WHAT SHOULD WE DO ABOUT HYPOTHESIS TESTING?

L. L. EBERHARDT, 1 2528 West Klamath, Kennewick, WA 99336, USA

Abstract: A sizable list of recently published papers criticized the use of hypothesis testing in ecology and wildlife management. I suggest that early neglect of statistical methods in those areas led to excesses, and that the real problem lies in the use of methods designed for controlled experiments in uncontrolled settings. Controlled experiments can be conducted in wildlife research, and the revolutionary use of controlled experiments in medical research indicates their utility and need. Some alternative paradigms (sampling, modeling, population analysis) are less troubled by the hypothesis-testing issues. The recent interest in a model-selection paradigm is warranted but should not be considered a universal solution.

JOURNAL OF WILDLIFE MANAGEMENT 67(2):241–247

Key words: experiments, hypothesis testing, model selection, modeling, population analysis, sampling.

A spate of recently published papers (Cherry 1998; Johnson 1999, 2002a; Anderson et al. 2000, 2001; Guthery et al. 2001; Robinson and Wainer 2002) raises questions about hypothesis testing in ecology and wildlife management, objecting to the use of “naked P-values,” “silly nulls,” and the limited value of the tests presented. Anderson et al. (2001) stated that “we estimate that there have been a minimum of several thousand P-values . . . in every volume of Ecology and JWM [The Journal of Wildlife Management] . . . in recent years.” Most of the complaints are justified and require attention. Signs indicate that at least The Journal of Wildlife Management is attempting to rectify the situation (Brennan 2001). But doing so requires some attention to root causes, and the list of papers given above seems to inadequately consider 2 such causes.

The first cause to consider is that 30 or 40 years ago, statistical methods and tests were largely neglected in both Ecology and JWM. As more people with adequate training in statistical methodology became involved as editors and referees, the pendulum swung the other way, so that the use of statistical methods began to be required. As usually happens, we went overboard, resulting in excessive use of these methods, exemplified for me by an associate editor’s instruction that “you must give a statistical test.”

The other cause is much more important, and perhaps not so readily defined. If one looks carefully at the various statements about silly nulls in the papers listed above, many of these reference papers in what has been called the “soft” sciences (i.e., those not involved with designed experimentation). Briefly stated, we utilize tests developed for experiments that use random assignment of treatments (and controls) to test plots or organisms in circumstances lacking both treatments and randomization. I found 3 quotes in the papers listed above that mention the underlying problem. Cherry (1998:951) pointed out that “many of the statistical methodologies . . . were developed for analysis of data from randomized controlled experiments. . . .” Johnson (1999:765) stated that “while such hypotheses are virtually always false for sampling studies, they may be reasonable for experimental studies.” Anderson et al. (2000:914) remarked that “The null distribution of the test statistic . . . may closely match the actual sampling distribution of that statistic in strict experiments but this property does not hold in observational studies. In these latter studies the distribution of the test statistic is unknown because randomization was not done. . . .”

The immediate issue is how to present useful and sensible results from field studies. A number of paradigms exist in which particular statistical techniques are justified and are based on solid precedents. I will list the more prominent schemes in the following sections. However, another misapprehension needs to be considered. One common theme in the papers that raise doubts about present-day hypothesis testing is that the problems can be solved by simply replicating studies, so that one might not need to do any hypothesis testing, given enough replicated studies from which “truth” can be deduced. This notion may be both fallacious and dangerous for observational studies. If one does a good job of planning such a study and uses respectable sample sizes, replicating the study could lead to virtually the same inferences, thus perpetuating a possibly faulty thesis. Because observational studies are not based on randomized selection of

1 E-mail: leberhardt@aol.com

241
treatments and controls, no rigorous analysis of the circumstances is possible.

Perhaps the best short text on the “Planning and analysis of observational studies” is that by Cochran (1983), an author known for his many contributions to both the design and analysis of experiments and to sampling methodology. So the problems in using observational studies are not new.

