are discussed.

1. Introduction

Cohen (1994) noted that there had been trenchant criticism of null hypothesis significance testing (NHST) for four decades. These criticisms have been aired amongst statisticians, and in psychological, sociological, and medical research, and elsewhere. There have been discussions in some fields about whether NHST should be banned completely. Despite these discussions and criticisms, the use of NHST in atmospheric science remains common. Reviewers and editors sometimes insist on NHST to "test" correlations or trends. Presumably, many authors include significance tests of their results in the belief that their absence would likely prejudice reviewers or editors against publication.

I will cite a few recent atmospheric science papers of which I am a coauthor to illustrate some of the ways NHST is used. In each case, better statistical testing could probably have been adopted, but the pervasiveness of NHST biases authors toward this approach. Many other examples of NHST could have been presented here—these have been selected to demonstrate the pervasiveness of NHST in climate research. They are drawn from three different journals (not just from

this Society). These examples will appear again later, in the discussion of criticisms of NHST.

The first example (Plummer et al. 1999) documented twentieth-century trends in a variety of climate extremes over Australia and New Zealand, and subjected these to NHST, highlighting the trends that were statistically significant. The second example (Manton et al. 2001) examined trends in extreme daily temperature and rainfall over Southeast Asia and the South Pacific in the period 1961–98. Most trends were illustrated just with maps with positive or negative symbols at stations to indicate the sign of the trend, with statistically significant trends denoted by bold symbols. The final example (Frederiksen et al. 2001) examined the skill of dynamical seasonal predictions during the 1997/98 El Niño. In maps of the skill of the forecasts, areas for which the skill did not reach statistically significant levels were left blank, even if these areas had positive skill. The use of NHST in each of these papers, and in many others, could be criticized. What is the basis for such criticisms?

2. Criticisms of NHST

To set the scene, consider a typical NHST for a specific study [a more complete description of hypothesis testing will be found in most statistics textbooks, e.g., Wilks (1995)]. The research could involve calculating a correlation between, say, the Southern Oscillation index (SOI) and winter snowfall at a specific location. The researcher might establish a null hypothesis, H_0 , that there is no relationship (i.e., zero correlation) between the SOI and snowfall, if the entire population of data could be examined. The researcher would then calculate the correlation on a sample of data, for example, data from the period 1950–99. Standard tables would be used to calculate p , the probability that this correlation or a more extreme correlation would arise if a sample of the same size was drawn from a population with zero correlation. The correla-

tion would be deemed “statistically significant” if p was less than 0.05 (or possibly 0.01 or even 0.001). If p was larger than the threshold, the correlation would be labeled as “not significant.” So what is wrong with such a procedure?

First, the test is arbitrary. For historical reasons, the test is usually performed as described above, with H_0 being “rejected” if $p < 0.05$. As Rosnell and Rosenthal (1989) noted however, “...surely, God loves the .06 nearly as much as the .05.” In other words, why should a sample correlation strong enough to return $p < 0.06$, but not $p < 0.05$, be denoted as not statistically significant? The same argument applies no matter what value of p is used as the cutoff for significance. This problem arises just because of the dichotomous nature of NHST. The “blinking out” of areas on maps where some feature does not reach statistical significance is quite often seen in atmospheric science (e.g., Frederiksen et al. 2001). In other papers results that do not reach statistical significance are not displayed or listed. This method of display means that information is lost. For example, the skill in a specific area may be useful and real even if it does not reach the arbitrary 5% significance level. This problem could be avoided by reporting the actual p values, rather than using the NHST to separate statistically significant results from nonsignificant results.

It is possible that a specific effect does not reach the 5% significance level simply because of the small sample size. NHST is affected as much by the size of the sample as by the strength of the effect (e.g., the correlation) being tested. So, a correlation of 0.60 would be deemed “not significant ($p > 0.05$)” if the sample size was 10, but a correlation of 0.10 would be significant in a sample of 400. In general, a correlation of 0.60 would be considered more physically significant or useful than a correlation of 0.10. Yet the dependence of NHST on sample size means that we may overlook the strong correlation, but accept as “real” the smaller, potentially less useful correlation. An example of the sample-size problems that arise by relying on NHST comes from the early tests of the use of citrus fruit to treat and prevent scurvy (Maltz 1994). In 1747 the physician James Lind carried out an experiment by providing dietary additions to scurvy patients on the ship *Salisbury*. He used six dietary additions, one of which was citrus fruit. Lind assigned two sailors ill from scurvy to each treatment. Those who received the citrus fruit were cured in a few days and were able to help nurse the other patients. Maltz (1994) points out that this result would not have been

significant at the 5% level, and thus would probably not be accepted by a modern journal that insists on NHST. So such results would probably be overlooked because of overreliance on significance testing.

