Invited Paper:

THE ROLE OF HYPOTHESIS TESTING IN WILDLIFE SCIENCE

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

Abstract: Statistical testing of null hypotheses recently has come under fire in wildlife sciences (Cherry 1998; Johnson 1999; Anderson et al. 2000, 2001). In response to this criticism, Robinson and Wainer (2002) provide some further background information on significance testing; they argue that significance testing in fact is useful in certain situations. I counter by suggesting that such situations rarely arise in our field. I agree with Robinson and Wainer that replication is the key to scientific advancement. I believe, however, that significance testing and resulting *P*-values frequently are confused with issues of replication. Any single study can yield a *P*-value, but only consistent results from truly replicated studies will advance our understanding of the natural world.

JOURNAL OF WILDLIFE MANAGEMENT 66(2):272–276

Key words: effect size, hypothesis test, null hypothesis, replication, significance test.

The wildlife literature recently has hosted a number of articles critical of the statistical testing of null hypotheses (hereafter, significance testing). Cherry (1998), Johnson (1999), and Anderson et al. (2000, 2001) argued that the practice is applied too often and inappropriately. Scientists in other disciplines have had similar discussions (see Johnson 1999 for references). Robinson and Wainer (2002; hereafter RW) provide some history of significance testing and its background. Understanding how significance testing arose is helpful in appreciating its strengths and weaknesses. Robinson and Wainer further offer some examples intended to illustrate situations in which significance testing of null hypotheses can be appropriate. I agree with much of what RW say. In fact, they restate and reinforce many of the points made in articles to which they are responding. For example, their emphasis on R. A. Fisher's view of the importance of replication is consistent with points made by Johnson (1999). While I do not dispute most of the points RW make, I question whether some of these points have utility to the wildlife profession. In this commentary, I focus on the issues on which I do not fully concur with RW. I offer comments of 2 types: (1) general responses to RW, and (2) remarks related to the points made by RW as they specifically apply to wildlife situations.

GENERAL RESPONSES

Robinson and Wainer (2002) claim to have seen no evidence that significance testing is misused any more often than any other statistical procedure. Exact numbers would be difficult to calculate, but by its very nature, significance testing of null hypotheses must be misused more than other procedures. This is true because significance testing is integral to the misuse of most other procedures. For example, perhaps the most easily misused statistical procedure is stepwise regression (e.g., Draper et al. 1971, Pope and Webster 1972, Hurvich and Tsai 1990, Thompson 1995). Stepwise regression includes or excludes variables in a model depending on their *P*-values. That is, model selection is based on tests of hypotheses that the effect of individual explanatory variables on a response variable is zero, given the other explanatory variables already included in the model. A second kit of statistical tools that are readily misused are multivariate methods (Armstrong 1967, Johnson 1981, Rexstad et al. 1988). Again, a key aspect involving their misuse is based on significance tests; multivariate procedures permit a plethora of null hypotheses to be tested. Other examples could be cited, but the mere fact that significance tests are central to so many statistical procedures per force implies that they are misused more than any single procedure.

As an example of misused statistics, RW suggest that the mean is inappropriate when the underlying distribution contains outliers. That statement may be true in some instances, but not in others; the appropriateness of any statistical procedure depends on the objective of the analysis. Consider an example involving a known population of 51 pheasant hunters in a county. I will simplify the example by supposing that 25 of them get no birds during the season, 25 of them bag 1 bird each, and

[1] E-mail: Douglas_H_Johnson@usgs.gov

a single individual shoots 179 birds. The mean bag is 4 birds. The median bag is 1 bird. Which of these 2 measures of central tendency is more appropriate? If the objective is to characterize the typical hunter, then the median (1 bird) clearly portrays the harvest of this hunter better than the mean does. So the median might be the measure of choice in a human dimensions study. Suppose, however, that interest was in the dynamics of the pheasant population. Then we would be concerned about the total harvest, which is the mean bag × the number of hunters. In this situation, the mean—not the median—is the appropriate measure, even though the distribution contains a wild outlier. The blanket statement by RW about the propriety of the mean is misleading: the objective of a particular investigation must be considered.

Robinson and Wainer (2002) agree with Guthery et al. (2001) that adoption of information–theoretic methods in place of significance testing would still involve an arbitrary numerical criterion to judge the strength of evidence in single studies. I disagree. One of the advantages of the information–theoretic approach is that it allows a set of models to be ranked, based on the support each model receives from the data. Further, information–theoretic methods lend themselves nicely to model averaging (Burnham and Anderson 1998, Anderson et al. 2000). Model averaging involves consideration of the full set of meaningful models that are supported by the data. Instead of settling on a single best model, all supported models are considered, with weights related to the strength of evidence for each model. That process is in stark contrast to what is usually done by significance-testing approaches to model selection, in which a single model is chosen based on tests of null hypotheses that the effects of variables are exactly zero.

