*Invited Paper:*

# THE FALL OF THE NULL HYPOTHESIS: LIABILITIES AND OPPORTUNITIES

FRED S. GUTHERY,[1] Department of Forestry, Oklahoma State University, Stillwater, OK 74078, USA
JEFFREY J. LUSK, Department of Forestry, Oklahoma State University, Stillwater, OK 74078, USA
MARKUS J. PETERSON, Department of Wildlife and Fisheries Sciences and George Bush School of Government and Public Service, Texas A&M University, College Station, TX 77843, USA

***Abstract:*** The collapse of null hypothesis significance testing as a statistical paradigm has created liabilities and opportunities in wildlife science. One liability is that some formalism for statistical hypothesis testing, such as likelihood with information theory, will simply replace null hypothesis significance testing as a rote approach to wildlife science. The principal intellectual instrument of the scientist remains the research hypothesis, not the statistical hypothesis. Accordingly, 1 opportunity arising from a change of statistical paradigms is that the research hypothesis will move to the foreground of wildlife science, the statistical hypothesis to the background. A second opportunity is a broadening of the suite of inferential methods considered orthodox in wildlife science. Realization of these opportunities should help wildlife scientists resist the social tendency to allow tools (means, statistical models) to supplant the search for reliable knowledge (end) as the benchmark of scientific endeavor. Science of the highest order, including virtually all discoveries that humankind extols today, is possible without the statistical hypothesis, but not without the research hypothesis.

***JOURNAL OF WILDLIFE MANAGEMENT 65(3):379–384***

***Key words:*** Akaike Information Criterion, inference, likelihood, null hypothesis, research hypothesis, statistical hypothesis.

Romesburg (1981) clarified for wildlife science the difference between statistical hypotheses (conjectures on the properties of data) and research hypotheses (conjectures on the processes of nature). Wildlife scientists responded with a tradition of crisply stating null hypotheses in the first person, active voice, and perhaps making a few predictions on the direction of treatment effects. Romesburg's (1981) main point was missed, almost completely. He advocated formulating research hypotheses followed by making deductions on events that would occur under observation or experimentation if the research hypotheses were true. The statistical hypothesis was simply a means of assessing the verity of deductions under the shadow of uncertainty.

Recent articles by Cherry (1998), Johnson (1999), and Anderson et al. (2000) alerted wildlife scientists to flaws in the long-held tradition of null hypothesis significance testing (hereafter, significance testing). These authors argued that significance levels (*P*-values) are over-reported; that statistical tests are often unnecessary because effects are known a priori and magnitude of effect is the issue of interest (Edwards 1992); and that the statistical null hypothesis (effect size = 0) must usually be false in the limit (e.g., large datasets analyzed on computers with finite-decimal capabilities). Wade (2000) showed that traditional significance testing—in addition to the problems mentioned above—deals poorly with issues of uncertainty and may therefore foster management error. Anderson et al. (2000) advocated likelihood in an information-theoretic venue as an alternative to significance testing. Hereafter, we use the word likelihood in the sense elaborated in Anderson et al.'s (2000) paper.

Under the arguments of Cherry (1998), Johnson (1999), and Anderson et al. (2000), wildlife science seems to be in the nascent stages of a change of paradigm concerning methods of analyzing data and drawing statistical inferences. We see liabilities and opportunities in this change. A major liability is that some rote approach to analysis and inference will simply replace significance testing. Cohen (1994:1001) observed, "[D]on't look for a magic alternative to [null hypothesis significance testing], some other objective mechanical ritual to replace it. It doesn't exist." A second liability is that the statistical hypothesis will retain its invalid priority relative to the research hypothesis. These liabilities can be turned into opportunities if the research community is willing to view the accumulation of reliable knowledge from a more heterodox perspective, and if the research hypothesis

[1] E-mail: forfsg@okstate.edu

takes precedence as the primary intellectual instrument of the scientist (Beveridge 1957:71).

