Statistical tests in publications of The Wildlife Society

Steve Cherry

The 1995 issue of *The Journal of Wildlife Management* (the *Journal*) has >2,400 *P*-values. I believe that is too many. In this article I argue that authors who publish in the *Journal* and in the *Wildlife Society Bulletin* (the *Bulletin*) are overusing and misusing hypothesis tests. They are conducting too many unnecessary tests, and they are making common mistakes in carrying out and interpreting the results of the tests they conduct. A major cause of the overuse of testing in the *Journal* and the *Bulletin* seems to be the mistaken belief that testing is necessary in order for a study to be valid or scientific.

The opinions presented below are not new or unique. The eagerness with which scientists embraced testing has been questioned before, with 1 critic labeling the attraction a religion (Salsburg 1985). It may be surprising to many wildlife researchers to learn that some of the strongest critics of testing are statisticians. The statistician Frank Yates wrote, "It [statistical testing] has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data . . . and too little to the estimates of the magnitude of the effects they are estimating" (Yates 1951:32). Cox (1977) began his article on significance testing with "Overemphasis on tests of significance at the expense especially of interval estimation has long been condemned...." In the President's Address at the 1996 meeting of the Western North American Region of the Biometric Society, Professor John Nelder lamented the "malign influence of *P*-values" and asked, "Why do editors think that *P*-value dominated analysis constitutes a scientific procedure?" There have been recent attacks on testing in psychology and epidemiology (Tukey 1969; Rothman 1986; Simon 1986; Rosnow and Rosenthal 1989; Cohen 1990, 1995; Goodman 1992, 1993). Others have criticized the use of testing in ecology (Yoccoz 1991, Johnson 1995). I will offer little new outside of bringing these ideas to wildlife professionals who publish, read, and review articles in the *Journal* and the *Bulletin*. I should add that many authors and reviewers learned how to do statistics from statisticians, and so statisticians who teach this material (and I am one of those) must share blame for the current state of affairs.

I will discuss 4 major problem areas. The first is the widespread use of testing in situations where it is not warranted. The second is the apparent confusion over power and its interpretation. This confusion is leading authors to draw the wrong conclusion when they observe high *P*-values and low power. The third is how authors assess the validity of assumptions of the tests they use. The fourth is the widespread use of fixed-level testing. I restrict my comments to simple procedures like *t*-tests, and simple linear regression, because most wildlife researchers are familiar with these.

I do not make a distinction between Fisherian significance testing and Neyman-Pearson hypothesis testing. The manner in which tests are carried out in the *Journal* and the *Bulletin* is a hybrid of these 2 approaches. The distinction has been addressed elsewhere and readers who are interested can refer to Cohen (1990), Goodman (1992), and Moore and McCabe (1993). It is worth pointing out that the 2 approaches are different. Fisher thought the alternative hypothesis, Type I and II errors, and power ridiculous concepts. *P*-values and their interpretation played no role in the Neyman-Pearson theory.

Unnecessary tests

Many of the tests reported in the *Journal* and the *Bulletin* are unnecessary. Three specific categories have been chosen to illustrate this point: testing in

Author's address: Department of Mathematics, Montana Tech of The University of Montana, Butte, MT 59701, USA.

Key words: confidence intervals, hypothesis testing, *P*-value, power, significance testing

habitat-use-availability studies, testing in regression, and testing biologically insignificant or obvious results.

Habitat-use–availability studies

Neu et al. (1974) has been cited hundreds of times in the published literature. The times it has been cited in unpublished theses and reports surely number in the thousands. The approach recommended in Neu et al. (1974) starts with a chi-square goodness-of-fit test to determine if animals are selecting habitats in proportion to their availability. If the null hypothesis of random use is rejected, then a set of simultaneous (typically) 95% confidence intervals is constructed to determine which habitats are being selected and which are being avoided.

The chi-square test followed by construction of the intervals if the null hypothesis is rejected is an example of a hierarchical, multiple-comparisons procedure. The testing procedure is logically flawed in this case because it is possible to reject the null and not find evidence of selection or avoidance in the intervals, and it is possible to fail to reject the null hypothesis and find evidence of selection or avoidance in the intervals. Hochberg and Tamhane (1987) referred to such problems as a lack of coherence and consonance and considered any procedure lacking coherence to be fatally flawed. This in itself leads to the conclusion that wildlife researchers should avoid the testing procedure, but there is a second problem with the approach that is even more fundamental: animals do not choose habitats in proportion to their availability. Why conduct a test to determine if they do?

The question of habitat-use-availability is more properly approached as an estimation problem. That is, researchers should not be asking the question "Are animals using habitats in proportion to their availability," but rather "How much time are animals spending in the available habitats." One approach to answering this question is to omit the test and simply construct the set of confidence intervals. If the relevant assumptions are met, the intervals are valid regardless of the outcome of the test (Byers et al. 1984, Cherry 1996). One can compare the available proportions with the endpoints of the intervals to see which habitats are being selected or avoided.