Replication

Robinson and Wainer (2002) remind us of the importance of replication, pointing out R. A. Fisher's perspective of science as a continuing process. They cite (2002:265) with evident approval Fisher's belief that significance testing "only made sense in the context of a continuing series of experiments that were aimed at confirming the size and direction of the effects of specific treatments." They also cite Tukey (1969) and note (2002:269) that "statistically significant results that are replicated provide the basis of scientific truth." I wholeheartedly concur. Studies conducted to understand phenomena generally are of too small a scale to yield results of unquestioned significance. Robinson and Wainer (2002) note that Fisher was not concerned with what we call Type I errors, claiming an effect to be real when it is not; he thought that continued replications would demonstrate that the effect was not real. In wildlife applications, however, we rarely seek to replicate studies. Many studies simply cannot be replicated because of conditions that vary markedly from 1 occasion to another. Further, we often (too often, in my opinion) urge managers to take action based on the results of a single study. Indeed, authors writing for *The Journal of Wildlife Management* (*JWM*) are strongly encouraged to include a Management Implications section. The authors call managers to arms based on results of their single—probably unreplicated—study, in which a Type I error may have occurred. This is not the situation Fisher envisaged.

Johnson (1999) emphasized that replication is a cornerstone of science and referred to Carver's (1978) point that statistical significance generally is interpreted as relating to replication. Johnson (1999:768) even noted Fisher's idea of "repeatedly getting results significant at 5%." Key to this issue is the comment by Bauernfeind (1968) that replicated results automatically make statistical significance testing unnecessary. Fisher argued that a result significant at 5% provided motivation to continue studying the phenomenon. Once such a series of experiments has been conducted, however—and most of them have provided *P*-values less than 0.05—those individual *P*-values are of little relevance.

In my view, the interpretation of any statistical evidence (e.g., *P*-values, estimated effect sizes, confidence intervals) makes sense only if the interpretation is grounded in the context of prior related findings. Even if no individual study obtained statistically significant results—but the effect sizes from a series of studies were consistent—important truth may have been discovered. Indeed, proponents of significance testing (e.g., Robinson and Levin 1997) and its opponents (e.g., Thompson 1996) agree on the importance of replication in research.

Detractors of significance testing, however, argue that too many researchers erroneously interpret statistical significance as necessary and sufficient evidence that results are replicable (Cohen 1994). Without statistical significance tests, such researchers would be forced to compare their effect sizes directly with those from similar studies or actually to conduct further

replicated studies. The researchers also would be forced to argue explicitly that their effects are practically—as opposed to only statistically—significant. Doing so would be a positive contribution to most areas of science.

Proving the Null Hypothesis

Robinson and Wainer (2002) note that critics of significance testing worry that researchers too commonly interpret results with $P > 0.05$ as indicating no effect. The critics are right: researchers do exactly that. A scan of the first few papers in a recent issue of *JWM* (Volume 65) found several examples in which authors determined that there was no effect after finding $P > 0.05$, even if P just barely exceeded 0.05. Among these instances were: "the probability of being detected in at least 1 month (p_i^*) did not differ from 1 ($P > 0.058$)"; "overlap between female and male core areas differed neither in early ($U = 30$, $P = 0.052$) nor in late spring ($U = 51.0$, $P = 0.144$)"; and "Daily nest survival was not significantly different between regeneration methods for ... yellow-breasted chat ($\chi^2 = 3.28$, df = 1, $P = 0.07$)." The problem of declaring no effect arises especially often when interactions are examined. If the interactions are real, even if not statistically significant, interpretation of main effects is confounded. Yet authors typically provide little information about interactions. One of the scanned articles in the *JWM* issue was characteristic: "[I]nteractions ... were not significant. Therefore, we ..." No evidence, not even a *P*-value, was provided to demonstrate that the interactions really were negligible and could safely be ignored.

Scientific versus Statistical Hypotheses

Robinson and Wainer (2002) state that not all *P*-values are unimportant and refer specifically, if obliquely, to Albert Einstein and a hypothesis about the speed of light. Without knowing exactly what hypothesis RW are referring to, I would suggest that it likely represents a scientific, as opposed to a statistical, hypothesis. Johnson (1999), among many others, also distinguished these kinds of hypotheses, citing Copernicus' hypothesis that the Earth revolves around the sun, in contrast to the hypothesis widely believed at the time that the sun revolved around the Earth. That scientific hypothesis was contrasted with the statistical hypotheses typically tested in *JWM* and many other scientific journals.

