Invited Paper:

SUGGESTIONS FOR PRESENTING THE RESULTS OF DATA ANALYSES

DAVID R. ANDERSON,[1,2] Colorado Cooperative Fish and Wildlife Research Unit, Room 201 Wagar Building, Colorado State University, Fort Collins, CO 80523, USA
WILLIAM A. LINK, U.S. Geological Survey, Patuxent Wildlife Research Center, Laurel, MD 20708, USA
DOUGLAS H. JOHNSON, U.S. Geological Survey, Northern Prairie Wildlife Research Center, Jamestown, ND 58401, USA
KENNETH P. BURNHAM,[1] Colorado Cooperative Fish and Wildlife Research Unit, Room 201 Wagar Building, Colorado State University, Fort Collins, CO 80523, USA

Abstract: We give suggestions for the presentation of research results from frequentist, information-theoretic, and Bayesian analysis paradigms, followed by several general suggestions. The information-theoretic and Bayesian methods offer alternative approaches to data analysis and inference compared to traditionally used methods. Guidance is lacking on the presentation of results under these alternative procedures and on nontesting aspects of classical frequentist methods of statistical analysis. Null hypothesis testing has come under intense criticism. We recommend less reporting of the results of statistical tests of null hypotheses in cases where the null is surely false anyway, or where the null hypothesis is of little interest to science or management.

JOURNAL OF WILDLIFE MANAGEMENT 65(3):373–378

Key words: AIC, Bayesian statistics, frequentist methods, information-theoretic methods, likelihood, publication guidelines.

For many years, researchers have relied heavily on testing null hypotheses in the analysis of fisheries and wildlife research data. For example, an average of 42 *P*-values testing the statistical significance of null hypotheses was reported for articles in *The Journal of Wildlife Management* during 1994–1998 (Anderson et al. 2000). This analysis paradigm has been challenged (Cherry 1998, Johnson 1999), and alternative approaches have been offered (Burnham and Anderson 2001). The simplest alternative is to employ a variety of classical frequentist methods (e.g., analysis of variance or covariance, or regression) that focus on the estimation of effect size and measures of its precision, rather than on statistical tests, *P*-values and arbitrary, dichotomous statements about statistical significance or lack thereof. Estimated effect sizes (e.g., the difference between the estimated treatment and control means) are the results useful in future meta-analysis (Hedges and Olkin 1985), while *P*-values are almost useless in these important syntheses. A second alternative is relatively new and based on criteria that estimate Kullback-Leibler information loss (Kullback and Leibler 1951). These information-theoretic approaches allow a ranking of various research hypotheses (represented by models) and several quantities can be computed to estimate a formal strength of evidence for alternative hypotheses. Finally, methods based on Bayes' theorem have become useful in applied sciences due mostly to advances in computer technology (Gelman et al. 1995).

Over the years, standard methods for presenting results from statistical hypothesis tests have evolved. The Wildlife Society (1995*a*,*b*), for example, addressed Type II errors, statistical power, and related issues. However, articles by Cherry (1998), Johnson (1999), and Anderson et al. (2000) provide reason to reflect on how research results are best presented. Anderson et al. (2000) estimated that 47% of the *P*-values reported recently in *The Journal of Wildlife Management* were naked (i.e., only the *P*-value is presented with a statement about its significance or lack of significance, without estimated effect size or even the sign of the difference being provided). Reporting of such results provides no information and is thus without meaning. Perhaps more importantly, there are thousands of null hypotheses tested and reported each year in biological journals that are clearly false on simple a priori grounds (Johnson 1999). These are called "silly nulls" and account for over 90% of the null hypotheses tested in *Ecology* and *The Journal of Wildlife Management* (Anderson et al. 2000). We seem to be failing by addressing so many trivial

[1] Employed by U.S. Geological Survey, Biological Resources Division.
[2] E-mail: anderson@cnr.colostate.edu

issues in theoretical and applied ecology. Articles that employ silly nulls and statistical tests of hypotheses known to be false severely retard progress in our understanding of ecological systems and the effects of management programs (O'Connor 2000). The misuse and overuse of *P*-values is astonishing. Further, there is little analogous guidance for authors to present results of data analysis under the newer information-theoretic or Bayesian methods.

