*Invited Paper:*

10/22/99

# THE INSIGNIFICANCE OF STATISTICAL SIGNIFICANCE TESTING

DOUGLAS H. JOHNSON,[1] U.S. Geological Survey, Biological Resources Division, Northern Prairie Wildlife Research Center, Jamestown, ND 58401, USA

***Abstract:*** Despite their wide use in scientific journals such as *The Journal of Wildlife Management*, statistical hypothesis tests add very little value to the products of research. Indeed, they frequently confuse the interpretation of data. This paper describes how statistical hypothesis tests are often viewed, and then contrasts that interpretation with the correct one. I discuss the arbitrariness of *P*-values, conclusions that the null hypothesis is true, power analysis, and distinctions between statistical and biological significance. Statistical hypothesis testing, in which the null hypothesis about the properties of a population is almost always known a priori to be false, is contrasted with scientific hypothesis testing, which examines a credible null hypothesis about phenomena in nature. More meaningful alternatives are briefly outlined, including estimation and confidence intervals for determining the importance of factors, decision theory for guiding actions in the face of uncertainty, and Bayesian approaches to hypothesis testing and other statistical practices.

***JOURNAL OF WILDLIFE MANAGEMENT 63(3):763–772***

***Key words:*** Bayesian approaches, confidence interval, null hypothesis, *P*-value, power analysis, scientific hypothesis test, statistical hypothesis test.

Statistical testing of hypotheses in the wildlife field has increased dramatically in recent years. Even more recent is an emphasis on power analysis associated with hypothesis testing (The Wildlife Society 1995). While this trend was occurring, statistical hypothesis testing was being deemphasized in some other disciplines. As an example, the American Psychological Association seriously debated a ban on presenting results of such tests in the Association's scientific journals. That proposal was rejected, not because it lacked merit, but due to its appearance of censorship (Meehl 1997).

The issue was highlighted at the 1998 annual conference of The Wildlife Society, in Buffalo, New York, where the Biometrics Working Group sponsored a half-day symposium on Evaluating the Role of Hypothesis Testing–Power Analysis in Wildlife Science. Speakers at that session who addressed statistical hypothesis testing were virtually unanimous in their opinion that the tool was overused, misused, and often inappropriate.

My objectives are to briefly describe statistical hypothesis testing, discuss common but incorrect interpretations of resulting *P*-values, mention some shortcomings of hypothesis testing, indicate why hypothesis testing is conducted, and outline some alternatives.

[1] E-mail: douglas_h_johnson@nbs.gov

## WHAT IS STATISTICAL HYPOTHESIS TESTING?

Four basic steps constitute statistical hypothesis testing. First, one develops a null hypothesis about some phenomenon or parameter. This null hypothesis is generally the opposite of the research hypothesis, which is what the investigator truly believes and wants to demonstrate. Research hypotheses may be generated either inductively, from a study of observations already made, or deductively, deriving from theory. Next, data are collected that bear on the issue, typically by an experiment or by sampling. (Null hypotheses often are developed after the data are in hand and have been rummaged through, but that's another topic.) A statistical test of the null hypothesis then is conducted, which generates a *P*-value. Finally, the question of what that value means relative to the null hypothesis is considered. Several interpretations of *P* often are made.

Sometimes *P* is viewed as the probability that the results obtained were due to chance. Small values are taken to indicate that the results were not just a happenstance. A large value of *P*, say for a test that $\mu = 0$, would suggest that the mean $\bar{x}$ actually recorded was due to chance, and $\mu$ could be assumed to be zero (Schmidt and Hunter 1997).

Other times, 1-*P* is considered the reliability of the result; that is, the probability of getting-

the same result if the experiment were repeated. Significant differences are often termed "reliable" under this interpretation.

Alternatively, $P$ can be treated as the probability that the null hypothesis is true. This interpretation is the most direct one, as it addresses head-on the question that interests the investigator.

These 3 interpretations are what Carver (1978) termed fantasies about statistical significance. None of them is true, although they are treated as if they were true in some statistical textbooks and applications papers. Small values of $P$ are taken to represent strong evidence that the null hypothesis is false, but workers demonstrated long ago (see references in Berger and Sellke 1987) that such is not the case. In fact, Berger and Sellke (1987) gave an example for which a $P$-value of 0.05 was attained with a sample of $n = 50$, but the probability that the null hypothesis was true was 0.52. Further, the disparity between $P$ and $\Pr[H_0 \mid \text{data}]$, the probability of the null hypothesis given the observed data, increases as samples become larger.

In reality, $P$ is the $\Pr[\text{observed or more extreme data} \mid H_0]$, the probability of the observed data or data more extreme, given that the null hypothesis is true, the assumed model is correct, and the sampling was done randomly. Let us consider the first 2 assumptions.

## What are More Extreme Data?

Suppose you have a sample consisting of 10 males and 3 females. For a null hypothesis of a balanced sex ratio, what samples would be more extreme? The answer to that question depends on the sampling plan used to collect the data (i.e., what stopping rule was used). The most obvious answer is based on the assumption that a total of 13 individuals was sampled. In that case, outcomes more extreme than 10 males and 3 females would be 11 males and 2 females, 12 males and 1 female, and 13 males and no females.