Cochran (1983) restricted attention to studies that concentrate “on a small number of procedures, programs or treatments, often only one, and takes one or more response measurements in order to estimate the effects.” Cochran (1983) listed the wearing of seat belts and smoking as well-known examples. In those and similar cases, repetitive studies likely can be trusted to eventually ascertain causal relationships. Very likely there is no universal solution, except that each member of a sequence of studies should be designed to try to test inferences drawn from the previous results, and replications of design may be most interesting in observational studies only if they do not yield the same inferences. Johnson (2002b) discussed replication in detail. It needs to be emphasized that this criticism applies only to observational studies. Replication and randomization supply the foundations for true experiments.

STATISTICAL HYPOTHESES VERSUS
SCIENTIFIC OR RESEARCH
HYPOTHESES

Several papers in the list above attempt to distinguish between “statistical” and “scientific” or “research” hypothesis, with the implication that statistical hypotheses constitute some real and different entity. Statistical methods provide tools for analyzing data and making inferences about underlying processes from observations. Theory and assumptions may be used to construct theoretical frequency distributions or probability models that can serve to test a specified hypothesis, but there is little to suggest that the hypotheses are statistical in nature. The scientist or researcher proposes the hypothesis and uses statistical methods to test it.

We would, of course, prefer to test the research hypothesis directly, but this may not be feasible, and we either study components or, too often, indulge in lengthy discussions of the data with no actual statistical tests performed (Eberhardt 1988). As implied by Eberhardt and Thomas (1991), Romesburg’s (1981) proposal to use the hypothetico-deductive method is not suitable for “noisy” data, in which the investigator may easily go off on the wrong course by chance alone. A more flexible viewpoint such as that proposed by Tukey (1960) is needed (the hypothetico-deductive method essentially assumes that every experiment gives an exact “yes” or “no” answer with no scope for Type I and II errors). Research hypotheses are often modified as time goes on and more data are gathered, so a continuing interplay between the research hypothesis and statistical testing is necessary.

THE EXPERIMENTAL APPROACH

The paradigm encompassing the design and analysis of experiments is well known and covers too much ground to be reviewed here. A few points are relevant to the papers listed in the Introduction. One of these is the statement that a hypothesis such as $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ is obviously false and “silly.” Given random assignments of treatments and a simple transformation such as $\mu_i - \mu$ (where $\mu$ is the overall mean), $H_0$ becomes the basis for the experimental approach and a wide range of statistical tests, with the analysis of variance providing perhaps the most commonly used example. The papers eliciting the present commentary mostly did not list specific statistical tests, but analysis of variance (ANOVA) is the most widely used approach.

Neyman-Pearson theory underlies most of the current testing procedures (Bayesian methods are not considered here, as they largely did not enter into the papers identified in the Introduction). Under Neyman-Pearson theory, one needs to identify a suitable alternative hypothesis to create a test of significance. Using a null hypothesis of no effect makes it possible to identify a probability distribution for the random variable used in a test. Common examples are $t$-tests and $F$-tests. If the alternative hypothesis is something other than “no effect,” then things get very complicated. Hence, the postulation of what seem to be some silly nulls. Power calculations are constructed through the use of “non-central” $t$- and $F$-distributions, but these mainly serve to help decide in advance of the study on the magnitude of an effect that might be detected with a given sample size (as remarked by some of the authors of the list of papers given in my Introduction, power calculations should not be used after a study is completed). Hence, I do not believe that we can entirely ban silly nulls, as they are often needed as a preliminary to estimating an effect.

Silly nulls may escape notice in an ANOVA because the $F$-test does establish that some significant effects are somewhere in the treatments uti-
lized in the study as compared to controls. Perhaps this is why ANOVA has been so popular in ecological studies. However, finding out just which effects are important and should be studied further may be difficult. Multiple comparisons have been used for this purpose. Scheffe’s S- test and Tukey’s T-test (Scheffe 1959, Tukey 1960) can be used to distinguish significant “contrasts” among treatments, but these tests are only valid in “fixed-effects” models (i.e., by restricting the analysis to only those treatments examined in the study). In most wildlife work, we deal with random-effects models for which the data-snooping methods are not appropriate.