The probability of rejecting the null hypothesis, with typically sized atmospheric data sets and effects, is actually rather smaller than many realize. Cohen (1990) provides an example that is very relevant to typical climate studies. He points out that the probability of a significance test of a correlation leading to the rejection of the null hypothesis is only 57% if the population, that is, *real* correlation, is 0.30, with a two-sided 0.05 alpha criterion and a sample size of 50. That is, the use of dichotomous NHST in such cases would lead to a real (and reasonably strong) correlation being discounted as not significant in nearly half the samples examined (if multiple samples were available). Despite the reasonable sample size (in climate terms) and the reasonably strong effect, the chance of wrongly not rejecting the null hypothesis is “a coin flip” (Cohen 1990).

Even if the null hypothesis is rejected, this is rarely very informative. The null hypothesis typically asserts that there is zero correlation between the variables, or zero difference in the mean between two treatments A and B. As noted by Tukey (1991), however, “It is foolish to ask ‘Are the effects of A and B different?’ They are always different—for some decimal place.” Cohen (1994) has labeled the null hypothesis the “nil hypothesis” to illustrate the “ridiculous” nature of typical NHST. In general, we are really not interested in finding a statistically significant effect or correlation. As noted above, if we increase the sample size we will eventually find such an effect. We are more interested in physically and/or socially significant effects. Fixation on statistical significance can misdirect us from physically important processes or effects.

If the null hypothesis is not rejected, then the test is even less informative. It is often forgotten that if NHST does not lead to the rejection of the null hypothesis, this does not mean that it can be concluded that the null is true. All that can be concluded, in such a case, is that it cannot be concluded that the null hypothesis is false. “In other words, you could hardly conclude anything” (Cohen 1990). However, it is easy to fall into the trap of “accepting the null hypothesis” in such a situation.

There are also problems with the “samples” subjected to NHST in atmospheric research. At times NHST is used on a correlation or trend calculated from the entire population. One example occurred in the

Second Assessment of the Intergovernmental Panel on Climate Change (Nicholls et al. 1996), where the trends in global temperatures from surface observations, radiosondes, and the Microwave Sounding Unit satellite-based instrument were compared over the period 1979–95. These trends were different and it was important for the study to examine the causes of these differences. A reviewer pointed out, however, that NHST applied to these three series would not have led to rejection of the null hypothesis that there was zero trend. He then asserted that the trends therefore were indistinguishable from zero and each other, so the difference in trends should not have even been discussed. As described earlier, significance tests are applied to a statistic estimated from a sample taken from a population. But since all the 1979–95 data (i.e., the total population, rather than a sample) had been used to calculate the trends, then a significance test does not provide any extra information about the reality of the trends over this period. The only uncertainty would arise from measurement errors, not from sampling problems. An effect calculated using all the data for the period should not be subjected to a “significance test,” since such a test is not, in fact, a test of statistical significance. The Plummer et al. (1999) and Manton et al. (2001) studies mentioned earlier also applied NHST to trends. The more important aspect of the magnitude of the trends can be lost in the question of whether they are statistically significant.

Samples of atmospheric data can diverge from the random samples assumed in NHST in other ways. In climate studies typically all the available data will be used for calculating correlations between, for instance, the SOI and snowfall. These data will typically exhibit serial correlation, and often also skewness, bimodality, and other deviations from well-behaved distributions. These deviations and problems can be overcome, and sometimes are overcome, through appropriate transformations and taking serial correlation into account. More often than not, however, it is simply assumed that these problems are not so severe as to invalidate classic NHST.