We suggest how to present results of data analysis under each of these 3 statistical paradigms: classical frequentist, information-theoretic, and Bayesian. We make no recommendation on the choice of analysis; instead, we focus on suggestions for the presentation of results of the data analysis. We assume authors are familiar with the analysis paradigm they have used; thus, we will not provide introductory material here.

Frequentist Methods

Frequentist methods, dating back at least a century, are much more than merely test statistics and *P*-values. *P*-values resulting from statistical tests of null hypotheses are usually of far less value than estimates of effect size. Authors frequently report the results of null hypothesis tests (test statistics, degrees of freedom, *P*-values) when—in less space—they could often report more complete, informative material conveyed by estimates of effect size and their standard errors or confidence intervals (e.g., the effect of the neck collars on winter survival probability was –0.036, SE = 0.012).

The prevalence of testing null hypotheses that are uninteresting (or even silly) is quite high. For example, Anderson et al. (2000) found an average of 6,188 *P*-values per year (1993–1997) in *Ecology* and 5,263 per year (1994–1998) in *The Journal of Wildlife Management* and suggested that these large frequencies represented a misuse and overuse of null hypothesis testing methods. Johnson (1999) and Anderson et al. (2000) give examples of null hypotheses tested that were clearly of little biological interest, or were entirely unsupported before the study was initiated. We strongly recommend a substantial decrease in the reporting of results of null hypothesis tests when the null is trivial or uninteresting.

Naked *P*-values (i.e., those reported without estimates of effect size, its sign, and a measure of precision) are especially to be avoided. Nonparametric tests (like their parametric counterparts) are based on estimates of effect size, although usually only the direction of the effect is reported (a nearly naked *P*-value). The problem with naked and nearly naked *P*-values is that their magnitude is often interpreted as indicative of effect size. It is misleading to interpret that small *P*-values indicate large effect sizes because small *P*-values can also result from low variability or large sample sizes. *P*-values are not a proper strength of evidence (Royall 1997, Sellke et al. 2001).

We encourage authors to carefully consider whether the information they convey in the language of null hypothesis testing could be greatly improved by instead reporting estimates and measures of precision. Emphasizing estimation over hypothesis testing in the reporting of the results of data analysis helps protect against the pitfalls associated with the failure to distinguish between statistical significance and biological significance (Yoccoz 1991).

We do not recommend reporting test statistics and *P*-values from observational studies, at least not without appropriate caveats (Sellke et al. 2001). Such results are suggestive rather than conclusive given the observational nature of the data. In strict experiments, these quantities can be useful, but we still recommend a focus on the estimation of effect size rather than on *P*-values and their supposed statistical significance.

The computer output of many canned statistical packages contains numerous test statistics and *P*-values, many of which are of little interest; reporting these values may create an aura of scientific objectivity when both the objectivity and substance are often lacking. We encourage authors to resist the temptation to report dozens of *P*-values only because these appear on computer output.

Do not claim to have proven the null hypothesis; this is a basic tenet of science. If a test yields a nonsignificant *P*-value, it may not be unreasonable to state that "the test failed to reject the null hypothesis" or that "the results seem consistent with the null hypothesis" and then discuss Type I and II errors. However, these classical issues are not necessary when discussing the estimated effect size (e.g, "The estimated effect of the treatment was small," and then give the estimate and a measure of precision).

Do not report estimated test power after a statistical test has been conducted and found to be nonsignificant, as such post hoc power is not meaningful (Goodman and Berlin 1994). A priori power and sample size considerations are important in planning an experimental design, but estimates of post hoc power should not be reported (Gerard et al. 1998, Hoenig and Heisey 2001).

Information-Theoretic Methods

These methods date back only to the mid-1970s. They are based on theory published in the early 1950s and are just beginning to see use in theoretical and applied ecology. A synthesis of this general approach is given by Burnham and Anderson (1998). Much of classical frequentist statistics (except the null hypothesis testing methods) underlie and are part of the information-theoretic approach; however, the philosophy of the 2 paradigms is substantially different.

As part of the Methods section of a paper, describe and justify the a priori hypotheses and models in the set and how these relate specifically to the study objectives. Avoid routinely including a trivial null hypothesis or model in the model set; all models considered should have some reasonable level of interest and scientific support (Chamberlin's [1965] concept of "multiple working hypotheses"). The number of models (R) should be small in most cases. If the study is only exploratory, then the number of models might be larger, but this situation can lead to inferential problems (e.g., inferred effects that are actually spurious; Anderson et al. 2001). Situations with more models than samples (i.e., $R > n$) should be avoided, except in the earliest phases of an exploratory investigation. Models with many parameters (e.g., $K \sim 30$–200) often find little support, unless sample size or effect sizes are large or if the residual variance is quite small.