RESEARCH IN MEDICINE VERSUS WILDLIFE RESEARCH

Robinson and Wainer (2002) discussed the Cochrane Collaboration, a database “containing more than 250,000 random assignment medical experiments . . .,” while Johnson (2002a) noted that “wildlifers can only envy databases like the Cochrane Collaboration. . . .” It is important to note that the notion of “random assignment” studies in medical research is a relatively recent development. Not long ago, ethical considerations were believed to prohibit the use of untreated or placebo controls in any substantial way. However, the need for experiments with controls in medical science eventually became apparent, and now many results are available. Wildlifers should take note. Many of the areas we need to investigate may not be amenable to observational studies, and we may well have to utilize designed experiments, with replication. Certainly the decision to do controlled experiments on humans was not easily reached, and we should seriously consider whether controlled experiments also are necessary in wildlife science.

The need for bona fide experimentation in ecology and wildlife management has long been apparent but seldom achieved in any substantial way. Hurlbert’s (1984) essay on pseudoreplication received a lot of attention for quite awhile and became a favorite reference for reviewers. Eberhardt and Thomas (1991) pointed out that researchers mostly observe nature by 1 means of sampling or another, and Cochran’s (1983) small book on planning and analysis of observational studies remains a valuable reference that defines the issues very well. I believe that 1 reason (other than cost) wildlife managers do not use designed experiments is that errors such as overharvesting usually can be rectified by tightening up regula-
tions for a few years. However, overharvesting in commercial fisheries has had some very serious and possibly permanent consequences. Much-improved monitoring (e.g., for waterfowl) has substantially helped to define needs and has included some experimental regulation. By and large, though, we simply do not perform controlled experiments, and I believe that we still avoid the issue of the need for experimentation in part by focusing attention on silly null hypotheses.

The sizable number of management units now extant in most states seems to provide a convenient opportunity to conduct true experiments for management issues (Eberhardt 1988). Some managers may not believe that controlled experimentation is worth the trouble, but some issues require an experimental approach. One example is the impact of predation by wolves and bears on moose. Gasaway et al. (1992) collected an impressive amount of empirical data that suggest a substantial impact, and Boutin (1992) outlined some prospective experiments. Very possibly his specific suggestions could be revised in light of more recent investigations, and the National Research Council (1997) has nicely summarized the relevant and recent studies on the issue. Because hunting by humans is involved, along with the other 2 major predators (wolves and bears), and the impacts are different due to selective killing, it seems evident that an experimental approach will be needed, is generally feasible within the framework of present hunting regulations in Alaska and Canada, and would fit into present management areas. No doubt a cooperative arrangement would be necessary, with several states and provinces involved. I believe that this is an important opportunity, and that a successful effort might go a long way toward encouraging more true experimentation in wildlife management.

An important aspect of the art of designing experiments lies in arranging things so that some factors extraneous to the current investigation are minimized and/or made part of the “error” term. For example, good “blocking” in randomized-block experiments can greatly reduce the impact of gradients in soil fertility. When experimentation is not feasible, as in the study of population indices, auxiliary variables may serve a similar purpose.

NON-EXPERIMENTAL STUDIES

The Sampling Paradigm

I believe that some of the concerns expressed in the papers listed in my Introduction lose force
if we approach observational studies in a sampling framework. Modern sampling theory and practice have a solid statistical basis and deal largely with problems of estimation. It remains true that null hypotheses like $H_0: \mu_1 = \mu_2 = \ldots = \mu_k$ cannot be true in the limit with finite populations, but random sampling permits us to make sound inferences about the degree of any differences detected (effect size), with useful confidence intervals for stated $1 - \alpha$ probability. That sizable samples do not demonstrate a significant difference between 2 populations is practically useful, even if an argument can be made that equality is most improbable for any 2 finite populations. Eberhardt and Thomas (1991) made an effort to classify the kinds of approaches that may be useful in environmental studies, listed 8 methods, and suggested how nonexperimental approaches might be dealt with statistically.