Confidence Intervals

Robinson and Wainer (2002) note that Anderson et al. (2001) recommended the use of a $(1 - \alpha)$ confidence interval for portraying the uncertainty of an estimate, in lieu of *P*-values. Robinson and Wainer suggest that a $(1 - \alpha)$ confidence interval is as arbitrary as rejecting or failing to reject a null hypothesis based on whether or not $P < \alpha$. It is true that a $(1 - \alpha)$ confidence interval gives as much information as knowing whether or not $P < \alpha$. But it gives much more information. The width of the confidence interval tells how well the parameter has been estimated. The distance from the hypothesized value of the parameter to the confidence interval gives a measure of the inconsistency of that value with the observed data. In contrast to a *P*-value, a confidence interval allows the reader to know if lack of statistical significance represents lack of effect or too small a sample size (Johnson 1999: Fig. 1). Further, the clear distinction between confidence intervals and significance testing can be seen in the realization that one cannot test statistical significance without a null hypothesis, but that confidence intervals can be obtained without nulls.

A major advantage of confidence intervals is that they allow (and even facilitate) thinking "meta-analytically" about effect size and effect size replicability across studies (Anderson et al. 2000, Cumming and Finch 2001). Confidence intervals have the additional appeal that they are readily amenable to graphical presentation. Unfortunately, confidence intervals are too rarely reported in many scientific journals.

APPLICABILITY TO WILDLIFE SITUATIONS

Some of the arguments made by RW are correct but apply to few situations in wildlife science. As an example, wildlifers can only envy databases like the Cochrane Collaboration, which is based on more than 250,000 medical experiments with random assignments (presumably of treatments to subjects) and for which enough information is provided to conduct meta-analyses. We have nothing comparable, but instead do as RW (2002:265) say: we "rarely replicate results where $P < 0.05$...."

Robinson and Wainer argue that testing of null hypotheses can be useful when attempting to determine only the sign of an effect, rather than its sign and magnitude. They illustrate this idea with a medical research example in which a new treatment is compared to an old one. Once a treatment has been demonstrated to be superior, ethical considerations demand that the inferior treatment not be applied to additional subjects. The magnitude of the difference between treat-

ments is not estimated; knowing the sign of the effect is sufficient to make a decision.

That example is valid but rarely relevant in the wildlife field. We generally need to know the magnitude, as well as the sign, of an effect. Consider the hypothesis: if we eliminate sport hunting on the North American mallard (*Anas platyrhynchos*) population, mallard survival rates will increase. Probably all of us believe this statement is true. The real question is: How much will survival rates increase? Is the increase in survival rate worthwhile compared to the loss of recreational opportunities? Similarly, we might all agree, even without study, that eliminating all animals that depredate nests of a species in an area will have a positive effect on the nesting success of that species there. But this information is not enough: we need to know how big that increase will be. Predator reduction is expensive and has social implications, and a conscientious manager wants to know what benefits will result from such costly and potentially controversial actions.

Evolutionary Operation

Evolutionary operation (EVOP) is proffered by RW as a situation in which interest lies only in the sign of an effect. As RW note, EVOP is applied in industrial settings when slight differences in manufacturing procedures (temperature, chemical inputs, etc.) are made and the direction of the effect on the product is noted (Box 1957, Box and Draper 1969). Only small changes from current settings are made, so that actual production is not compromised. Hence, effects are likely to be small, too. Robinson and Wainer note that only the direction of the change is important: did the quality of the product improve or worsen?

Box and Draper (1982) provided an overview of EVOP. Interestingly, in their examples, they presented estimated effects and their standard errors, but no *P*-values or hypothesis tests. Evolutionary operation is akin to adaptive resource management (Walters 1986), which has gained increased popularity in wildlife and fisheries management. Both methodologies focus on learning about the system at the same time the system is managed. That is, managers want to manipulate inputs to the system to seek optimal combinations of those inputs while not varying things so much as to cause a serious reduction in the output. The major difference between the methodologies, in my view, is that natural systems have far more uncontrollable, and often unknowable, inputs than do the industrial systems for which evolutionary operation was designed. It remains to be seen whether adaptive resource management will be as successful as evolutionary operation has been.