We address certain concerns over the change of analytical paradigm as it seems to be developing. We point out shortcomings of likelihood as advocated in Anderson et al. (2000) and in their more general coverage of the subject and related issues of science (Burnham and Anderson 1998). We argue that likelihood merely changes the nature of statistical hypothesis testing and thus places intellectual effort on the wrong aspect of science. We will revisit hypotheses (null and research) and argue, contrary to Johnson (1999), that research hypotheses are omnipresent in wildlife science though seldom stated explicitly. We discuss the human tendency to switch the relative priorities of means (e.g., statistical tests) with ends (e.g., reliable knowledge). We conclude with a brief discussion of astatistical science, which has produced all of the major discoveries currently extolled by humankind. Our main message is simple: statistical hypothesis testing of any kind is a means, not an end, and advances in understanding come largely if not solely through the formulation and testing of research hypotheses.

## LIKELIHOOD AND INFORMATION THEORY

Burnham and Anderson (1998) and Anderson et al. (2000) presented compelling arguments in support of likelihood as a powerful and conceptually sound replacement for significance testing. In the course of these arguments, however, the idiosyncrasies of likelihood were not well elaborated. Potential users of the methodology should appreciate its limitations, especially in the general context of scientific accomplishment.

Likelihood retains some of the pitfalls of significance testing. It is by definition a parametric approach to inference because one cannot compute a likelihood without a discrete or continuous probability model. Thus, in many applications, the researcher using likelihood must make strong assumptions regarding the probability distributions that they obtain in nature. If the assumptions are wrong, the inferences drawn from a dataset are not likely to be useful.

The use of likelihood provides no protection against trivial hypotheses, a point condemned for significance testing. Of course, no statistical methodology can provide this protection, but the fact remains that a hypothesis trivial under significance testing remains trivial under likelihood. Consider the example provided by Anderson et al. (2000). The research question was whether spectacled eiders (*Somateria fischeri*) with lead poisoning had annual survival rates different from those without lead poisoning; the analytical approach was essentially control versus treatment (exposed to lead versus not exposed). Not surprisingly, poisoned birds survived at lower rates than did control birds. Perhaps this example was poorly chosen by Anderson et al. (2000), but the results confirmed the obvious, as significance testing would have. The analysis provided no information on dose-response relationships and accordingly rendered magnitude of effects arguments possibly contingent on the specific sample analyzed. Further, another variable in the eider models (site) was not scientifically informative. Science cannot build a theory of eider demographics based on knowledge of site effects unless it deals with properties of sites, not sites per se.

Anderson et al. (2000) argued that multiple hypotheses were a component of good science and cited Chamberlin (1965) in support of these arguments. There are 2 problems here. First, Chamberlin did not advocate multiple statistical hypotheses (the situation in likelihood, see below) but rather multiple research hypotheses. That is, Chamberlin advocated working with multiple explanations of how a process operates in nature, not multiple statistical models based on essentially the same underlying idea on how nature works. For example, using Anderson et al.'s eider example, the multiple hypotheses tested were merely rearrangements of the same underlying driving variables.

Second, Chamberlin advocated multiple research hypotheses 80 years before the melding of information theory and likelihood into an analytical package. He did not advocate multiple research hypotheses in anticipation of likelihood, but rather as a means for individual scientists to protect themselves from personal bias. Chamberlin believed scientists would be less likely to become enamored of a favorite research hypothesis, and thus to compromise their objectivity, if they had a set of competing hypotheses. He therefore recognized "the universal [human] tendency to notice instances that corroborate a favorite belief more readily than those that contradict it" (Dewey 1910:22).