A common mistake is the use of Akaike's Information Criterion (AIC) rather than the second-order criterion, AIC_c. Use AIC_c (a generally appropriate small-sample version of AIC) unless the number of observations is at least 40 times the number of explanatory variables (i.e., $n/K > 40$ for the biggest K over all R models). If using count data, provide some detail on how goodness of fit was assessed and, if necessary, an estimate of the variance inflation factor (c) and its degrees of freedom. If evidence of overdispersion (Liang and McCullagh 1993) is found, the log-likelihood must be computed as $\log_e(\mathcal{L})/\hat{c}$ and used in $QAIC_c$, a selection criterion based on quasi-likelihood theory (Anderson et al. 1994). When the appropriate criterion has been identified (AIC, AIC_c, or $QAIC_c$), it should be used for all the models in the set.

Discuss or reference the use of other aspects of the information-theoretic approach, such as model averaging, a confidence set on models, and examination of the relative importance of variables. Define or reference the notation used (e.g., K, Δ_i, and w_i). Ideally, the variance component due to model selection uncertainty should be included in estimates of precision (i.e., unconditional vs. conditional standard errors) unless there is strong evidence favoring the best model, such as an Akaike weight (w_i) > about 0.9.

For well-designed, true experiments in which the number of effects or factors is small and factors are orthogonal, use of the full model will often suffice (rather than considering more parsimonious models). If an objective is to assess the relative importance of variables, inference can be based on the sum of the Akaike weights for each variable, across models that include that variable, and these sums should be reported (Burnham and Anderson 1998:140–141). Avoid the implication that variables not in the selected (estimated best) model are unimportant.

The results should be easy to report if the Methods section outlines convincingly the science hypotheses and associated models of interest. Show a table of the value of the maximized log-likelihood function ($\log(\mathcal{L})$), the number of estimated parameters (K), the appropriate selection criterion (AIC, AIC_c, or $QAIC_c$), the simple differences (Δ_i), and the Akaike weights (w_i) for models in the set (or at least the models with some reasonable level of support, such as where $\Delta_i < 10$). Interpret and report the evidence for the various science hypotheses by ranking the models from best to worst, based on the differences (Δ_i), and on the Akaike weights (w_i). Provide quantities of interest from the best model or others in the set (e.g., $\hat{\sigma}^2$, coefficients of determination, estimates of model parameters and their standard errors). Those using the Bayesian Information Criterion (BIC; Schwarz 1978) for model selection should justify the existence of a true model in the set of candidate models (Methods section).

Do not include test statistics and P-values when using the information-theoretic approach since this inappropriately mixes differing analysis paradigms. For example, do not use AIC_c to rank models in the set and then test if the best model is significantly better than the second best model (no such test is valid). Do not imply that the information-theoretic approaches are a test in any sense. Avoid the use of terms such as significant and not significant, or rejected and not rejected; instead view the results in a strength of evidence context (Royall 1997).

If some analysis and modeling were done after the a priori effort (often called data dredging), then make sure this procedure is clearly explained when such results are mentioned in the

Discussion section. Give estimates of the important parameters (e.g., effect size) and measures of precision (preferably a confidence interval).

Bayesian Methods

Although Bayesian methods date back over 2 centuries, they are not familiar to most biologists. Bayesian analysis allows inference from a posterior distribution that incorporates information from the observed data, the model, and the prior distribution (Schmitt 1969, Ellison 1996, Barnett 1999, Wade 2000). Bayesian methods often require substantial computation and have been increasingly applied since computers have become widely available (Lee 1997).

In reporting results, authors should consider readers' lack of familiarity with Bayesian summaries such as odds ratios, Bayes factors, and credible intervals. The clarity of a presentation can be greatly enhanced by a simple explanation after the first references to such quantities: "A Bayes factor of 4.0 indicates that the ratio of probabilities for Model 1 and Model 2 is 4 times larger when computed using the posterior rather than the prior."

