Invited Paper:

ON THE PAST AND FUTURE OF NULL HYPOTHESIS SIGNIFICANCE TESTING

DANIEL H. ROBINSON,[1] Department of Educational Psychology, University of Texas, Austin, TX 78712, USA
HOWARD WAINER, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA

Abstract: Recent criticisms of null hypothesis significance testing (NHST) have appeared in wildlife research journals (Cherry 1998; Johnson 1999; Anderson et al. 2000, 2001; Guthery et al. 2001). In this essay, we discuss these criticisms with regard to both current usage of NHST and plausible future use. We suggest that the historical use of such procedures was reasonable and that current users might spend time profitably reading some of Fisher's applied work. However, modifications to NHST, and to the interpretations of its outcomes, might better suit the needs of modern science. Our primary conclusion is that NHST most often is useful as an adjunct to other results (e.g., effect sizes) rather than as a stand-alone result. We cite some examples, however, where NHST can be profitably used alone. Last, we find considerable experimental support for a less dogmatic attitude toward the interpretation of the probability yielded from such procedures.

JOURNAL OF WILDLIFE MANAGEMENT 66(2):263–271

Key words: Bayesian statistics, effect sizes, null hypothesis testing, *P*-values, significance testing, statistical significance testing.

In nearly 300 years since its introduction by John Arbuthnot (1710), null hypothesis significance testing (NHST) has become an important tool for scientists. During the early 20th century, the founders of modern statistics (R. A. Fisher, Jerzy Neyman, and Egon Pearson) showed how to apply this tool in widely varying circumstances, often in agriculture, and nearly all in applications that were far afield from Dr. Arbuthnot's attempt to prove the existence of God. Cox (1977) termed Fisher's procedure "significance testing" to differentiate it from Neyman and Pearson's "hypothesis testing." He drew distinctions between these 2 ideas, but those distinctions are sufficiently fine that modern users lose little if we ignore them. The ability of statisticians to construct schemes that require humans to make distinctions that appear to be smaller than the threshold of comprehension for most humans is a theme we will address when we discuss α levels.

With the advantage of increasing use, practitioners became accustomed to the darker reality as the shortcomings of NHST became apparent. The reexamination of the viability of NHST was described by Anderson et al. (2000), who showed that during the past 60 years, an increasing number of articles have questioned the utility of NHST. It is revealing to notice that over the same time period Thompson's (2001) database (Fig. 1) also showed a concomitant increase in the number of articles defending the utility of NHST. In view of the breadth of the current discussion concerning the utility of NHST in wildlife research (see also Cherry 1998, Johnson 1999, Anderson et al. 2001, Guthery et al. 2001), it seems worthwhile to provide a balanced, up-to-date summary of the situations for which NHST still remains a viable tool for research as well as describing those situations for which alternative procedures seem better suited. We conclude with some recommendations for improving the practice of NHST.

Most of the criticisms of NHST focus on its misuse by researchers rather than on inherent weaknesses. Johnson (1999) claimed that misuse was an intrinsic weakness of NHST and that somehow the tool itself encourages misuse. However, Johnson, perhaps because of a well-developed sense of polite diplomacy, did not cite specific circumstances of individual scientists misusing NHST. We agree that any statistical procedure, including NHST, can be misused, but we have seen no evidence that NHST is misused any more often than any other statistical procedure. For example, the most common statistical measure, the mean, is usually inappropriate when the underlying distribution contains outliers. This is an easy mistake. For example, such an error was made by Graunt (1662) and took more than 300 years to be uncovered (Zabell 1976).

[1] E-mail: dan.robinson@mail.utexas.edu

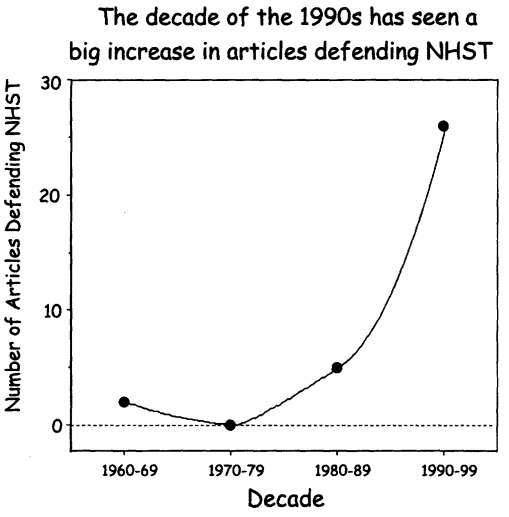

Fig. 1. Number of articles appearing in journals that have defended the utility of null hypothesis significance testing (NHST). From a database compiled by Thompson (2001).