However, the investigator might have decided to stop sampling as soon as he encountered 10 males. Were that the situation, the possible outcomes more extreme against the null hypothesis would be 10 males and 2 females, 10 males and 1 female, and 10 males and no females. Conversely, the investigator might have collected data until 3 females were encountered. The number of more extreme outcomes then are infinite: they include 11 males and 3 females, 12 males and 3 females, 13 males and 3 females, etc. Alternatively, the investigator might have collected data until the difference between the numbers of males and females was 7, or until the difference was significant at some level. Each set of more extreme outcomes has its own probability, which, along with the probability of the result actually obtained, constitutes $P$.

The point is that determining which outcomes of an experiment or survey are more extreme than the observed one, so a $P$-value can be calculated, requires knowledge of the intentions of the investigator (Berger and Berry 1988). Hence, $P$, the outcome of a statistical hypothesis test, depends on results that were not obtained; that is, something that did not happen, and what the intentions of the investigator were.

## Are Null Hypotheses Really True?

$P$ is calculated under the assumption that the null hypothesis is true. Most null hypotheses tested, however, state that some parameter equals zero, or that some set of parameters are all equal. These hypotheses, called point null hypotheses, are almost invariably known to be false before any data are collected (Berkson 1938, Savage 1957, Johnson 1995). If such hypotheses are not rejected, it is usually because the sample size is too small (Nunnally 1960).

To see if the null hypotheses being tested in *The Journal of Wildlife Management* can validly be considered to be true, I arbitrarily selected 2 issues: an issue from the 1996 volume, the other from 1998. I scanned the results section of each paper, looking for $P$-values. For each $P$-value I found, I looked back to see what hypothesis was being tested. I made a very biased selection of some conclusions reached by rejecting null hypotheses; these include: (1) the occurrence of sheep remains in coyote (*Canis latrans*) scats differed among seasons ($P = 0.03$, $n = 467$), (2) duckling body mass differed among years ($P < 0.0001$), and (3) the density of large trees was greater in unlogged forest stands than in logged stands ($P = 0.02$). (The last is my personal favorite.) Certainly we knew before any data were collected that the null hypotheses being tested were false. Sheep remains certainly must have varied among seasons, if only between 61.1% in 1 season and 61.2% in another. The only question was whether or not the sample size was sufficient to detect the difference. Likewise, we know before data are collected that there are real differences in the other examples, which are what Abelson (1997) referred to as "gratuitous" significance testing—testing what is already known.



Three comments in favor of the point null hypothesis, such as $\mu = \mu_0$. First, while such hypotheses are virtually always false for sampling studies, they may be reasonable for experimental studies in which subjects are randomly assigned to treatment groups (Mulaik et al. 1997). Second, testing a point null hypothesis in fact does provide a reasonable approximation to a more appropriate question: is $\mu$ nearly equal to $\mu_0$ (Berger and Delampady 1987, Berger and Sellke 1987), if the sample size is modest (Rindskopf 1997). Large sample sizes will result in small $P$-values even if $\mu$ is nearly equal to $\mu_0$. Third, testing the point null hypothesis is mathematically much easier than testing composite null hypotheses, which involve noncentrality parameters (Steiger and Fouladi 1997).

The bottom line on $P$-values is that they relate to data that were not observed under a model that is known to be false. How meaningful can they be? But they are objective, at least; or are they?

## $P$ is Arbitrary

If the null hypothesis truly is false (as most of those tested really are), then $P$ can be made as small as one wishes, by getting a large enough sample. $P$ is a function of (1) the difference between reality and the null hypothesis, and (2) the sample size. Suppose, for example, that you are testing to see if the mean of a population ($\mu$) is, say, 100. The null hypothesis then is $H_0$: $\mu = 100$, versus the alternative hypothesis of $H_1$: $\mu \neq 100$. One might use Student's $t$-test, which is

$$t = \frac{(\bar{x} - 100)}{S} \times \sqrt{(n-1)},$$

where $\bar{x}$ is the mean of the sample, $S$ is the standard deviation of the sample, and $n$ is the sample size. Clearly, $t$ can be made arbitrarily large (and the $P$-value associated with it arbitrarily small) by making either $(\bar{x} - 100)$ or $\sqrt{(n-1)}$ large enough. As the sample size increases, $(\bar{x} - 100)$ and $S$ will approximately stabilize at the true parameter values. Hence, a large value of $n$ translates into a large value of $t$. This strong dependence of $P$ on the sample size led Good (1982) to suggest that $P$-values be standardized to a sample size of 100, by replacing $P$ by $P\sqrt{n/10}$ (or 0.5, if that is smaller).