Perhaps a more fundamental problem with the samples in the atmospheric sciences, especially in climate studies, is that classic NHST does not apply to exploratory studies (Flueck and Brown 1993). In classic NHST, the researcher specifies the null and alternative hypotheses before examining the data. In many climate studies, the alternative hypothesis (e.g., that the SOI and snowfall are correlated) arises from an initial examination of the data. These same data can-

not, then, be used as the sample for classic, a priori significance testing. Madden and Julian (1971) in their original paper on the 40–50-day oscillation note this problem with exploratory studies. Flueck and Brown point out that the “flexibility and searching nature of this type of study has dire effects on classical statistical inference statements.” Most studies searching for climate trends are of the exploratory kind (e.g., Manton et al. 2001, Plummer et al. 1999). Trends are found in the data, which are then used as the “random” sample for NHST. This exploratory approach invalidates the use of classic NHST.

A common criticism of NHST is that it is often misinterpreted. Thus many researchers believe that significance tests answer the question “Given these data and the correlation calculated with them, what is the probability that H_0 is true (i.e., that there is zero correlation between the variables)?” In fact, the p value in NHST tells us “Given that H_0 is true, what is the probability of finding a correlation this strong (or stronger)?” These questions are not the same (Cohen 1994). However, it is easy to slip into the assumption that the questions are identical, thereby misinterpreting the NHST.

Another common misinterpretation is to equate “ $1-p$ ” with the probability that the null hypothesis would be rejected if a second sample was available (i.e., the replicability of rejection of the null hypothesis). This fallacy is perhaps less common in climate work, mainly because of the difficulty of conceiving of a second sample becoming available, except after many years have passed. In fact, the probability of a second sample replicating the rejection of the null hypothesis, with the sample sizes and effect sizes typical in most research, is much lower than $1-p$. In typical cases in the behavioral sciences (with similar effect and sample sizes to climate research) a p of 0.01 would typically mean that in five replications the chances of as many as three of them being significant would only be about 50% (Cohen 1994).

Rozeboom (1960) points out that NHST is essentially unscientific. The use of NHST implies that scientists should elect to adopt one belief or another, as a result of the data analysis and testing. Carver (1978) labeled NHST as a “corrupt form of the scientific method” because of this feature. Scientific experiments should help us make an appropriate adjustment in the degree to which we believe the hypothesis being tested, rather than a simple dichotomous decision about statistical significance. Reliance on NHST can lead us to dismiss data that actually support the re-

search hypothesis, simply because it does not reach statistical significance. This can mean that such data may not be combined with other, independent tests of the research hypothesis. An example of how this may affect atmospheric research relates to anthropogenic climate change detection. Most detection studies apply NHST to a sample of data, and determine whether to reject the null hypothesis of zero trend in the atmospheric variable under consideration. The data may, in one case, be the horizontal distribution of surface temperature trends; the next case may use the vertical distribution of temperature changes. Since these are somewhat independent pieces of evidence they should both help us adjust our beliefs in the probability that human influences are affecting climate. But treating these separately and individually to NHST means that two separate and unrelated “decisions” are made, and that the two data analyses do not provide a cumulative adjustment to our belief in climate change.

Sterling (1959) pointed out that fixation on NHST may lead to a bias in published scientific results. Thus, if a research project yields nonsignificant results, it will probably not be published. This research will, therefore, remain unknown to other researchers who may repeat the investigation on a different sample. Eventually one of these repeated investigations will yield a statistically significant result, and will be published. So, in a field where NHST is dominant, the published literature will consist of false conclusions arising from errors of the “first kind,” that is, wrongly rejecting the null hypothesis.

Cohen (1990) and von Storch and Zwiers (1999) present other objections to the use of NHST. Loftus (1996) concludes his discussion of NHST with the statement that “the common belief that the precise quantity 0.05 refers to anything meaningful or interesting is illusory.” Summarizing the arguments, Falk (1986) noted that “significance tests do not provide the information that scientists need, neither do they solve the crucial questions that they are characteristically believed to answer. The one answer that they do give, is not a question that we have asked. The conclusion that this practice should be dropped altogether seems inevitable.” Although many of the problems with NHST arise because of misconceptions about what it means, these misconceptions are widespread, perhaps close to universal, and difficult to change (Falk and Greenbaum 1995). It seems unlikely that we will be able to overcome these misconceptions. So what can we do instead of NHST?