## HYPOTHESES

### The Null Effect

That a statistical null hypothesis is inevitably false (Cohen 1994) based on the mathematical structure of null hypothesis testing does not prove that a research null hypothesis is always

wrong where theory meets practice. Wildlife scientists must be careful not to confuse null hypotheses and null effects because nature must be ripe with null effects that are significant. If community stability is a nondecreasing function of community diversity, for example, there would seem to be levels of diversity over which stability is constant. Altering diversity in these domains would have null effects on stability. Null effects would seem to hold, within the appropriate domain, for all variables in nature that are asymptotically related. For example, if a wildlife population conforms to logistic growth, there will be a range (domain) of population sizes that has null effects on the rate of growth of the population. We could even conjecture that, in nature, all continuous, curvilinear relationships with extrema have neighborhoods of nullness—large or small—where change in a forcing variable or set of forcing variables is not associated with change in a response variable or set of response variables. For example, yield is a parabolic function of population size under the logistic model; there is a neighborhood of nullness near the single population size associated with maximum sustained yield. Within this neighborhood, change in population abundance is associated with trivial changes in yield. The related idea of compensatory harvest mortality is a classic example of a null effect for harvest rates ranging from 0 to some threshold level. From a nonlinear perspective, neighborhoods of nullness (and hence, null effects within specified domains) would seem to be general, interesting, and important properties of nature.

Likewise, the practice of wildlife management undoubtedly is ripe with null effects. For example, wildlife managers seem to operate under the fundamental assumption (research hypothesis) that wildlife populations are food-limited. Under this assumption, a natural management response is to increase food supplies. Yet for northern bobwhites (*Colinus virginianus*), there is no convincing evidence that adding food to occupied areas increases abundance (Guthery 1997); i.e., effect size = 0. Whereas effect size = 0 may be incompatible with the theory of significance testing, it may be quite compatible with the reality of the field and may lead to better understanding of nature.

In a similar vein, a research hypothesis of no effect is a legitimate challenge to untested assumptions and dogmatic principles of management. Leopold's (1933:132) law of dispersion (principle of edge), for example, was elevated "into a theorem by usage rather than by more scientific methods" (Giles 1978:139). In such situations, where the processes of human social behavior lead to "proof by acclimation," a null effect hypothesis may be properly aggressive from a scientific perspective.

Our point is that the somewhat tautological silliness of the statistical null hypothesis in no way abrogates the importance of null effects in ecology and management. Indeed, neighborhoods of nullness serve as props for understanding the stability of populations in response to harvest, variation in habitat features, and certain management perturbations.

## Statistical versus Research Hypotheses

We appreciate the value of quantitative models in wildlife science; indeed, we have published several mechanistic and phenomenological models of wildlife population processes. We also recognize that mathematical models (including statistical models as special cases) may be taken as research hypotheses in certain situations. However, we suspect that, in general, statistical and research hypotheses are best described as members of fuzzy as opposed to crisp sets (Kosko 1992). This means that a statistical hypothesis may be to some degree a research hypothesis, and a research hypothesis may be to some degree a statistical hypothesis.

Given the fuzzy nature of hypotheses, one could argue that the multiple statistical hypotheses advocated under likelihood are to some degree research hypotheses. We contend these multiple hypotheses are less like research hypotheses and more like statistical hypotheses. A probability model and a numerical criterion of favor and disfavor attend inference under likelihood; the Akaike Information Criterion or a related index is the functional analog (not homologue) of a *P*-value in model-selection exercises. It is a numerical criterion by which one judges (i.e., tests) the strength of evidence in support of a statistical model. Therefore, likelihood does not deal explicitly with research hypotheses any better than did significance testing. Of course, it was not meant to do so.

Krebs (2000:9) lamented, "Almost all statistical tests reported in the literature address low-level hypotheses of minor importance to the ecological issues of our day, not the major unsolved problems of ecological science." The tests emanating from likelihood, although softened by ambiguity and strengthened by compromise between parsimony and explanation, play about the same role in the "ecological issues of our day" as did null hypothesis significance testing. Likelihood, like significance testing, is but a tool derived to assist

deductive science in evaluating the merit of research hypotheses under the uncertainty that attends laboratory and field experiments.

Johnson (1999) reasoned that (explicitly stated) research hypotheses were rare in wildlife science because of the interconnectedness of components in an ecosystem, leading to nebulosity over what could be considered an effect. We offer a simpler explanation: by dint of training, tradition, and the feedback loop engendered by training and tradition, wildlife scientists have come to see the statistical hypothesis as more important than the research hypothesis—a prime example of what Burke (1984) termed a trained incapacity. We point out that one cannot design a study or conduct a statistical test without a research hypothesis because the research hypothesis governs what perturbations are invoked, what variables are measured, and what tests are conducted. So a research hypothesis is always present, at least implicitly.