The possibility of erroneous conclusions generated by the misuse of statistical procedures suggests several corrective alternatives. One Draconian extreme might be to ban all such procedures from use. Another approach might be to adopt the strict *caveat emptor* philosophy of the free market. Both approaches seem unnecessarily stringent. It is hard to imagine any open-minded person adopting either extreme. The former approach is not acceptable because it would essentially eliminate everything, and the latter approach fails because some quality control over scientific discourse is essential. We favor a middle path—a mixed plan that includes both enlightened standards for journal editors as well as a program to educate users of statistical procedures. This article is an attempt to contribute to that education.

Some researchers (Schmidt 1996) have felt that the misuse of NHST was sufficiently widespread to justify its ban within the journals of the American Psychological Association (APA). The APA formed a task force in 1997 to recommend appropriate statistical practices. As a small part of its deliberations, the task force briefly considered this extreme proposal (i.e., banning NHST) as well. Johnson (1999), citing Meehl (1997), surmised that the proposal to ban NHST ultimately was rejected due to the appearance of censorship, and not because the proposal was without merit. This was not the case; banning NHST was never deemed to be a credible option.

Aristotle, in his *Metaphysics*, pointed out that we understand best those things that we see grow from their very beginnings. Thus, in our summary of both the misuses and proper uses of NHST, let us begin with the original intent of 1 of its earliest modern progenitors, Sir Ronald Fisher.

FISHER'S ORIGINAL PLAN FOR NHST

Fisher understood science as a continuous and continuing process and viewed NHST in that context. He often used NHST to test the potential usefulness of agricultural innovations. He understood that scientific investigations often begin with small-scale studies whose purpose is to discover phenomena. Small-scale studies typically do not have the power to yield results of unquestioned significance. Moreover, Fisher recognized that the cost of getting rid of a false positive result was small compared to the cost of missing something that was potentially useful. He knew that if someone incorrectly found that some sort of innovation improved yields, others would quickly try to replicate. If replication repeatedly failed, the innovation would be dismissed.

Fisher (1926:504) adopted a generous α of 0.05 to screen for potentially useful innovations "and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance" (emphasis in original). He understood that if a smaller α was used—say 0.001—then less dramatic improvements would be missed and might not be rediscovered for a long time. Thus, 0.05 was used in the context of screening for innovations that would then be replicated if found to be significant. Fisher (1929:189) went on to say

> In the investigation of living beings by biological methods, statistical tests of significance are essential. Their function is to prevent us being deceived by accidental occurrences, due not to causes we wish to study, or are trying to detect, but to a combination of many other circumstances which we cannot control. An observation is judged significant, if it would rarely have been produced, in the absence of a real cause of the kind we are seeking. It is common practice to judge a result significant, if it is of such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical

> investigator, but it does not mean that he allows himself to be deceived once every twenty experiments. The test of significance only tells him what to ignore, namely all experiments in which significant results are not obtained. He should only claim that a phenomenon is experimentally demonstrable when he knows how to design an experiment so that it will rarely fail to give a significant result. Consequently, isolated significant results which he does not know how to reproduce are left in suspense pending further investigation.

There are 2 key parts of this quote—the trivial "once every twenty experiments" and the more important, "he knows how to design an experiment so that it will rarely fail to give a significant result." Fisher believed NHST made sense only in the context of a continuing series of experiments that were aimed at confirming the size and direction of the effects of specific treatments. Throughout Fisher's work, he used statistical tests to come to 1 of 3 conclusions. (1) When P was small (<0.05), he declared that an effect has been demonstrated. (2) When it was large ($P > 0.2$), he concluded that if there was an effect, it was too small to be detected with an experiment this size. (3) When it lies between these extremes, he discussed how to design the next experiment to estimate the effect better.