3. Alternatives to NHST

a Confidence intervals

One way to avoid some of the problems associated with NHST is to focus on the strength of the effect, rather than its statistical significance. This way we could focus further testing (a new sample, for instance, or testing in a physical model) on the strong effects rather than on weak but “significant” effects. This could be done, as well as retaining an indication of the confidence we should have in the results from the sample, by reporting confidence intervals around the effect size calculated from the sample. Gardner and Altman (1989) describe methods for calculating confidence intervals for correlations, regressions, differences of means, and a variety of other statistical measures. Wilks (1995) illustrates the computation and use of confidence intervals in atmospheric science research, and notes that, in a sense, confidence interval estimation is the inverse operation to NHST.

The reporting of confidence intervals would allow readers to address the question “Given these data and the correlation calculated with them, what is the probability that H_0 is true?” rather than “Given that H_0 is true, what is the probability of finding a correlation this strong (or stronger)?” If the interval included zero correlation then one may conclude that the evidence from the sample may not be strong enough, on its own, to determine the sign of the effect.

An important advantage of presenting a confidence interval is that it allows questions more relevant than the null hypothesis to be addressed. Reporting confidence intervals allows the testing of more useful hypotheses such as “Is there more than a weak correlation between the variables?” It includes the point estimate of the magnitude of the effect (instead of just reporting results as significant, as in, e.g., Manton et al. 2001). Other advantages of confidence intervals, as noted by Krantz (1999), are that a wide interval would reveal ignorance and thus dampen “optimism about replicability,” and that if a second sample produced a confidence interval overlapping that of the original sample, this would not be considered a failure to replicate. This would not usually be the case with NHST where a nonsignificant effect on a second sample is often considered as nonreplication. As well, confidence intervals are formally valid, do not depend on a prior hypotheses, and do not result in trivial knowledge (Brandstätter 1999).

One problem with confidence intervals is that they, like NHST, are arbitrary, in the sense that the “width”

of the interval (e.g., 95%) must be specified. In some cases this arbitrariness might be overcome, to some degree, by having more than one interval (e.g., 50% and 95%), just as the same problem with NHST can be overcome by quoting several significance levels. But confidence intervals, despite this arbitrariness, do provide specific information concealed by NHST. Thus confidence intervals clearly reveal the precision of parameter estimates, with small intervals indicating more exact estimates than larger ones; they provide information on the magnitude of the effect of interest; and they are easier to understand than significance tests (Brandstätter 1999).

b. *Permutation tests*

Confidence intervals do not overcome the problem that many atmospheric empirical datasets are not random, well-behaved samples, and in some cases are not even samples. From this point of view, confidence intervals suffer similar problems to NHST, in that intervals or significance levels are not formally valid. In such circumstances randomization tests, including bootstrap and Monte Carlo methods are a more effective way of addressing questions such as “What is the probability that H_0 is true?” Ludbrook and Dudley (1998) discuss the advantages of such tests, and the issue of the nonrandomness of samples in the atmospheric sciences is also discussed by Fluek and Brown (1993).

c. *Cross validation and a posteriori testing*

Confidence intervals also do not overcome the problems raised by the exploratory nature of much atmospheric research that invalidates the application of a priori tests such as NHST. Cross validation (e.g., Drosowsky and Chambers 2001) can partly overcome this problem, producing a more realistic estimation of the reliability of an estimate of an effect in an exploratory study. Madden and Julian (1971) point out that in some cases a posteriori tests can also be devised to overcome this problem.

4. Are there situations where NHST could be used?

Hunter (1997) points out that NHST was originally designed for “debunking” studies of the following type. Suppose a researcher has strong reason to believe that a certain theory is false (e.g., that volcanic activity “triggers” El Niño events). The researcher will use a sample of data to examine the relationship between

the variables. If the population effect size is actually zero, sampling error will still cause the observed result to differ from zero. The significance test was developed to see if the deviation from zero effect is too large to be credible. However, most uses of significance tests are not debunking studies, but in fact start from the hypothesis that there is an effect (sometimes based on earlier research by others). For such studies (labeled “confirmatory studies” by Hunter 1997), the significance tests are inappropriate because the earlier work almost guarantees that the null hypothesis is not true. That is, the earlier work or theoretical considerations suggest the existence of an effect of some magnitude—the purpose of the “confirmatory” test, then, is mainly to estimate the strength of this effect. This is more readily done with confidence intervals. NHST helps little in such studies.