Even more arbitrary in a sense than $P$ is the use of a standard cutoff value, usually denoted $\alpha$. $P$-values less than or equal to $\alpha$ are deemed significant; those greater than $\alpha$ are nonsignificant. Use of $\alpha$ was advocated by Jerzy Neyman and Egon Pearson, whereas R. A. Fisher recommended presentation of observed $P$-values instead (Huberty 1993). Use of a fixed $\alpha$ level, say $\alpha = 0.05$, promotes the seemingly nonsensical distinction between a significant finding if $P = 0.049$, and a nonsignificant finding if $P = 0.051$. Such minor differences are illusory anyway, as they derive from tests whose assumptions often are only approximately met (Preece 1990). Fisher objected to the Neyman-Pearson procedure because of its mechanical, automated nature (Mulaik et al. 1997).

## Proving the Null Hypothesis

Discourses on hypothesis testing emphasize that null hypotheses cannot be proved, they can only be disproved (rejected). Failing to reject a null hypothesis does not mean that it is true. Especially with small samples, one must be careful not to accept the null hypothesis. Consider a test of the null hypothesis that a mean $\mu$ equals $\mu_0$. The situations illustrated in Figure 1 both reflect a failure to reject that hypothesis. Figure 1A suggests the null hypothesis may well be false, but the sample was too small to indicate significance; there is a lack of power. Conversely, Figure 1B shows that the data truly were consistent with the null hypothesis. The 2 situations should lead to different conclusions about $\mu$, but the $P$-values associated with the tests are identical.

Taking another look at the 2 issues of *The Journal of Wildlife Management*, I noted a number of articles that indicated a null hypothesis was proven. Among these were (1) no difference in slope aspect of random snags ($P = 0.112$, $n = 57$), (2) no difference in viable seeds ($F_{2,6} = 3.18$, $P = 0.11$), (3) lamb kill was not correlated to trapper hours ($r_{12} = 0.50$, $P = 0.095$), (4) no effect due to month ($P = 0.07$, $n = 15$), and (5) no significant differences in survival distributions ($P$-values $\geq 0.014$!, $n$ variable). I selected the examples to illustrate null hypotheses claimed to be true, despite small sample sizes and $P$-values that were small but (usually) >0.05. All examples, I believe, reflect the lack of power (Fig. 1A) while claiming a lack of effect (Fig. 1B).

## Power Analysis

Power analysis is an adjunct to hypothesis testing that has become increasingly popular

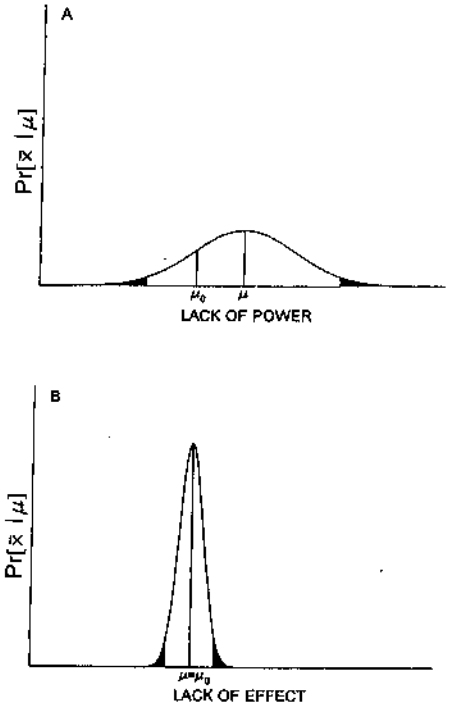


Fig. 1. Results of a test that failed to reject the null hypothesis that a mean $\mu$ equals $\mu_0$. Shaded areas indicate regions for which hypothesis would be rejected. (A) suggests the null hypothesis may well be false, but the sample was too small to indicate significance; there is a lack of power. (B) suggests the data truly were consistent with the null hypothesis.

(Peterman 1990, Thomas and Krebs 1997). The procedure can be used to estimate the sample size needed to have a specified probability (power = $1 - \beta$) of declaring as significant (at the $\alpha$ level) a particular difference or effect (effect size). As such, the process can usefully be used to design a survey or experiment (Gerard et al. 1998). Its use is sometimes recommended to ascertain the power of the test after a study has been conducted and nonsignificant results obtained (The Wildlife Society 1995). The notion is to guard against wrongly declaring the null hypothesis to be true. Such retrospective power analysis can be misleading, however. Steidl et al. (1997:274) noted that power estimated with the data used to test the null hypothesis and the observed effect size is meaningless, as a high *P*-value will invariably result in low estimated power. Retrospective power estimates may be meaningful if they are computed with effect sizes different from the observed effect size. Power analysis programs, however, assume the input values for effect and variance are known, rather than estimated, so they give misleadingly high estimates of power (Steidl et al. 1997, Gerard et al. 1998). In addition, although statistical hypothesis testing invokes what I believe to be 1 rather arbitrary parameter ($\alpha$ or *P*), power analysis requires 3 of them ($\alpha$, $\beta$, effect size). For further comments see Shaver (1993:309), who termed power analysis "a vacuous intellectual game," and who noted that the tendency to use criteria, such as Cohen's (1988) standards for small, medium, and large effect sizes, is as mindless as the practice of using the $\alpha = 0.05$ criterion in statistical significance testing. Questions about the likely size of true effects can be better addressed with confidence intervals than with retrospective power analyses (e.g., Steidl et al. 1997, Steiger and Fouladi 1997).