Null hypothesis significance testing as it is used today hardly resembles Fisher's original idea. Its critics worry that researchers too commonly interpret results that show $P > 0.05$ as evidence of no effect and rarely replicate results where $P < 0.05$ in a series of experiments designed to firmly establish the size and direction of the experimental effects. This conception is of a science built largely on single-shot studies in which researchers choose to reach conclusions based on these obviously arbitrary criteria. We should mention, however, that a strong counter-current to this conception is reflected in the Cochrane Collaboration—a database containing more than 250,000 random assignment medical experiments in which all of the included studies provide the information necessary for a meta-analysis. Such meta-analyses allow the formal concatenation of results, which then can yield more powerful inferences than would be possible from a single study. Robert Boruch, at the University of Pennsylvania, is currently organizing a parallel database for the social sciences; this effort is called the Campbell Collaboration.

We find it curious that NHST has been criticized by Anderson et al. (2000) and Johnson (1999) for using arbitrary cutoff levels when at the same time Anderson et al. (2001) recommended that authors should report the $(1 - \alpha)$ confidence level, also an arbitrary cutoff level of precision. Furthermore, we agree with Guthery et al. (2001) that even if researchers were to adopt the information–theoretic methods recommended by Anderson et al. (2001), an arbitrary numerical criterion is still used to judge the strength of evidence in single studies. The practice of basing scientific conclusions on single studies using arbitrary criteria, if widespread, could naturally give NHST, or any other method, a bad name that could be avoided if researchers simply emulated Fisher's original plan. Nevertheless, additional ways exist in which NHST can be improved still further. In the next section, we examine how NHST has been misused and/or criticized and how it might be improved or used more appropriately.

SILLY NULL HYPOTHESES

Cherry (1998), Johnson (1999), and Anderson et al. (2000, 2001) echo a common complaint (Schmidt 1996, Thompson 1996) that the typical null hypothesis is almost always false. We agree that NHST is being misused when it tests a null hypothesis in which the effect can only go in a single direction. Reporting a P-value for a correlation that was computed for reliability and validity coefficients represents vacuous information (Abelson 1997) and constitutes what Brennan (2001:172) called "excessive use of P-values." If P-values add nothing to the interpretation of results, leave them out, although sometimes a significant P-value may just be scientific shorthand for a substantial effect size. This occurs if one's reaction to a significant P-value is "if the difference is statistically significant with that small a sample, it must be a huge effect." Obviously, communicating effect size with P-value and sample size is indirect, but sometimes such shorthand aids in efficient communication.

Not all P-values, however, are unimportant. Wainer (1999) mentioned several examples of research hypotheses in which simply being able to reject the null would be a considerable contribution to science. For example, if physicists had been able to design an experiment that could reject the null hypothesis that the speed of light is equal in different reference frames that are moving at very different speeds, Albert Einstein—then a young Swiss patent clerk who sug-

gested otherwise—might have remained obscure. Nevertheless, we agree that many of the null hypotheses tested in the research literature are false only in the *statistical* sense of the word, but, as a practical matter, could be treated as if they were true with little likelihood of any negative consequences. Newtonian physics jumps to mind as 1 example of a false hypothesis that under very broad conditions could profitably be treated as true. Guthery et al. (2001) also argued that although most statistical null hypotheses are false, there are many research null hypotheses in wildlife science in which stating no effect constitutes a legitimate challenge to untested assumptions.

The probabilistic appendage to a statement such as "The foraging patterns were not the same for all months ($P < 0.05$)" seems unnecessary since everyone would agree that it is extraordinarily unlikely that 12 population means would be identical. Usually, if large enough samples are obtained, *P*-values can be made arbitrarily small.

This criticism of NHST seems to be a valid one. If the only purpose of a hypothesis test is to canonize a small difference whose size and direction are of no interest, we agree that NHST is unnecessary. Further, we generally agree with critics who suggest that it is exactly the size and direction of observed differences that should be reported, and not "naked *P*-values" (Anderson et al. 2001:374). We depart from complete agreement with such sentiments for those (admittedly more rare) circumstances in which such differences are of secondary importance (e.g., H_0: I am pregnant) and simply being able to reject the null hypothesis (or not) is what is of principal interest.