Model Selection

Akaike’s Information Criterion (AIC) was advocated by Anderson et al. (2000) as “a practical alternative to null hypothesis testing” and was described in much more detail in the book by Burnham and Anderson (1998, 2002). The approach has become very popular in recent years. The _a priori Model Set._—Anderson et al. (2000) and Burnham and Anderson (1998, 2002) strongly recommend starting with a “set of a priori candidate models (hypotheses) that are well supported.” They use the pejorative term “data dredging” to refer to instances where a wide range of starting models are considered. Because some other authorities (e.g., Draper and Smith 1998) are much less dogmatic about approaches to model selection, more work on this issue would prove useful.

In many cases, supplying a set of well-supported models a priori is difficult. The method sorts out the models with the smallest AIC scores, thus discriminating against the remaining models. Burnham and Anderson (1998:128) recommended calculating the difference between AIC scores for individual models and that of the “best” (smallest AIC score) model and mostly considering those models with a difference ($A_i$) of less than 4–7 units, providing a “rule of thumb for an approximate 95% confidence set.” If only models with small $A_i$ are used, a large set of starting models may be narrowed down to just a few. Consequently, the number of starting models may not matter so much. Of course, if hypothesis testing is used to select candidate models (e.g., by stepwise regression) from a large set of possible models, the objections of Burnham and Anderson (1998) become very important indeed.

Data Dredging.—Anderson et al. (2000) list 2 types of data dredging: (1) “an iterative approach where patterns . . . are chased by repeatedly building new models with these effects included" and (2) “analysis of all possible models (unless, perhaps, if model averaging is used).” Unfortunately, researchers do not have a priori information in many situations where models may be useful. Therefore the recommendation of Anderson et al. (2000) that “both types of data dredging are best reserved for more exploratory investigations that probably should often remain unpublished” seems to me to be questionable advice. Failure to report the results of an exploratory study means the loss of useful information, if some promising results are obtained, and makes a priori grounds difficult to achieve.

I have been interested in the use of indices of population trend (Eberhardt et al. 1999) and the possible use of various auxiliary (explanatory) variables to improve an index. Many such variables measure environmental conditions when the population counts are made. Thus, one might suppose that some a priori reason exists to appraise such variables, but usually only an analysis of actual data turns up 1 or more that substantially increase $R^2$, and AIC is very useful in identifying those variables that are worthwhile. Is this data dredging? Certainly, not reporting the full set of results would be misleading.

Confidence Sets and Model Selection._—The notion of a “confidence set” for models can be misleading, and having some alternative criteria for examining models is important. Anderson et al. (2000:917) noted that for least-squares models

$$AIC = n \log(\hat{\sigma}^2) + 2K.$$

In using multiple regression, $\hat{\sigma}^2$ is essentially the estimate of variance about regression (except that maximum likelihood estimation divides by $n$, rather than $n - 1$ or $n - p$). Often, models with the same number of variables are compared. So setting $A_i = 4$ and comparing a candidate model with that having the minimum AIC gives:

$$4 = n \log(\hat{\sigma}^2) - n \log(\hat{\sigma}^2_{\text{min AIC}}).$$

Because sample sizes ($n$) are necessarily equal, this yields:

$$\exp\left(\frac{4}{n}\right) = \frac{\hat{\sigma}^2}{\hat{\sigma}^2_{\text{min AIC}}}.$$
Thus, if \( n = 40 \), the ratio of estimated variances about regression is about 1.1 (or 1.2 for \( \Delta_k = 7 \)). Thus, the confidence set is rather narrow, and one might suppose that the model with minimum AIC is well supported. But, if \( \hat{\sigma}^2_{\text{min}} \) is fairly large, \( R^2 \) may be small and the model unsatisfactory. Hence, for multiple regression models, one should also use \( R^2 \) for further support.