Table 1. Reaction of investigator to results of a statistical significance test (after Nester 1996).

| Practical importance of observed difference | Statistical significance | |
|---|---|---|
| | Not significant | Significant |
| Not important | Happy | Annoyed |
| Important | Very sad | Elated |

## Biological Versus Statistical Significance

Many authors make note of the distinction between statistical significance and subject-matter (in our case, biological) significance. Unimportant differences or effects that do not attain significance are okay, and important differences that do show up significant are excellent, for they facilitate publication (Table 1). Unimportant differences that turn out significant are annoying, and important differences that fail statistical detection are truly depressing. Recalling our earlier comments about the effect of sample size on *P*-values, the 2 outcomes that please the researcher suggest the sample size was about right (Table 2). The annoying unimportant differences that were significant indicate that too large a sample was obtained. Further, if an important difference was not significant, the investigator concludes that the sample was insufficient and calls for further research. This schizophrenic nature of the interpretation of significance greatly reduces its value.

Table 2. Interpretation of sample size as related to results of a statistical significance test.

| Practical importance of observed difference | Statistical significance | |
|---|---|---|
| | Not significant | Significant |
| Not important | *n* okay | *n* too big |
| Important | *n* too small | *n* okay |

## Other Comments on Hypothesis Tests

Statistical hypothesis testing has received an enormous amount of criticism, and for a rather long time. In 1963, Clark (1963:466) noted that it was "no longer a sound or fruitful basis for statistical investigation." Bakan (1966:436) called it "essential mindlessness in the conduct of research." The famed quality guru W. Edwards Deming (1975) commented that the reason students have problems understanding hypothesis tests is that they may be trying to think. Carver (1978) recommended that statistical significance testing should be eliminated; it is not only useless, it is also harmful because it is interpreted to mean something else. Guttman (1985) recognized that "In practice, of course, tests of significance are not taken seriously." Loftus (1991) found it difficult to imagine a less insightful way to translate data into conclusions. Cohen (1994:997) noted that statistical testing of the null hypothesis "does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does!" Barnard (1998:47) argued that ". . . simple *P*-values are not now used by the best statisticians." These examples are but a fraction of the comments made by statisticians and users of statistics about the role of statistical hypothesis testing. While many of the arguments against significance tests stem from their misuse, rather than their intrinsic values (Mulaik et al. 1997), I believe that 1 of their intrinsic problems is that they do encourage misuse.

## WHY ARE HYPOTHESIS TESTS USED?

With all the deficiencies of statistical hypothesis tests, it is reasonable to wonder why they remain so widely used. Nester (1996) suggested several reasons: (1) they appear to be objective and exact; (2) they are readily available and easily invoked in many commercial statistics packages; (3) everyone else seems to use them; (4) students, statisticians, and scientists are taught to use them; and (5) some journal editors and thesis supervisors demand them. Carver (1978) recognized that statistical significance is generally interpreted as having some relation to replication, which is the cornerstone of science. More cynically, Carver (1978) suggested that complicated mathematical procedures lend an air of scientific objectivity to conclusions. Shaver (1993) noted that social scientists equate being quantitative with being scientific. D. V. Lindley (quoted in Matthews 1997) observed that "People like conventional hypothesis tests because it's so easy to get significant results from them."

I attribute the heavy use of statistical hypothesis testing, not just in the wildlife field but in other "soft" sciences such as psychology, sociology, and education, to "physics envy." Physicists and other researchers in the "hard" sciences are widely respected for their ability to learn things about the real world (and universe) that are solid and incontrovertible, and also yield results that translate into products that we see daily. Psychologists, for 1 group, have difficulty developing tests that are able to distinguish 2 competing theories.

In the hard sciences, hypotheses are tested; that process is an integral component of the hypothetico–deductive scientific method. Under that method, a theory is postulated, which generates several predictions. These predictions are treated as scientific hypotheses, and an experiment is conducted to try to falsify each hypothesis. If the results of the experiment refute the hypothesis, that outcome implies that the theory is incorrect and should be modified or scrapped. If the results do not refute the hypothesis, the theory stands and may gain support, depending on how critical the experiment was.

In contrast, the hypotheses usually tested by wildlife ecologists do not devolve from general theories about how the real world operates. More typically they are statistical hypotheses (i.e., statements about properties of populations; Simberloff 1990). Unlike scientific hypotheses, the truth of which is truly in question, most statistical hypotheses are known a priori to be false. The confusion of the 2 types of hypotheses has been attributed to the pervasive influence of R. A. Fisher, who did not distinguish them (Schmidt and Hunter 1997).

Scientific hypothesis testing dates back at least to the 17th century: in 1620, Francis Bacon discussed the role of proposing alternative explanations and conducting explicit tests to distinguish between them as the most direct route

to scientific understanding (Quinn and Dunham 1983). This concept is related to Popperian inference, which seeks to develop and test hypotheses that can clearly be falsified (Popper 1959), because a falsified hypothesis provides greater advance in understanding than does a hypothesis that is supported. Also similar is Platt's (1964) notion of strong inference, which emphasizes developing alternative hypotheses that lead to different predictions. In such a case, results inconsistent with predictions from a hypothesis cast doubt of its validity.