We also depart from the critics in our belief that we should modify NHST to suit our modern understanding rather than eliminate it. We discuss some plausible modifications in later sections.

THE ROLE OF EFFECT SIZES IN NHST

An ordinal claim regarding the direction of the difference or relationship can be a substantial contribution (Frick 1996). In some cases, however, knowing the direction of the effect is not sufficient in deciding whether an intervention is cost-effective. In these situations, calculating the size of the effect usually is critical. Conducting NHST does not preclude the researcher from calculating effect sizes. Whereas NHST is useful in determining statistical significance, effect sizes are useful in determining practical importance. Of course, we would prefer to see all effect sizes accompanied with a confidence interval that indicates the precision (or imprecision) with which that effect has been estimated. Nonetheless, we find the notion that one must somehow choose between conducting NHST or calculating effect sizes and confidence intervals to be absurd. Both a frying pan and butter are useful on their own, but together they can do things that neither can do alone. So, too, is the case with NHST, effect sizes, and confidence intervals. Researchers should be able to use any statistical technique that will help shed light on the interesting aspects of their data. Tukey (1969:83) recommended that "we ought to try to calculate what will help us most to understand our data, and their indications. We ought not to be bound by preconceived notions—or preconceived analyses."

Thompson (2000) reported that over the past few years, more than a dozen journals in education-related fields have instituted policies that require authors to provide effect sizes in addition to *P*-values (e.g., *Contemporary Educational Psychology, Educational and Psychological Measurement, Journal of Agricultural Education, Journal of Applied Psychology, Journal of Consulting and Clinical Psychology, Journal of Early Intervention, Journal of Experimental Education, Journal of Learning Disabilities, Language Learning, Measurement and Evaluation in Counseling and Development, The Professional Educator*, and *Research in the Schools*). The reporting of effect sizes matches 1 recommendation of the APA Task Force on Statistical Inference who suggested that authors "always present effect sizes [and] add brief comments that place these effect sizes in a practical and theoretical context" (Wilkinson et al. 1999:599). However, the most recent edition of the *APA Style Manual* stops short of recommending that authors always present effect sizes because the issue of whether such a requirement is necessary is far from being resolved.

Obviously, one should provide effect sizes or, for that matter, any other type of statistical information that yields useful insights into the characteristics of data. However, requiring authors to always provide effect size information may be overkill in those situations in which such information adds little to the correct interpretation of the data, and more dangerously, if it distracts or misleads readers. For example, a major use of NHST is in testing model fit, such as using a likelihood ratio to compare a restricted model to its more general parent. What does effect size mean in this context? Also, in some instances (e.g., medical research), it is a practical impossibility to obtain good estimates of effect size, because once

a treatment is determined to be superior, researchers are ethically forbidden from using the inferior one. This particular circumstance provides a good illustration of 2 important ideas.

First, Will Rogers' colorful caveat "What we don't know won't hurt us; it's what we do know that ain't so" has important application in hypothesis testing. Indeed, finding a significant, but inaccurate, direction of a difference or relationship has been called a Type III error by Henry Kaiser in his (1970) Psychometric Society presidential address, an idea discussed many years earlier by Wald (1947). This suggests that accompanying an effect size by a suitably small *P*-value is more than just an adornment.

The second issue worth mentioning is the question "what is the effect whose size we are reporting?" In medical research, 1 measure of a treatment's effectiveness might be the number of people who don't get the disease who would have otherwise, or the number of people cured who would not have been—in short, the causal effect of the treatment. Let us consider the ethical conundrum of trying to get a good estimate of the effect of a treatment. Obviously, we want to know the direction of the effect of the treatment, and once we know it with reasonable certainty, we are ethically bound not to use the inferior treatment. But how far can we continue with the experiment to be "sure enough?" Anscombe (1963) proposed a modification to the typical Neyman-Pearson formulation that is more in keeping with medical needs and forms a model for the flexibility of approach we support. Anscombe pointed out that we are not interested in the asymptotic probability of error; rather, he observed that for any medical treatment, there would be a finite number of patients treated. A small number of them will be treated as part of the clinical trial; the rest will be given the treatment that the clinical trial decides is "best." If we use too small a number of patients in the trial, the decision of which treatment is best is more likely to be in error. And if so, all the rest of the patients will be given the wrong treatment. If we use too many patients in the trial, then all the patients in the trial on the other treatments will have been mistreated unnecessarily. Anscombe proposed that 1 criterion of analysis, 1 "effect," should be minimizing the total number of patients (both those in the trial and those treated afterwards) who are given the poorer treatment.