Burnham and Anderson (1998:77) recognized the need to assess model fit, stating that "If . . . [the] model still fits poorly, the information–theoretic methods will only select the best of the set of poor-fitting models." Hence, a good model is needed in the set considered, as judged by a criterion other than AIC (e.g., \( R^2 \)). I believe that this point may not be appreciated by some users of AIC. Researchers usually consider a 95% confidence interval as a device that will include the true, but unknown, mean in 95% of a very large number of repetitions of the same experiment. The confidence set of models applies only to the particular group of models examined and is thus quite different from the usual concept of a confidence interval.

Model Truth.—Anderson et al. (2000) and Burnham and Anderson (1998, 2002) maintain that "no true models in the biological sciences" exist and "truth in the biological sciences has essentially infinite dimensions." I prefer a somewhat more pragmatic view that regards some models as close enough to truth for most practical purposes. I think that the theoretical work of A. J. Lotka nearly a century ago provides a model for population growth that appears to approximate truth well enough for practical use. Four cases in which populations were evidently counted quite accurately over time yielded an \( R^2 \) exceeding 0.99 (see Table 4 in Eberhardt 2002). In 9 examples of index calculations (see Table 5 in Eberhardt 2002), that true model was not selected as the best approximating model by AIC, no doubt due to the effects of various biases and sampling errors on the indices. For the 9 trend indices, \( R^2 \) did not exceed 0.83, although sizable gains in \( R^2 \) existed over the use of time alone as auxiliary variable.

Population Analysis Paradigm

Another paradigm generally less affected by the concerns about hypothesis testing deals with the analysis of population dynamics. The methodology used traces back to early work by A. J. Lotka and the development of matrix methods by P. H. Leslie, so a handy shorthand cognomen is the Lotka-Leslie model. Eberhardt (1985) provided a useful approximation to Lotka’s original equation, while Caswell (2001) gave a detailed assessment of the matrix approach. A central statistical issue has been that the basic Lotka-Leslie model must be solved iteratively. Hence, convenient confidence limits for the measure of population change (\( \lambda \)) were not available until the advent of the bootstrapping technique described by Efron and Tibishirani (1993).

The “delta method” (Seber 1982) provides a handy way to address the issues of sample sizes for the components of the model (Eberhardt 2002). Effect sizes (i.e., components of the overall variance of \( \lambda \)) can be addressed by bootstrapping (Eberhardt 2002), but no well-developed methodology is available as yet for testing hypotheses about components of the underlying model. Very possibly, further developments in bootstrapping may ultimately serve that purpose. As advocated by many of the authors listed in my Introduction, confidence limits appear to be the best presently available tool for this paradigm.

For at least large mammals, rather consistent patterns exist in survival and reproduction (examples appear in Eberhardt 1985, 2002). One might thus be tempted to use the model selection paradigm to select an overall model for the survival (or reproduction) process. Whether this is advisable may depend very much on the purpose of the study. If the goal is to estimate \( \lambda \) from reproductive and survival data, I think it preferable to either derive estimates directly from age-specific survival and reproductive rates, if possible, or to use an approximation such as that suggested by Eberhardt (1985). Modeling the data may not always be compatible with estimation.

Simulation and Mathematical Modeling

Simulation modeling may constitute another paradigm largely free of the concerns about statistical aspects of hypothesis testing, and in fact, has been used to test all sorts of hypotheses. I am pessimistic about many of the claims made for simulation models (Eberhardt 1975), but the field has nonetheless thrived and now has its own journals (e.g., Ecological Modelling). I do believe that simulation models can help explore and understand ecological systems, but I have been concerned about the uncertainties introduced by the use of components and parameters not well supported by actual data. Both Hall and DeAngelis (1985), and Caswell (1988) have discussed the role of modeling and its virtues and weaknesses. A number of more recent commentaries are listed by Mitro (2001).
Increased use of the bootstrapping technique (Efron and Tibshirani 1993) promises to make possible the production of models with a minimum of arbitrary parameters and with assumptions based on actual field data. Combined with Monte Carlo methods for introducing random variation from weather or other environmental conditions, such a model can serve to test a hypothesis and may be falsified by comparison with observed population data.