Examples of scientific hypotheses, which were considered credible, include Copernicus' notion $H_A$: the Earth revolves around the sun, versus the conventional wisdom of the time, $H_0$: the sun revolves around the Earth. Another example is Fermat's last theorem, which states that for integers $n$, $X$, $Y$, and $Z$, $X^n + Y^n = Z^n$ implies $n \leq 2$. Alternatively, a physicist may make specific predictions about a parameter based on a theory, and the theory is provisionally accepted only if the outcomes are within measurement error of the predicted value, and no other theories make predictions that also fall within that range (Mulaik et al. 1997). Contrast these hypotheses, which involve phenomena in nature, with the statistical hypotheses presented in *The Journal of Wildlife Management*, which were mentioned above, and which involve properties of populations.

Rejection of a statistical hypothesis would constitute a piece of evidence to be considered in deciding whether or not to reject a scientific hypothesis (Simberloff 1990). For example, a scientific hypothesis might state that clutch sizes of birds increase with the age of the bird, up to some plateau. That idea would generate a hypothesis that could be tested statistically within a particular population of birds. A single such test, regardless of its *P*-value, would little affect the credibility of the scientific hypothesis, which is far more general. A related distinction is that scientific hypotheses are global, applying to all of nature, while statistical hypotheses are local, applying to particular systems (Simberloff 1990).

Why do we wildlife ecologists rarely test scientific hypotheses? My view is that we are dealing with systems more complex than those faced by physicists. A saying in ecology is that everything is connected to everything else. (In psychology, "everything correlates with everything," giving rise to what David Lykken called the "crud factor" for such ambient correlation noise [Meehl 1997]). This saying implies that all variables in an ecological system are intercorrelated, and that any null hypothesis postulating no effect of a variable on another will in fact be false; a statistical test of that hypothesis will be rejected, as long as the sample is sufficiently large. This line of reasoning does not denigrate the value of experimentation in real systems; ecologists should seek situations in which variables thought to be influential can be manipulated and the results carefully monitored (Underwood 1997). Too often, however, experimentation in natural systems is very difficult if not impossible.

## REPLICATION

Replication is a cornerstone of science. If results from a study cannot be reproduced, they have no credibility. Scale is important here. Conducting the same study at the same time but at several different sites and getting comparable results is reassuring, but not nearly so convincing as having different investigators achieve similar results using different methods in different areas at different times. R. A. Fisher's idea of solid knowledge was not a single extremely significant result, but rather the ability of repeatedly getting results significant at 5% (Tukey 1969). Shaver (1993:304) observed that "The question of interest is whether an effect size of a magnitude judged to be important has been consistently obtained across valid replications. Whether any or all of the results are statistically significant is irrelevant." Replicated results automatically make statistical significance testing unnecessary (Bauernfeind 1968).

Individual studies rarely contain sufficient information to support a final conclusion about the truth or value of a hypothesis (Schmidt and Hunter 1997). Studies differ in design, measurement devices, samples included, weather conditions, and many other ways. This variability among studies is more pervasive in ecological situations than in, for example, the physical sciences (Ellison 1996). To have generality, results should be consistent under a wide variety of circumstances. Meta-analysis provides some tools for combining information from repeated studies (e.g., Hedges and Olkin 1985) and can reduce dependence on significance testing by examining replicated studies (Schmidt and Hunter 1997). Meta-analysis can be dangerously misleading, however, if nonsignificant results or results that did not conform to the conventional wisdom were less likely to have been published.



## WHAT ARE THE ALTERNATIVES?

What should we do instead of testing hypotheses? As Quinn and Dunham (1983) pointed out, it is more fruitful to determine the relative importance to the contributions of, and interactions between, a number of processes. For this purpose, estimation is far more appropriate than hypothesis testing (Campbell 1992). For certain other situations, decision theory is an appropriate tool. For either of these applications, as well as for hypothesis testing itself, the Bayesian approach offers some distinct advantages over the traditional methods. These alternatives are briefly outlined below. Although the alternatives will not meet all potential needs, they do offer attractive choices in many frequently encountered situations.

### Estimates and Confidence Intervals

Four decades ago, Anscombe (1956) observed that statistical hypothesis tests were totally irrelevant, and that what was needed were estimates of magnitudes of effects, with standard errors. Yates (1964) indicated that "The most commonly occurring weakness in the application of Fisherian methods is undue emphasis on tests of significance, and failure to recognize that in many types of experimental work estimates of the treatment effects, together with estimates of the errors to which they are subject, are the quantities of primary interest." Further, because wildlife ecologists want to influence management practices, Johnson (1995) noted that, "If ecologists are to be taken seriously by decision makers, they must provide information useful for deciding on a course of action, as opposed to addressing purely academic questions." To enforce that point, several education and psychological journals have adopted editorial policies requiring that parameter estimates accompany any *P*-values be presented (McLean and Ernest 1998).