One possible use of NHST in atmospheric science is where correlations (or differences in means) are displayed as a map (e.g., the correlation between the SOI and global sea surface temperatures). Here the use of confidence intervals would result in a very complex map. Perhaps in this instance shading could identify those regions where the correlations were larger than might be expected if H_0 were true. But even in this instance, such mapping should only be a preliminary to the calculation of confidence intervals around what appear to be the most interesting correlations, and the map should still display correlations even where H_0 was not rejected with NHST.

In general, however, NHST tells us little of what we need to know and is inherently misleading. We should be less enthusiastic about insisting on its use.

Acknowledgments. Bob Seaman, Lynda Chambers, Malcolm Haylock, Kevin Hennessy, Neil Plummer, Scott Power, Ian Jolliffe, and three anonymous referees kindly provided comments on earlier versions of this manuscript.

References

- Brandstätter, E., 1999: Confidence intervals as an alternative to significance testing. *Methods Psychol. Res. Online*, **4**, 33–46.
- Carver, R. P., 1978: The case against statistical significance testing. *Harvard Educ. Rev.*, **48**, 378–399.
- Cohen, J., 1990: Things I have learned (so far). *Amer. Psychol.*, **45**, 1304–1312.
- , 1994: The Earth is round ($p < .05$). *Amer. Psychol.*, **49**, 997–1003.
- Drosowsky, W., and L. E. Chambers, 2001: Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate*, in press.

- Falk, R., 1986: Misconceptions of statistical significance. *J. Struct. Learn.*, **9**, 83–96.
- , and C. W. Greenbaum, 1995: Significance tests die hard. The amazing persistence of a probabilistic misconception. *Theory Psychol.*, **5**, 75–98.
- Flueck, J. A., and T. J. Brown, 1993: Criteria and methods for performing and evaluating solar-weather studies. *J. Climate*, **6**, 373–385.
- Frederiksen, C. S., H. Zhang, R. C. Balgovind, N. Nicholls, W. Drosowsky, and L. Chambers, 2001: Dynamical seasonal forecasts during the 1997/98 ENSO using persisted SST anomalies. *J. Climate*, in press.
- Gardner, M. J., and D. G. Altman, 1989: *Statistics with Confidence—Confidence Intervals and Statistical Guidelines*. British Medical Journal, 140 pp.
- Hunter, J. E., 1997: Needed: A ban on the significance test. *Psychol. Sci.*, **8**, 3–7.
- Krantz, D. H., 1999: The Null Hypothesis testing controversy in psychology. *J. Amer. Stat. Assoc.*, **44**, 1372–1381.
- Loftus, G. R., 1996: Psychology will be a much better science when we change the way we analyze data. *Curr. Direct. Psychol. Sci.*, **5**, 161–171.
- Ludbrook, J., and H. Dudley, 1998: Why permutation tests are superior to *t* and *F* tests in biomedical research. *Amer. Stat.*, **52**, 127–132.
- Madden, R. A., and P. R. Julian, 1971: Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.*, **28**, 702–708.
- Maltz, M. D., 1994: Deviating from the mean: The declining significance of significance. *J. Res. Crime Delinquency*, **31**, 434–463.
- Manton, M. J., and Coauthors, 2001: Trends in extreme daily rainfall and temperature in Southeast Asia and the South Pacific: 1961–1998. *Int. J. Climatol.*, in press.
- Nicholls, N., G. V. Gruza, J. Jouzel, T. T. Karl, L. A. Ogallo, and D. E. Parker, 1996: Observed climate variability and change. *Climate Change 1995: The Science of Climate Change*, J. T. Houghton, et al., Eds., Cambridge University Press, 133–192.
- Plummer, N., and Coauthors, 1999: Changes in climate extremes over the Australian region and New Zealand during the twentieth century. *Climatic Change*, **42**, 183–202.
- Rosnell, R. L., and R. Rosenthal, 1989: Statistical procedures and the justification of knowledge and psychological science. *Amer. Psychol.*, **44**, 1276–1284.
- Rozeboom, W. W., 1960: The fallacy of the Null-Hypothesis Significance test. *Psychol. Bull.*, **57**, 416–428.
- Sterling, T. D., 1959: Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Amer. Stat. Assoc.*, **54**, 30–34.
- Tukey, J. W., 1991: The philosophy of multiple comparisons. *Stat. Sci.*, **6**, 100–116.
- von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

NEVILLE NICHOLLS
 BUREAU OF METEOROLOGY RESEARCH CENTRE
 MELBOURNE, VICTORIA, AUSTRALIA