Presentations of Bayesian analyses should report the sensitivity of conclusions to the choice of the prior distribution. This portrayal of sensitivity can be accomplished by including overlaid graphs of the posterior distributions for a variety of reasonable priors or by tabular presentations of credible intervals, posterior means, and medians.

An analysis based on flat priors representing limited or vague prior knowledge should be included in the model set. When the data seem to contradict prevailing thought, the strength of the contradiction can be assessed by reporting analyses based on priors reflecting prevailing thought.

Generally, Bayesian model-checking should be reported. Model-checks vary among applications, and there are a variety of approaches even for a given application (Carlin and Lewis 1996, Gelman and Meng 1996, Gelman et al. 1995). One particularly simple and easily implemented check is a posterior predictive check (Rubin 1984). The credibility of the results will be enhanced by a brief description of model-checks performed, especially as these relate to questionable aspects of the model. A lengthy report of model-checks will usually not be appropriate, but the credibility of the published paper will often be enhanced by reporting the results of model-checks.

The implementation of sophisticated methods for fitting models, such as Markov Chain Monte Carlo (MCMC; Geyer 1992) should be reported in sufficient detail. In particular, MCMC requires diagnostics to indicate that the posterior distribution has been adequately estimated.

General Considerations Concerning the Presentation of Results

The standard deviation (SD) is a descriptive statistic, and the standard error (SE) is an inferential statistic. Accordingly, the SD can be used to portray the variation observed in a sample: $\hat{\mu} = 100$, SD = 25 suggests a much more variable population than does $\hat{\mu} = 100$, SD = 5. The expected value (i.e., an average over a large number of replicate samples) of the SD^2 equals σ^2 and depends very little on sample size (n). The SE is useful to assess the precision (repeatability) of an estimator. For example, in a comparison of males (m) and females (f), $\hat{\mu}_m = 100$, SE = 2 and $\hat{\mu}_f = 120$, SE = 1.5 would allow an inference that the population mean value μ is greater among females than among males. Such an inference rests on some assumptions, such as random sampling of a defined population. Unlike the SD, the SE decreases with increasing sample size.

When presenting results such as $a \pm b$, always indicate if b is a SD or a SE or is $t \times$ SE (indicating a confidence limit), where t is from the t distribution (e.g., 1.96 if the degrees of freedom are large). If a confidence interval is to be used, give the lower and upper limits as these are often asymmetric about the estimate. Authors should be clear concerning the distinction between precision (measured by variances, standard errors, coefficients of variation, and confidence intervals) and bias (an average tendency to estimate values either smaller or larger than the parameter; see White et al. 1982:22–23).

The Methods section should indicate the $(1-\alpha)\%$ confidence level used (e.g., 90, 95, or 99%). Information in tables should be arranged so that numbers to be compared are close to each other. Excellent advice on the visual display of quantitative information is given in Tufte (1983). Provide references for any statistical software and specific options used (e.g., equal or unequal variances in t-tests, procedure TTEST in SAS, or a particular Bayesian procedure in BUGS). The Methods section should always provide sufficient detail so that the reader can understand what was done.

In regression, discriminant function analysis, and similar procedures, one should avoid the term independent variables because the variables are rarely independent among themselves or with the response variable. Better terms include

explanatory or predictor variables (see McCullagh and Nelder 1989:8).

Avoid confusing low frequencies with small sample sizes. If one finds only 4 birds on 230 plots, the proportion of plots with birds can be precisely estimated. Alternatively, if the birds are the object of study, the 230 plots are irrelevant, and the sample size (4) is very small.

It is important to separate analysis of results based on questions and hypotheses formed before examining the data from results found after sequentially examining the results of data analyses. The first approach tends to be more confirmatory, while the second approach tends to be more exploratory. In particular, if the data analysis suggests a particular pattern leading to an interesting hypothesis then, at this midway point, few statistical tests or measures of precision remain valid (Lindsey 1999*a*,*b*; White 2000). That is, an inference concerning patterns or hypotheses as being an actual feature of the population or process of interest are not well supported (e.g., likely to be spurious). Conclusions reached after repeated examination of the results of prior analyses, while interesting, cannot be taken with the same degree of confidence as those from the more confirmatory analysis. However, these post hoc results often represent intriguing hypotheses to be readdressed with a new, independent set of data. Thus, as part of the Introduction, authors should note the degree to which the study was exploratory versus confirmatory. Provide information concerning any post hoc analyses in the Discussion section.