Ordinary confidence intervals provide more information than do *P*-values. Knowing that a 95% confidence interval includes zero tells one that, if a test of the hypothesis that the parameter equals zero is conducted, the resulting *P*-value will be >0.05. A confidence interval provides both an estimate of the effect size and a measure of its uncertainty. A 95% confidence interval of, say, (−50, 300) suggests the parameter is less well estimated than would a confidence interval of (120, 130). Perhaps surprisingly, confidence intervals have a longer history than statistical hypothesis tests (Schmidt and Hunter 1997).

With its advantages and longer history, why have confidence intervals not been used more than they have? Steiger and Fouladi (1997) and Reichardt and Gollob (1997) posited several explanations: (1) hypothesis testing has become a tradition; (2) the advantages of confidence intervals are not recognized; (3) there is some ignorance of the procedures available; (4) major statistical packages do not include many confidence interval estimates; (5) sizes of parameter estimates are often disappointingly small, even though they may be very significantly different from zero; (6) the wide confidence intervals that often result from a study are embarrassing; (7) some hypothesis tests (e.g., chi-square contingency table) have no uniquely defined parameter associated with them; and (8) recommendations to use confidence intervals often are accompanied by recommendations to abandon statistical tests altogether, which is unwelcome advice. These reasons are not valid excuses for avoiding confidence intervals in lieu of hypothesis tests in situations for which parameter estimation is the objective.

### Decision Theory

Often experiments or surveys are conducted to help make some decision, such as what limits to set on hunting seasons, if a forest stand should be logged, or if a pesticide should be approved. In those cases, hypothesis testing is inadequate, for it does not take into consideration the costs of alternative actions. Here a useful tool is statistical decision theory: the theory of acting rationally with respect to anticipated gains and losses, in the face of uncertainty. Hypothesis testing generally limits the probability of a Type I error (rejecting a true null hypothesis), often arbitrarily set at $\alpha = 0.05$, while letting the probability of a Type II error (accepting a false null hypothesis) fall where it may. In ecological situations, however, a Type II error may be far more costly than a Type I error (Toft and Shea 1983). As an example, approving a pesticide that reduces the survival rate of an endangered species by 5% may be disastrous to that species, even if that change is not statistically detectable. As another, continued overharvest in marine fisheries may result in the collapse of the ecosystem even while statistical tests are unable to reject the null hypothesis that fishing has no effect (Dayton 1998). Details on

decision theory can be found in DeGroot (1970), Berger (1985), and Pratt et al. (1995).

### Model Selection

Statistical tests can play a useful role in diagnostic checks and evaluations of tentative statistical models (Box 1980). But even for this application, competing tools are superior. Information criteria, such as Akaike's, provide objective measures for selecting among different models fitted to a dataset. Burnham and Anderson (1998) provided a detailed overview of model selection procedures based on information criteria. In addition, for many applications it is not advisable to select a "best" model and then proceed as if that model was correct. There may be a group of models entertained, and the data will provide different strength of evidence for each model. Rather than basing decisions or conclusions on the single model most strongly supported by the data, one should acknowledge the uncertainty about the model by considering the entire set of models, each perhaps weighted by its own strength of evidence (Buckland et al. 1997).

### Bayesian Approaches

Bayesian approaches offer some alternatives preferable to the ordinary (often called frequentist, because they invoke the idea of the long-term frequency of outcomes in imagined repeats of experiments or samples) methods for hypothesis testing, as well as for estimation and decision-making. Space limitations preclude a detailed review of the approach here; see Box and Tiao (1973), Berger (1985), and Carlin and Louis (1996) for longer expositions, and Schmitt (1969) for an elementary introduction.

Sometimes the value of a parameter is predicted from theory, and it is more reasonable to test whether or not that value is consistent with the observed data than to calculate a confidence interval (Berger and Delampady 1987, Zellner 1987). For testing such hypotheses, what is usually desired (and what is sometimes believed to be provided by a statistical hypothesis test) is $\Pr[H_0 \mid \text{data}]$. What is obtained, as pointed out earlier, is $P = \Pr[\text{observed or more extreme data} \mid H_0]$. Bayes' theorem offers a formula for converting between them.

$$\Pr[H_0 \mid \text{data}] = \frac{\Pr[\text{data} \mid H_0]\Pr[H_0]}{\Pr[\text{data}]}.$$

This is an old (Bayes 1763) and well-known theorem in probability. Its use in the present situation does not follow from the frequentist view of statistics, which considers $\Pr[H_0]$ as unknown, but either zero or 1. In the Bayesian approach, $\Pr[H_0]$ is determined before data are gathered; it is therefore called the prior probability of $H_0$. $\Pr[H_0]$ can be determined either subjectively (what is your prior belief about the truth of the null hypothesis?) or by a variety of objective means (e.g., Box and Tiao 1973, Carlin and Louis 1996). The use of subjective probabilities is a major reason that Bayesian approaches fell out of favor: science must be objective! (The other main reason is that Bayesian calculations tend to get fairly heavy, but modern computer capabilities can largely overcome this obstacle.)