Research hypotheses may range from the everyday to the elegant and their predictions from the sophomoric to the sophisticated. A food plot study, for example, operates under the research hypothesis that a population is food-limited and the researchers predict (deduce) increased density with the addition of foods to an area. They may further predict that increases in density will be proportional to the amount of food provided.

In other situations, the processes leading to an outcome are more involved and the deduction(s) less obvious. These circumstances may hold for basic and applied studies. For example, in a basic study, Kohlmann and Risenhoover (1998:178) reasoned that a forager would tend to selectively exploit the richest patches if it had "preharvest information" on patch richness (as garnered through experience), and if the spatial scale of available patches matched the forager's cognitive representation of the environment. This research hypothesis led to a prediction of selective foraging behavior. They reasoned further that a lack of preharvest information, mismatched cognitive and real spatial scales, and/or high variability in resource availability would lead to a prediction of nonselective foraging behavior. These research hypotheses were evaluated in the laboratory, and their validity was judged with statistical tests.

In the field, experimental manipulation of variables thought to govern processes is more difficult than in the laboratory. However, this does not preclude the application of deductive science. For example, Spears et al. (1993), finding empirical contradictions to the hypothesis that bobwhites are a lower successional species, reasoned that commonalities in habitat structure govern the abundance of bobwhites in different regions (research hypothesis). They observed that a given structure occurs in different seral stages on sites of different productivity, as mediated by soils and climate. They deduced, under the research hypothesis, the presence of an interaction effect between bobwhite abundance and seral stage on sites of different productivity and tested for this effect with field data. An interesting circumstance in this study was that the data were not amenable to analysis by inferential statistics; e.g., a "low seral stage" treatment in a xeric environment differed compositionally and structurally from a "low seral stage" treatment in a mesic environment. The scientific strength of the work arose because empirical findings supported an a priori prediction on bobwhite population behavior, although not unambiguously.

The handicap principle (Zahavi and Zahavi 1997) serves as an example of an inspired research hypothesis that we see as more elegant than a suite of Akaike Information Criteria in a multi-model likelihood setting. The principle states that in the animal kingdom a signal receiver (e.g., a sexually receptive female) judges the quality of a signaler (e.g., a male) according the risk (handicap) associated with signaling. In other words, the risk associated with a signal reliably quantifies the fitness of the signaler (nature's method of assuring truth in advertising). As a point of interest, scientific challenges to the handicap principle have taken place in a game theoretic context (Johnstone 1998), devoid of statistical inference but not devoid of reason and mathematical logic.

The handicap principle might ultimately fall to challenges from skeptical scientists, but it now serves as a simple, unifying theme that explains the behavior of a diversity of organisms in a broad range of contexts (Johnstone 1998). Science is a search for such themes, an attempt to collapse the chaos of nature (Cohen and Stewart 1994) into a set of simple, explanatory models. The search may be aided by the statistical hypothesis, but it starts and ends with the research hypothesis.

## MEANS VERSUS ENDS

Cherry (1998), Johnson (1999), and Anderson et al. (2000) clearly demonstrated that significance testing (a tool or means) was transmuted into an end by wildlife science. Postman (1992) maintained that modern human institutions,

including science, are uniquely gifted at such transmutations. Tools, he argued, were originally objects invented either to solve specific and urgent problems of physical life or serve in the symbolic world of art, myth, or religion. As a technological culture develops, however, tools play a more central role in its thought-world. In essence, technologies become the culture. Suppose, for example, that we require a house to protect us from inclement weather and that a hammer is needed for construction. The tendency of modern society eventually would be to focus on perfecting the hammer, rather than building the house—thus transmuting means into ends.