Finally, some situations do not warrant obtaining a large or practical effect. W. J. McKeachie at the University of Michigan (personal communication) noted that effect sizes are mostly useful for

> ...research that is directed toward decisions with some immediate practical consequences. As I see it, much research is concerned with developing or testing theory. If it is to test an existing theory, even a small difference should increase one's confidence that the theory has some validity. Similarly if you are contributing to theory development, the size of the result is not so important as its heuristic value in stimulating thinking, which may then be tested by further research.

Thus, in some cases, researchers can or should look only for significance of direction and not necessarily effect size.

In fact, there are even very practical situations in which effect size is known in advance to be very small and only direction is of interest. Consider, for example, an application of what Box and Wilson (1951) called evolutionary operation (EVOP) in which slight variations in manufacturing procedures are tried and the direction of their effect noted (does it improve matters or make them worse?). The variations are never large because the costs of a major screw-up are too serious. If the direction of change is an improvement, then further changes of that sort are made. If things get worse, subsequent changes are made in another direction. As an example, suppose a manufacturer of paper is using EVOP and introduces experiments into the production run. The humidity, speed, sulphur, and temperature are modified slightly in various ways. The resulting change in paper strength cannot be great and still produce a salable product. Yet some of these slight modifications may yield a significant increase, which then becomes the stage for another experiment. The results of each stage in EVOP are compared with previous stages. Experiments with seemingly anomalous results are rerun. The experiments go on forever, for there is no final "correct" solution. This matches closely the scientific enterprise in which the sequence of experiments followed by examination and reexamination of data has no end. Walters (1986) has proposed a similar methodology, adaptive resource management, which has received considerable attention in wildlife and fisheries management.

Thus, it is not always necessary for authors to provide effect size information. It is worse than inappropriate if that information subsequently

misleads readers about the accuracy of the results. Anderson et al. (2001:374) argued "emphasizing estimation over hypothesis testing . . . helps protect against pitfalls associated with the failure to distinguish between statistical significance and biological significance." They went on to say that if a test yields a nonsignificant *P*-value, authors should discuss the estimated effect size and then give the estimate and a measure of precision. Unfortunately, this recommendation may be potentially dangerous if readers fail to distinguish between significant and nonsignificant effects for conclusions based on a single study. For example, suppose a small, spurious effect is reported with a confidence interval and the author goes on to discuss the size of the effect as if it were meaningful. This mismatch between the results of statistical tests and researchers' interpretations of them has been termed a Type IV error (Marascuilo and Levin 1970). We recommend that authors follow a 2-step procedure in which the likelihood of an effect (small *P*-value) is established before discussing how impressive it is (Robinson and Levin 1997).

Requiring authors to provide effect size information may also be inappropriate if that information subsequently misleads readers about the importance of the results. Robinson et al. (in press) had college students read research abstracts and found that the inclusion of effect sizes led readers to overestimate the importance of research results. We suspect that while effect sizes often are an important facet of an experiment, sometimes they may not be important and that the authors (with editorial guidance) may be the ones best suited to choose both when and what type of effect size information should be included in a research paper.

ARBITRARY α LEVELS

Both Anderson et al. (2000) and Johnson (1999) complained that NHST involves using an arbitrary cutoff point. We agree that researchers should not be bound by the chains of $\alpha = 0.05$. The fact that many persons misuse NHST by simply making reject/fail-to-reject decisions on single studies probably is due to the Neyman-Pearson legacy of such dichotomous decisions. We recommend that *P*-values should be reported as Fisher suggested. But researchers and readers still have to interpret those *P*-values. Researchers should select an α level for a statistical test a priori and explain why it was chosen. The level of α chosen should correspond to the researcher's threshold for the dismissal of the idea of chance (Alberoni 1962) for that particular null hypothesis. A person's threshold may certainly change given the stakes of the hypothesis that is tested. Fisher (1925:42) himself stated that

> ...no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.