Briefly consider parameter estimation. Suppose you want to estimate a parameter $\theta$. Then replacing $H_0$ by $\theta$ in the above formula yields

$$\Pr[\theta \mid \text{data}] = \frac{\Pr[\text{data} \mid \theta]\Pr[\theta]}{\Pr[\text{data}]},$$

which provides an expression that shows how initial knowledge about the value of a parameter, reflected in the prior probability function $\Pr[\theta]$, is modified by data obtained from a study, $\Pr[\text{data} \mid \theta]$, to yield a final probability function, $\Pr[\theta \mid \text{data}]$. This process of updating beliefs leads in a natural way to adaptive resource management (Holling 1978, Walters 1986), a recent favorite topic in wildlife science (e.g., Walters and Green 1997).

Bayesian confidence intervals are much more natural than their frequentist counterparts. A frequentist 95% confidence interval for a parameter $\theta$, denoted $(\theta_L, \theta_U)$, is interpreted as follows: if the study were repeated an infinite number of times, 95% of the confidence intervals that resulted would contain the true value $\theta$. It says nothing about the particular study that was actually conducted, which led Howson and Urbach (1991:373) to comment that "statisticians regularly say that one can be '95 per cent confident' that the parameter lies in the confidence interval. They never say why." In contrast, a Bayesian confidence interval, sometimes called a credible interval, is interpreted to mean that the probability that the true value of the parameter lies in the interval is 95%. That statement is much more natural, and is what people think a confidence interval is, until they get the notion drummed out of their heads in statistics courses.

For decision analysis, Bayes' theorem offers a very logical way to make decisions in the face of uncertainty. It allows for incorporating beliefs, data, and the gains or losses expected from possible consequences of decisions. See Wolfson et al. (1996) and Ellison (1996) for recent overviews of Bayesian methods with an ecological orientation.



## CONCLUSIONS

Editors of scientific journals, along with the referees they rely on, are really the arbiters of scientific practice. They need to understand how statistical methods can be used to reach sound conclusions from data that have been gathered. It is not sufficient to insist that authors use statistical methods—the methods must be appropriate to the application. The most common and flagrant misuse of statistics, in my view, is the testing of hypotheses, especially the vast majority of them known beforehand to be false.

With the hundreds of articles already published that decry the use of statistical hypothesis testing, I was somewhat hesitant about writing another. It contains nothing new. But still, reading *The Journal of Wildlife Management* makes me realize that the message has not really reached the audience of wildlife biologists. Our work is important, so we should use the best tools we have available. Rarely, however, is that tool statistical hypothesis testing.

## ACKNOWLEDGMENTS

W. L. Thompson and C. A. Ribic deserve thanks for organizing the symposium that prompted this article. I appreciate the encouragement and comments on the manuscript provided by D. R. Anderson, J. O. Berger, D. L. Larson, M. R. Nester, W. E. Newton, T. L. Shaffer, S. L. Sheriff, B. Thompson, and C. C. White, who nonetheless remain blameless for any misinterpretations contained herein. B. R. Euliss assisted with the preparation of the manuscript.

## LITERATURE CITED

ABELSON, R. P. 1997. A retrospective on the significance test ban on 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.

ANSCOMBE, F. J. 1956. Discussion on Dr. David's and Dr. Johnson's paper. Journal of the Royal Statistical Society 18:24–27.

BAKAN, D. 1966. The test of significance in psychological research. Psychological Bulletin 66:423–437.

BARNARD, G. 1998. Pooling probabilities. New Scientist 157:47.

BAUERNFEIND, R. H. 1968. The need for replication in educational research. Phi Delta Kappan 50: 126–128.

BAYES, T. 1763. An essay toward solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society, London 53:370–418.

BERGER, J. O. 1985. Statistical decision theory and Bayesian analysis. Springer-Verlag, Berlin, Germany.

———, AND D. A. BERRY. 1988. Statistical analysis and illusion of objectivity. American Scientist 76: 159–165.

———, AND M. DELAMPADY. 1987. Testing precise hypotheses. Statistical Science 2:317–352.

———, AND T. SELLKE. 1987. Testing a point null hypothesis: the irreconcilability of $P$ values and evidence. Journal of the American Statistical Association 82:112–122.

BERKSON, J. 1938. Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association 33:526–542.

BOX, G. E. P. 1980. Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society 143:383–430.

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

BUCKLAND, S. T., K. P. BURNHAM, AND N. H. AUGUSTIN. 1997. Model selection: an integrated part of inference. Biometrics 53:603–618.

BURNHAM, K. P., AND D. R. ANDERSON. 1998. Model selection and inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

CAMPBELL, M. 1992. Confidence intervals. Royal Statistical Society News and Notes 18:4–5.

CARLIN, B. P., AND T. A. LOUIS. 1996. Bayes and empirical Bayes methods for data analysis. Chapman & Hall, London, United Kingdom.

CARVER, R. P. 1978. The case against statistical significance testing. Harvard Educational Review 48:378–399.

CLARK, C. A. 1963. Hypothesis testing in relation to statistical methodology. Review of Educational Research 33:455–473.