Statistical approaches are increasingly important in many areas of applied science. The field of statistics is a science, with new discoveries leading to changing paradigms. New methods sometimes require new ways of effectively reporting results. We should be able to evolve as progress is made and changes are necessary. We encourage wildlife researchers and managers to capitalize on modern methods and to suggest how the results from such methods might be best presented. We hope our suggestions will be viewed as constructive.

LITERATURE CITED

ANDERSON, D. R., K. P. BURNHAM, W. R. GOULD, AND S. CHERRY. 2001. Concerns about finding effects that are actually spurious. Wildlife Society Bulletin 29:311–316.

———, ———, AND W. L. THOMPSON. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912–923.

———, ———, AND G. C. WHITE. 1994. AIC model selection in overdispersed capture–recapture data. Ecology 75:1780–1793.

BARNETT, V. 1999. Comparative statistical inference. John Wiley & Sons, 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.

———, AND ———. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research 28: in press.

CARLIN, B. P., AND T. A. LEWIS. 1996. Bayes and empirical Bayes methods for data analysis. Chapman & Hall, 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.

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

GELMAN, A., J. B. CARLIN, H. S. STERN, AND D. B. RUBIN. 1995. Bayesian data analysis. Chapman & Hall, New York, USA.

———, AND X.-L. MENG. 1996. Model checking and model improvement. Pages 189–201 *in* W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. Markov Chain Monte Carlo in practice. Chapman & Hall, New York, USA.

GERARD, P. D., D. R. SMITH, AND G. WEERAKKODY. 1998. Limits of retrospective power analysis. Journal of Wildlife Management 62:801–807.

GEYER, C. J. 1992. Practical Markov chain Monte Carlo (with discussion). Statistical Science 7:473–503.

GOODMAN, S. N., AND J. A. BERLIN. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals of Internal Medicine 121:200–206.

HEDGES, L. V., AND I. OLKIN. 1985. Statistical methods for meta-analysis. Academic Press, London, United Kingdom.

HOENIG, J. M., AND D. M. HEISEY. 2001. The abuse of power: the pervasive fallacy of power calculation for data analysis. The American Statistician 55:19–24.

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

KULLBACK, S., AND R. A. LEIBLER. 1951. On information and sufficiency. Annals of Mathematical Statistics 22:79–86.

LEE, P. M. 1997. Bayesian statistics, an introduction. Second edition. John Wiley & Sons, New York, USA.

LIANG, K-Y., AND P. MCCULLAGH. 1993. Case studies in binary dispersion. Biometrics 49:623–630.

LINDSEY, J. K. 1999*a*. Some statistical heresies. The Statistician 48:1–40.

———. 1999*b*. Revealing statistical principles. Oxford University Press, New York, USA.

MCCULLAGH, P., AND J. A. NELDER. 1989. Generalized linear models. Chapman & Hall, London, United Kingdom.

O'CONNOR, R. J. 2000. Why ecology lags behind biology. The Scientist 14(20):35.

ROYALL, R. M. 1997. Statistical evidence: a likelihood

paradigm. Chapman & Hall, London, United Kingdom.

RUBIN, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Annals of Statistics 12:1151–1172.

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

SCHWARZ, G. 1978. Estimating the dimension of a model. Annals of Statistics 6:461–464.

SELLKE, T., M. J. BAYARRI, AND J. O. BERGER. 2001. Calibration of *p* values for testing precise null hypotheses. The American Statistician 55:62–71.

TUFTE, E. R. 1983. The visual display of quantitative information. Graphic Press, Cheshire, Connecticut, USA.

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

WHITE, G. C., D. R. ANDERSON, K. P. BURNHAM, AND D. L. OTIS. 1982. Capture–recapture and removal methods for sampling closed populations. Los Alamos National Laboratory Report LA-8787-NERP, Los Alamos, New Mexico, USA.

WHITE, H. 2000. A reality check for data snooping. Econometrica 68:1097–1126.

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

———. 1995*b*. Journal news. Journal of Wildlife Management 59:630.

YOCCOZ, N. G. 1991. Use, overuse, and misuse of significance tests in evolutionary biology and ecology. Bulletin of the Ecological Society of America 72:106–111.