Environmental scientists are far from immune to this process. Recently, Peterson (1997:86–118) evaluated an Environmental Assessment in Canada wherein governmental wildlife and animal disease experts recommended killing all bison (*Bison bison*) in Wood Buffalo National Park to reduce the risk of disease transmission from these bison to cattle or disease-free bison herds. She found that local residents, many of whom had little formal education, were much less likely to conflate means and ends than were biological "experts." This resulted in a broader array of options available to them than to the experts. Although the culture of expertise personified by biologists predominated throughout the official hearings, the local residents subverted this culture of expertise and prevented the bison slaughter recommended by the biologists. Whereas focusing on methodology may lend an air of objectivity to wildlife science, eventually the methodology threatens to subsume the goal for which the methods were originally developed.

## ASTATISTICAL SCIENCE

The heading of this section may seem an oxymoron to wildlife scientists. Our profession has become so smitten with statistical testing of 1 form or another (Cherry 1998, Johnson 1999, Anderson et al. 2000) that science without statistics is almost inconceivable. In fact, science without statistics is unparalleled.

The greatest accomplishments in the history of science are debatable, but Brody and Brody (1996) list these: gravity and the basic laws of physics, the structure of the atom, the principle of relativity, the Big Bang and the formation of the universe, evolution and the principle of natural selection, the cell and genetics, and the structure of the DNA molecule (note that 3 of the 7 accomplishments are biological). Mendelian genetics offers a specific example of this type of astatistical science. Without the aid of statistics, and with nothing more sophisticated than simple frequency calculations, Fr. Mendel was able to deduce the fundamental principles of heredity. A more recent but related example is the mapping of the human genome (International Human Genome Sequencing Consortium 2001). The completion of this monumental task completely rewrote ideas (research hypotheses) about the number of genes in the human genome. Instead of the 100,000–150,000 genes originally hypothesized, the completed sequences indicated a mere 30,000. In fact, none of these scientific contributions was made possible by statistical hypotheses. The contributions arose from simple descriptive science embellished with research hypotheses that explained the patterns apparent in descriptive data. The development of these retroductively derived hypotheses (Romesburg 1981) required years of testing before they reached the stature of scientific principles, theories, or laws (Krebs 2000), but did not hinge on statistical hypotheses.

Although these important scientific contributions did not rely on statistical hypothesis testing, we should not infer that no evaluative methods were used. Instead, methodologies free from statistical hypothesis testing were employed. In recent years, ecological science typically has ignored these approaches, used them only in conjunction with inferential statistics, or relegated their use to popular literature. We maintain, however, that logical arguments, such as those used by Cherry (1998), Johnson (1999), and Anderson et al. (2000), and/or mathematical arguments can further the cause of wildlife science. Similarly, conducting sensitivity analyses using ecological simulation models can provide useful, reliable knowledge (Maguire et al. 1995, Wisdom and Mills 1997, Peterson et al. 1998). Numerous other largely astatistical approaches also deserve consideration, including graphic depictions of data, qualitative analyses, and soft systems. Just as Chamberlin (1965) advocated multiple research hypotheses to protect against personal bias, so might wildlife scientists expand the suite of analytical practices considered orthodox to protect against the transmutation of means and ends.

## CONCLUSION

We see likelihood as a formalism for statistical inference that is better than null hypothesis significance testing where statistical hypotheses have

merit, but that performs the same secondary service to the scientific community: tests of statistical hypotheses. As wildlife science changes the manner in which it deals with statistical inference, practitioners need to keep in mind that the statistical hypothesis is a means, not an end.

The opportunities associated with the nascent paradigm shift in wildlife science—assuming we can avoid mindlessly replacing one form of inferential statistics with another—include reemphasizing research hypotheses, deductions related to these hypotheses, and creative manipulative or mensurative experiments designed to test these deductively derived predictions (Romesburg 1981, Hurlbert 1984). This process can be enhanced if academicians, referees, and editors are willing to place logical argument, effect estimation, descriptive statistics, graphic approaches, qualitative analysis, and other methods in the realm of scientific orthodoxy.