COHEN, J. 1988. Statistical power analysis for the behavioral sciences. Second edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

———. 1994. The earth is round ($p < .05$). American Psychologist 49:997–1003.

DAYTON, P. K. 1998. Reversal of the burden of proof in fisheries management. Science 279:821–822.

DEGROOT, M. H. 1970. Optimal statistical decisions. McGraw-Hill, New York, New York, USA.

DEMING, W. E. 1975. On probability as a basis for action. American Statistician 29:146–152.

ELLISON, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6:1036–1046.

GERARD, P. D., D. R. SMITH, AND G. WEERAKKODY.

1998. Limits of retrospective power analysis. Journal of Wildlife Management 62:801–807.

GOOD, I. J. 1982. Standardized tail-area probabilities. Journal of Statistical Computation and Simulation 16:65–66.

GUTTMAN, L. 1985. The illogic of statistical inference for cumulative science. Applied Stochastic Models and Data Analysis 1:3–10.

HEDGES, L. V., AND I. OLKIN. 1985. Statistical methods for meta-analysis. Academic Press, New York, New York, USA.

HOLLING, C. S., editor. 1978. Adaptive environmental assessment and management. John Wiley & Sons, Chichester, United Kingdom.

HOWSON, C., AND P. URBACH. 1991. Bayesian reasoning in science. Nature 350:371–374.

HUBERTY, C. J. 1993. Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. Journal of Experimental Education 61:317–333.

JOHNSON, D. H. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998–2000.

LOFTUS, G. R. 1991. On the tyranny of hypothesis testing in the social sciences. Contemporary Psychology 36:102–105.

MATTHEWS, R. 1997. Faith, hope and statistics. New Scientist 156:36–39.

MCLEAN, J. E., AND J. M. ERNEST. 1998. The role of statistical significance testing in educational research. Research in the Schools 5:15–22.

MEEHL, P. E. 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.

MULAIK, S. A., N. S. RAJU, AND R. A. HARSHMAN. 1997. There is a time and a place for significance testing. Pages 65–115 *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.

NESTER, M. R. 1996. An applied statistician's creed. Applied Statistics 45:401–410.

NUNNALLY, J. C. 1960. The place of statistics in psychology. Educational and Psychological Measurement 20:641–650.

PETERMAN, R. M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences 47:2–15

PLATT, J. R. 1964. Strong inference. Science 146:347–353.

POPPER, K. R. 1959. The logic of scientific discovery. Basic Books, New York, New York, USA.

PRATT, J. W., H. RAIFFA, AND R. SCHLAIFER. 1995. Introduction to statistical decision theory. MIT Press, Cambridge, Massachusetts, USA.

PREECE, D. A. 1990. R. A. Fisher and experimental design: a review. Biometrics 46:925–935.

QUINN, J. F., AND A. E. DUNHAM. 1983. On hypothesis testing in ecology and evolution. American Naturalist 122:602–617.

REICHARDT, C. S, AND H. F. GOLLOB. 1997. When confidence intervals should be used instead of statistical tests, and vice versa. Pages 259–284 *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.

RINDSKOPF, D. M. 1997. Testing "small," not null, hypotheses: classical and Bayesian approaches. Pages 319–332 *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.

SAVAGE, I. R. 1957. Nonparametric statistics. Journal of the American Statistical Association 52:331–344.

SCHMIDT, F. L., AND J. E. HUNTER. 1997. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. Pages 37–64 *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.

SCHMITT, S. A. 1969. Measuring uncertainty: an elementary introduction to Bayesian statistics. Addison-Wesley, Reading, Massachusetts, USA.

SHAVER, J. P. 1993. What statistical significance testing is, and what it is not. Journal of Experimental Education 61:293–316.

SIMBERLOFF, D. 1990. Hypotheses, errors, and statistical assumptions. Herpetologica 46:351–357.

STEIDL, R. J., J. P. HAYES, AND E. SCHAUBER. 1997. Statistical power analysis in wildlife research. Journal of Wildlife Management 61:270–279.

STEIGER, J. H., AND R. T. FOULADI. 1997. Noncentrality interval estimation and evaluation of statistical models. Pages 221–257 *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.

THE WILDLIFE SOCIETY. 1995. Journal news. Journal of Wildlife Management 59:196–198.

THOMAS, L., AND C. J. KREBS. 1997. Technological tools. Bulletin of the Ecological Society of America 78:126–139.

TOFT, C. A., AND P. J. SHEA. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. American Naturalist 122: 618–625.

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

UNDERWOOD, A. J. 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge, United Kingdom.

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

———, AND R. GREEN. 1997. Valuation of experimental management options for ecological systems. Journal of Wildlife Management 61:987–1006.

WOLFSON, L. J., J. B. KADANE, AND M. J. SMALL. 1996. Bayesian environmental policy decisions: two case studies. Ecological Applications 6:1056–1066.

YATES, F. 1964. Sir Ronald Fisher and the design of experiments. Biometrics 20:307–321.

ZELLNER, A. 1987. Comment. Statistical Science 2: 339–341.