Tukey (1969:85) discussed the potential problems of using different α levels for different contexts:

> Need we—should we—stick to $P = 0.05$ if what we seek is a relatively pure list of appearances? No matter where our cutoff comes, we will not be sure of all appearances. Might it not be better to adjust the critical *P* moderately—say to 0.03 or 0.07—whenever such a less standard value seems to offer a greater fraction of presumably real appearances among those significant at the critical *P*? We would then use different modifications for different sets of data. No one, to my knowledge, has set himself the twin problems of how to do this and how well doing this in a specific way performs.

If researchers are conducting small-scale studies that are to be included as part of a continuing series of studies, then using 0.05 as an α level seems appropriate as a screening device. However, if researchers are conducting 1-time studies that have serious consequences, much smaller α levels should certainly be used. But it is likely to be an unusual circumstance where high stakes decisions are based on a single study.

WHAT IF $P = 0.06$?

Johnson (1999) and Anderson et al. (2000) properly complain that referring to outcomes in which $P < 0.05$ as significant and where $P > 0.05$ as nonsignificant is problematic when *P*-values are close to 0.05, like 0.06. Johnson recently scanned the first few papers in Volume 65 of *The Journal of Wildlife Management* and found several examples of authors interpreting results where *P* barely exceeded 0.05 as indicating no effect. As previously noted, Fisher used the 0.05 level as a heuristic threshold because he knew that if a potentially useful treatment were discovered, someone would replicate it and show it to be use-

ful. We feel that *P*-values should be interpreted in the context of a series of experiments. If $P = 0.06$, then the researcher should ask if the effect is of potential interest to explore further. Fisher always attempted to improve the design when *P*-values were between 0.05 and 0.2.

In quantitative research, consistent small probabilities from several studies that indicate effects in the same direction are needed to conclude a direction of an effect. Statistically significant results that are replicated provide the basis of scientific truth (Tukey 1969). As for describing results where P is greater than 0.05 but still small, say less than 0.25, Tukey (1991) proposed that we might use additional words besides significant or nonsignificant to describe our reluctance to bet on the direction of the true difference or relationship. For example, if P is greater than 0.05 but less than 0.15, we could say that the difference *leans* in a certain direction. If P is greater than 0.15 but less than 0.25, we could say that there is a *hint* about the true direction. Tukey was not suggesting that we should use 0.25 as the level of significance. Rather, he was telling us to stop treating statistical testing as an all-or-nothing procedure and instead use appropriate wording to describe degrees of uncertainty.

Tukey's advice incorporates a great deal about what modern psychological investigations have told us about how humans understand probability. Modern concepts of probability began with Kolmogorov's mathematical definition of probability as a measure of sets in an abstract space of events. While all mathematical properties of probability can be derived from this definition, it is of little value in helping us to apply probability to real-life situations. Understanding how humans understand probability was helped enormously by the concept of "personal probability" that was proposed almost a half century ago by both Savage (1954) and De Finetti (1974), who contended that probability is a common concept that people can use coherently if their inferences using it follow a few simple rules. Unfortunately, in a series of ingenious experiments, the psychologists Tversky and Kahneman (1974) found no one whose probabilistic judgments met Savage's criteria for coherence. They found, instead, that most people did not have the ability even to keep a consistent view of what different numerical probabilities meant. They reported that the best humans could manage was a vastly simplified probability model (which they attribute to Patrick Suppes) that met Kolmogorov's axioms and fit their data. Suppes' model, which he presented during a discussion of an earlier version of Tversky and Kahneman's paper at the 1974 meeting of the Royal Statistical Society, has only 5 probabilities:

(1) Surely true
(2) More probable than not
(3) As probable as not
(4) Less probable than not
(5) Surely false

While Suppes' model has the benefit of fitting Tversky and Kahneman's data, it also leads to a remarkably uninteresting mathematical theory with only a few possible theorems. Indeed, if Suppes' model is, in fact, the only 1 that fits personal probability, then many of the techniques of statistical analysis that are standard practice are useless, since they serve only to produce distinctions below the level of human perception. In view of these results, Tukey's approach to interpreting *P*-values may indeed be the only sensible way to proceed; arguing about 0.04 or 0.05 or 0.06 is a poor use of one's time.