DISCUSSION AND CONCLUSIONS

Awareness of the need to use statistical methods in ecology and wildlife management resulted in widespread use of techniques originally developed in the context of controlled experimentation. Application of these techniques to non-experimental circumstances has resulted in the use of inappropriate (silly) null hypotheses. Naked P-values stem from efforts to condense outcomes of statistical tests to a bare minimum (the P-value). I believe that naked P-values can be disposed of quite readily by introducing (and enforcing) a prohibition in a journal’s instructions to authors. Of course, the prohibition should not prevent use of P-values and the usual conventions (single and double asterisks) in large tables, where they may serve a useful purpose.

Several paradigms exist that may be more appropriate for the analysis of observational data, including various kinds of descriptive techniques in which observations are taken by controlled sampling. These observational methods do not provide the basis for strong inference supplied by true experiments. Although medical scientists were once hampered by ethical considerations, they ultimately embraced designed experimentation. Wildlife investigators should consider using the many management units now existing in most states as experimental units (Eberhardt 1988). Robinson and Wainer (2002) noted that evolutionary operations (EVOP) and adaptive management (Walters 1986) provide methods for making incremental experimental changes in ongoing management. Very likely, wildlife managers and ecologists do not really recognize the weaknesses of their present non-experimental approaches.

Model selection using information–theoretic methods, specifically the AIC criterion, has been proposed as an alternative to null hypothesis testing (Anderson et al. 2000). Certainly, this technique is very useful in discriminating among (and perhaps averaging over) a set of candidate models. Selecting an appropriate and limited initial set poses some difficulties, and I believe that supplementary methods are needed to ensure that the selected set contains some good models. Also, once the best model (or an averaged set) is selected, the issue of how to use it remains. In most cases, researchers want to project some trend into the immediate future and usually have no way to project future behavior of the components of a model.

Experiments usually are designed to try out various management techniques and thus accommodate the future. Simulations attempt to explain the past and project into the future but lack inferential strength. Bootstrapping and Monte Carlo methods may permit projections based on actual data. We should not forget that “hard” science proceeds with a mix of modeling and experimentation. Models are tested by experiments designed to falsify the model if possible, and new models are then erected and tested experimentally. When natural systems can somehow be perturbed, observational studies might well use the same general approach. An inevitable conclusion, I believe, is that we should think in terms of a sequence of studies and not to expect to solve important problems in a single study. Hence, I believe that Tukey’s (1960) advice is about as good as we can hope to obtain. I concur that we should emphasize estimation of effects but expect some silly nulls to be a necessary preliminary.

The title of this paper asks “What should we do about hypothesis testing?” My answer is to do legitimate experiments if at all possible, but use existing management frameworks and adaptive methods where possible. When true experiments are not possible, a number of paradigms have been suggested here that contribute to knowledge of the relevant systems without relying on hypothesis testing as a major research tool. Modeling and model selection are relatively new and useful tools but no panacea.

ACKNOWLEDGMENTS

Helpful comments on an earlier draft by S. Cherry, F. Guthery, and D. Johnson are hereby acknowledged, as are reviews by D. R. Anderson, F. Guthery, and an anonymous reviewer. Needless to say, some areas of disagreement remain. Support from NSF DEB00-74444 also is appreciated.

LITERATURE CITED

WHAT SHOULD WE DO ABOUT HYPOTHESIS TESTING?

J. Wildl. Manage. 67(2):2003


Received 19 August 2002.
Accepted 29 January 2003.
Associate Editor: Lubow.