## ACKNOWLEDGMENTS

This paper benefited from review comments by T. R. Peterson, E. C. Hellgren, S. D. Fuhlendorf, G. A. Baldassarre, L. A. Brennan, D. R. Anderson, and an anonymous statistician. Our acknowledgment of these reviewers does not indicate that they agree in toto with our outlook on wildlife science. This work was supported by the Game Bird Research Fund, Bollenbach Endowment, the Environmental Institute at Oklahoma State University (OSU), the OSU Foundation, and the Oklahoma and Texas Agricultural Experiment Stations. This manuscript is approved for publication by the Oklahoma Agricultural Experiment Station.

## LITERATURE CITED

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.

Beveridge, W. I. B. 1957. The art of scientific investigation. Vintage Books, New York, USA.

Brody, D. E., and A. R. Brody. 1996. The science class you wish you had. Perigee Books, New York, USA.

Burke, K. 1984. Permanence and change. Third edition. University of California Press, Berkeley, USA.

Burnham, K. P., and D. R. Anderson. 1998. Model selection and inference. Springer, New York, USA.

Chamberlin, T. C. 1965 (1890). The method of multiple working hypotheses. Science 148:754–759. (Reprint of 1890 paper in *Science*.)

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

Cohen, J. 1994. The earth is round ($p < 0.5$). American Psychologist 49:997–1003.

———, and I. Stewart. 1994. The collapse of chaos: discovering simplicity in a complex world. Penguin Books, New York, USA.

Dewey, J. 1910. How we think. D. C. Heath, Boston, Massachusetts, USA.

Edwards, A. W. F. 1992. Likelihood. Johns Hopkins University Press, Baltimore, Maryland, USA.

Giles, R. H., Jr. 1978. Wildlife management. W. H. Freeman, San Francisco, California, USA.

Guthery, F. S. 1997. A philosophy of habitat management for northern bobwhites. Journal of Wildlife Management 61:291–301.

Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54:187–211.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

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

Johnstone, R. A. 1998. Game theory and communication. Pages 94–117 *in* L. A. Dugatkin and H. K. Reeve, editors. Game theory and animal behavior. Oxford University Press, New York, USA.

Kothmann, S. G., and K. L. Risenhoover. 1998. Effects of resource distribution, patch spacing, and preharvest information on foraging decisions by northern bobwhites. Behavioral Ecology 2:177–186.

Kosko, B. 1992. Neural networks and fuzzy systems. Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Krebs, C. J. 2000. Hypothesis testing in ecology. Pages 1–14 *in* L. Boitani and T. K. Fuller, editors. Research techniques in animal ecology. Columbia University Press, Columbia, New York, USA.

Leopold, A. 1933. Game management. Charles Scribner's Sons, New York, USA.

Maguire, L. A., G. F. Wilhere, and Q. Dong. 1995. Population viability analysis for red-cockaded woodpeckers in the Georgia piedmont. Journal of Wildlife Management 59:533–542.

Peterson, M. J., W. E. Grant, and N. J. Silvy. 1998. Simulation of reproductive stages limiting productivity of the endangered Attwater's prairie chicken. Ecological Modelling 111:283–295.

Peterson, T. R. 1997. Sharing the earth: the rhetoric of sustainable development. University of South Carolina Press, Columbia, USA.

Postman, N. 1992. Technopoly: the surrender of culture to technology. Alfred A. Knopf, New York, USA.

Romesburg, H. C. 1981. Wildlife science: gaining reliable knowledge. Journal of Wildlife Management 45:293–313.

Spears, G. S., F. S. Guthery, S. M. Rice, S. J. DeMaso, and B. Zaiglin. 1993. Optimum seral stage for northern bobwhites as influenced by site productivity. Journal of Wildlife Management 47:805–811.

Wade, P. R. 2000. Bayesian methods in conservation biology. Conservation Biology 14:1308–1316.

Wisdom, M. J., and L. S. Mills. 1997. Sensitivity analysis to guide population recovery: prairie-chickens as an example. Journal of Wildlife Management 61:302–312.

Zahavi, A., and A. Zahavi. 1997. The handicap principle: a missing piece of Darwin's puzzle. Oxford University Press, New York, USA.