ONE EXPANDED VIEW OF NHST

Recently, Jones and Tukey (2000), expanding on an old idea (e.g., Wald 1947, Lehmann 1959), suggested a better way to interpret significant and nonsignificant *P*-values. If P is less than 0.05, researchers can conclude that the direction of a difference was determined (either the mean of group 1 is greater than the mean of group 2, or vice versa). If P is greater than 0.05, the conclusion is simply that the sign of the difference is not yet determined. This trinary decision approach (either $\mu_1 > \mu_2$, $\mu_2 > \mu_1$, or do not know yet) has the advantages of stressing that research is a continuing activity and never having to fail to reject a null hypothesis that is likely untrue. Fisher (1929:192) also commented on NHST's inability to support a null hypothesis as true:

> For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning ... it would therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data: but that they are never capable of establishing them as certainly true...

Rather than concluding that "there was no difference among the treatments ($P = 0.07$)" or that "the 2 variables were not correlated ($P = 0.06$)," authors should instead simply state that "the direction of the differences among the treatments was undetermined" or that "the sign of the correlation among the 2 variables was undetermined." This language avoids leaving the impression that the null hypothesis was accepted—or more appropriately, failed to be rejected—and suggests rather that more data are needed before a determination can be made.

CONCLUSIONS AND RECOMMENDATIONS

Null hypothesis significance testing, as currently constituted, is a tool of limited usefulness. It is useful in determining the direction of a treatment or comparative effect. It can be a valued accompaniment to effect sizes and confidence intervals by providing information about the trustworthiness of estimates of the size of the effect. It is not very useful when sample sizes are extremely large. On the other hand, effect sizes are not particularly helpful when testing model fit. In addition, accurate estimates of effect sizes sometimes are impossible to obtain, as for example, in medical research where the continued use of a control group is not ethical.

Modified versions of NHST can be used to good effect, as in tests on means with a trinary hypothesis. Such procedures have been in use for decades in sequential analysis (e.g., it's better, it's worse, or keep on testing).

Null hypothesis significance testing is well used in conjunction with a series of sequential investigations. Replicated significant results serve as the foundation of scientific justification of the direction of an effect. Replications with extensions also serve to enhance the generalizability of results while at the same time adding to the evidence for the effect. Research studies that are unique ventures are not well modeled by any statistical procedure whose goal is to predict long-term frequencies of occurrence.

Last, it has been our informal experience that many users of NHST interpret the result as the probability of the null hypothesis based on the data observed. That is, $P[H_0 \mid \text{data}]$, when formally what is actually yielded is $P[\text{data} \mid H_0]$. This error suggests that users really want to make a different kind of inference—a probabilistic statement of the likelihood of the hypothesis. To be able to make such inferences requires transforming the usual $P[\text{data} \mid H_0]$ with a straightforward application of Bayes' theorem. Bayesian hypothesis testing is reasonably well developed (Box and Tiao 1973, Novick and Jackson 1974, Winkler 1993) and well worth inclusion in the arsenal of any data analyst.

ACKNOWLEDGMENTS

This paper was collaborative in every respect and the order of authorship is alphabetical.

LITERATURE CITED

ABELSON, R. P. 1997. A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). Pages 117–141 *in* L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors. What if there were no significance tests? Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

ALBERONI, F. 1962. Contribution to the study of subjective probability. Part I. Journal of General Psychology 66:241–264.

ANDERSON, D. R., K. P. BURNHAM, AND W. L. THOMPSON. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

———, W. A. LINK, D. H. JOHNSON, AND K. P. BURNHAM. 2001. Suggestions for presenting the results of data analyses. Journal of Wildlife Management 65:373–378.

ANSCOMBE, F. 1963. Tests of goodness of fit. Journal of the Royal Statistical Society Series B 25:81–94.

ARBUTHNOT, J. 1710. An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. Philosophical Transactions of the Royal Society of London 27:186–190.

BOX, G. E. P., AND G. C. TIAO. 1973. Bayesian inference in statistical analysis. Addison-Wesley, Reading, Massachusetts, USA.