CONCLUSIONS

Testing hypotheses is an important component of scientific endeavor. Indeed, it is integral to the hypothetico-deductive method, which is a powerful way of learning (Romesburg 1981). Key to this concept is that the hypotheses being tested are scientific, not merely statistical. That is, scientific hypotheses address fundamental, global predictions that derive from theory. Statistical hypotheses, in contrast, address local questions, usually about single populations or systems (Simberloff 1990), and the null hypotheses usually are meaningless and known a priori to be false. Most hypotheses tested in *JWM* are statistical in nature, not scientific. Wildlife researchers should be encouraged to employ scientific hypotheses more often and statistical hypotheses less often.

It is widely acknowledged that virtually all statistical null hypotheses are known to be false, even before any data are collected or any tests conducted (Johnson 1995, 1999; Cherry 1998; Anderson et al. 2001). Why then should a significance test be conducted? As it turns out, significance testing can be useful for determining if the null hypothesis is approximately true, if the sample size is not too small and not too large (Berger and Delampady 1987). For example, the null hypothesis that the means of 2 populations are the same ($\mu_1 = \mu_2$) is almost certainly false in any finite population. Nonetheless, the hypothesis will be accepted if the sample is too small. The hypothesis $\mu_1 \sim \mu_2$ is much more reasonable to consider; how similar the means need to be depends on the context. This hypothesis will be accepted if the sample is too small and will be rejected if the sample is very large, but for moderate samples, the test can be meaningful.

Testing null hypotheses is seldom useful or necessary. The examples cited by RW rarely are germane to the wildlife field. The fundamental need, as RW mention and as R. A. Fisher emphasized, is for true replication. Researchers and managers should not rely on single studies conducted in a single area even over a few years, but instead should require results that are replicated by different researchers using a variety of methods. As Cohen (1994), Thompson (1996), and others have strongly emphasized—contrary to common misperceptions—*P*-values do not reflect

the repeatability of study results; actual replications are required to definitively establish repeatability. Any single study can yield a *P*-value, but only consistency among replicated studies will advance our science.

ACKNOWLEDGMENTS

I am grateful to D. R. Anderson, M. M. Rowland, and G. A. Sargeant for valuable comments on earlier drafts of this commentary. Special thanks for his contributions to B. Thompson, Department of Educational Psychology, Texas A&M University, a leader in clarifying the role of statistical hypothesis testing.

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.

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

ARMSTRONG, J. S. 1967. Derivation of theory by means of factor analysis, or Tom Swift and his electric factor analysis machine. American Statistician 21:17–21.

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

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

BOX, G. E. P. 1957. Evolutionary operation: a method for increasing industrial productivity. Applied Statistics 6:81–101.

———, AND N. R. DRAPER. 1969. Evolutionary operation: a statistical method for process improvement. Wiley, New York, USA.

———, AND ———. 1982. Evolutionary operation (EVOP). Pages 564–572 *in* S. Kotz and N. L. Johnson, editors-in-chief. Encyclopedia of statistical sciences. Volume 2. Wiley, New York, USA.

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

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

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

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

CUMMING, G., AND S. FINCH. 2001. A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. Educational and Psychological Measurement 61:532–574.

DRAPER, N. R., I. GUTTMAN, AND H. KANEMASU. 1971. The distribution of certain regression statistics. Biometrika 58:295–298.

GUTHERY, F. S., J. J. LUSK, AND M. J. PETERSON. 2001. The fall of the null hypothesis: liabilities and opportunities. Journal of Wildlife Management 65:379–384.

HURVICH, C. M., AND C.-L. TSAI. 1990. The impact of model selection on inference in linear regression. American Statistician 44:214–217.

JOHNSON, D. H. 1981. The use and misuse of statistics in wildlife habitat studies. Pages 11–19 *in* D. E. Capen, editor. The use of multivariate statistics in studies of wildlife habitat. U.S. Forest Service General Technical Report RM-87.

———. 1995. Statistical sirens: the allure of nonparametrics. Ecology 76:1998–2000.

———. 1999. The insignificance of statistical significance testing. Journal of Wildlife Management 63:763–772.

POPE, P. T., AND J. T. WEBSTER. 1972. The use of an *F*-statistic in stepwise regression procedures. Technometrics 14:327–340.

REXSTAD, E. A., D. MILLER, C. FLATHER, E. ANDERSON, J. HUPP, AND D. R. ANDERSON. 1988. Questionable multivariate statistical inference in wildlife and community studies. Journal of Wildlife Management 52:794–798.

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

———, AND H. WAINER. 2002. On the past and future of null hypothesis significance testing. Journal of Wildlife Management 66:263–271.

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

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

THOMPSON, B. 1995. Stepwise regression and stepwise discriminant analysis need not apply here: a guidelines editorial. Educational and Psychological Measurement 55:525–534.

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

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

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