———, AND K. B. WILSON. 1951. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society Series B 135:1–45.

BRENNAN, L. 2001. Journal news. Journal of Wildlife Management 65:171–172.

CHERRY, S. 1998. Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin 26:947–953.

COX, D. R. 1977. The role of significance tests. Scandinavian Journal of Statistics 4:49–63.

DE FINETTI, B. 1974. Theory of probability. A. Machí and A. Smith, translators. Wiley, London, United Kingdom. Translation of Teoria delle probabilità.

FISHER, R. A. 1925. Theory of statistical estimation. Proceedings of the Cambridge Philosophical Society 22:700–725.

———. 1926. The arrangement of field experiments. Journal of the Ministry of Agriculture 33:503–513.

———. 1929. The statistical method in psychical research. Proceedings of the Society for Psychical Research 39:189–192.

FRICK, R. W. 1996. The appropriate use of null hypothesis testing. Psychological Methods 1:379–390.

GRAUNT, J. 1662. Natural and political observations on the bills of mortality. John Martin and James Allestry, London, United Kingdom.

GUTHERY, F. S., J. J. LUSK, AND M. J. PETERSON. 2001. The

fall of the null hypothesis: liabilities and opportunities. Journal of Wildlife Management 65:379–384.

JOHNSON, D. H. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772.

JONES, L. V., AND J. W. TUKEY. 2000. A sensible formulation of the significance test. Psychological Methods 5:411–414.

KAISER, H. 1970. A second generation little jiffy. Psychometrika 35:411–436.

LEHMANN, E. 1959. Testing statistical hypotheses. Wiley, New York, USA.

MARASCUILO, L. A., AND J. R. LEVIN. 1970. Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: the elimination of Type IV errors. American Educational Research Journal 7:397–421.

MEEHL, P. 1997. The problem is epistemology, not statistics: replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. Pages 393–425 *in* L. L. Harlow, S. A. Mulaik, and J. H. Steiger, editors. What if there were no significance tests? Lawrence Erlbaum Associates, Mahwah, New Jersey, USA.

NOVICK, M. R., AND J. E. JACKSON. 1974. Statistical methods for educational and psychological research. McGraw-Hill, New York, USA.

ROBINSON, D. H., R. T. FOULADI, N. J. WILLIAMS, AND S. J. BERA. In press. Some effects of providing effect size and "what if" information. Journal of Experimental Education.

———, AND J. R. LEVIN. 1997. Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher 26(5):21–26.

SAVAGE, L. J. 1954. The foundations of statistics. Wiley, New York, USA.

SCHMIDT, F. L. 1996. Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. Psychological Methods 1:115–129.

THOMPSON, B. 1996. AERA editorial policies regarding statistical significance testing: three suggested reforms. Educational Researcher 25(2):26–30.

———. 2000. Various editorial policies regarding statistical significance tests and effect sizes. http://www.coe.tamu.edu/~bthompson

THOMPSON, W. 2001. 36 articles/book chapters supporting (to varying degrees) use of statistical hypothesis tests. http://biology.uark.edu/Coop/thompson6.html

TUKEY, J. W. 1969. Analyzing data: sanctification or detective work. American Psychologist 24:83–91.

———. 1991. The philosophy of multiple comparisons. Statistical Science 6:98–116.

TVERSKY, A., AND D. KAHNEMAN. 1975. Judgment under uncertainty. Science 185:1124–1131.

WAINER, H. 1999. One cheer for null hypothesis significance testing. Psychological Methods 4:212–213.

WALD, A. 1947. Sequential analysis. John Wiley & Sons, New York, USA.

WALTERS, C. 1986. Adaptive management of renewable resources. Macmillan, New York, USA.

WILKINSON, L., AND TASK FORCE ON STATISTICAL INFERENCE. 1999. Statistical methods in psychology journals: guidelines and explanations. American Psychologist 54:594–604.

WINKLER, R. L. 1993. Bayesian statistics: an overview. Pages 201–232 *in* G. Keren and C. Lewis, editors. A handbook for data analysis in the behavioral sciences: statistical issues. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

ZABELL, S. 1976. Arbuthnot, Heberden and the bills of mortality. Technical Report Number 40, University of Chicago, Illinois, USA.