I believe that it is better to avoid the question of testing altogether. However, if one treats the confidence interval procedure as a test with the same null and alternative hypotheses as in Neu et al. (1974) the probability of a Type I error is being controlled (conservatively) at a rate equal to 1 minus the simultaneous confidence level. The standard large sample intervals used in Neu et al. (1974), and explained further in Byers et al. (1984), may not be the best to use. Cherry (1996) discussed some alternatives. Tukey (1991) offered further justification for the use of intervals in multiple-comparisons procedures. Many of the alternatives to the Neu et al. (1974) method suffer from the same problem, notably the increasingly popular method of Aebischer and Robertson (1993). There are approaches to the problem of habitat selection that rely on more sophisticated methods, but constructing a set of intervals is easy to do and there will be many times when it is a perfectly reasonable thing to do.

Testing in regression

One of the more common figures to appear in the *Journal* and the *Bulletin* shows the results of conducting a simple linear regression of a response variable on an explanatory variable. The figure shows the data and some or all of the following: a regression line and the estimated equation, a *P*-value, and an *R*-squared value. The *P*-value provides evidence that a significant relationship exists and the *R*-squared value is a measure, seemingly, of the goodness of the fit. In particular, the *R*-squared value seems to be understood as a measure of the predictive capability of the model. The null hypothesis is that the slope parameter in the model is equal to 0. The alternative is that the slope is not equal to 0. In the majority of cases in the *Journal* and the *Bulletin*, these are uninteresting hypotheses. The null is false; a relationship does exist. The more interesting question is "What is the relationship?" In general, there is little mention of the standard error of the slope of the regression line, but that is far more important for answering the real question of interest than the *P*-value or the *R*-squared value. The lack of this appropriate measure of uncertainty is particularly puzzling in articles where the main use of the published regression results is prediction (e.g., Millspaugh and Brundige 1996). The *R*-squared value does not provide an adequate measure of the predictive capability of a model (Neter et al. 1996:82–83). It is possible to have a high *R*-squared value and still have prediction intervals so wide as to be practically useless. A regression equation to be used in prediction should always be accompanied by the associated prediction intervals, or enough information to allow readers to construct them.

Testing obvious hypotheses

The 2 problem areas discussed above dealt with this topic in some fashion. I wish to go further and address the unnecessary testing (or at least the unnecessary reporting of the results of testing) in 2 cases: (1) testing when there is no difference or ef-

fect of biological significance, and (2) testing when the results are obvious. The main message is this: it is not necessary to test every result.

Introductory statistics texts (e.g., Moore and McCabe 1993) often point out that statistically significant results do not necessarily mean that one has found something of importance. Observed biologically insignificant effects do not need to be tested. I realize that the determination of what constitutes a biologically significant result is subjective. But researchers should at least look at the results of their studies and ask themselves if the observed effects are biologically meaningful before carrying out tests. One clear-cut example can be found in Franklin and Johnson (1994:255) where a 2-sample *t*-test was conducted when the 2-sample means had the same value.

Pre-test power calculations can provide some guidance in determining if an observed effect is biologically significant (insignificant). There is little reason to test an observed effect if it is less than the biologically meaningful effect specified in the power calculation.

Testing can be avoided when differences are obviously going to be statistically significant or insignificant. Leif (1994) compared survival rates, predation rates, and nesting success of wild versus pen-raised pheasants (*Phasianus colchicus*). Most of the observed differences were so large (e.g., pen-reared survival rate estimated at 7.8 ± 2.4% SE vs. wild survival rate estimated at 54.6 ± 6.6% SE) that testing was not needed to conclude the differences were real. The reported rates and standard errors were sufficient. Another example can be found in Hyvärinen and Nygrén (1993) in which a comparison of copper and zinc concentrations in the livers of newborn and older moose (*Alces alces*) calves is made. Their Figure 2 showed quite clearly that there was a difference. The graphical evidence was overwhelming, and a test was not required.

I am not advocating that researchers pore over their data in a search for significance. This is not appropriate. But researchers should have well-defined questions to answer, and some idea of the biological differences and effects that are of interest to them before beginning a study. They do not necessarily need a statistical test to determine if the differences or effects they are looking for exist or do not exist.

Some might contend that it does no harm to carry out tests and report the *P*-values anyway, but I disagree. Tests should be conducted only when and where necessary. Cluttering up the pages of the *Journal* and the *Bulletin* with hundreds of unnecessary *P*-values leads to a perception that they are a required part of valid scientific research and that investigators cannot be trusted to draw a valid conclusion on the basis of the biological or visual evidence alone. This is simply not true. Another consequence of the overemphasis on testing is that investigators may be forced to carry out difficult, sophisticated analyses in order to get a result published even when the effect is obvious. As Cox and Snell (1981: 24) put it, "...pressures to apply...tests of significance to totally obvious results should be resisted."

Misunderstanding power

There has been an increase in reporting the power of statistical tests in the *Journal* and the *Bulletin*. Power is the probability of rejecting the null hypothesis when it is false. It is a function of the Type I error rate, the sample size, and the effect size. Typically power calculations are carried out prior to running a study. Acceptable levels of error for both Type I and Type II errors are specified and the sample size necessary to achieve those levels is determined. Typical levels for Type I error rates are 0.05 or 0.10, and typical acceptable error rates for Type II errors are 0.20 or 0.10 (corresponding to power of 0.80 or 0.90, respectively). Also needed in this preliminary power calculation is a decision by the investigator(s) as to what constitutes a meaningful effect size, i.e., a preliminary power calculation requires researchers to consider in an explicit way just what constitutes a biologically significant result. That is one of the most beneficial aspects of computing power. If the results of the initial power calculations reveal that one needs an unacceptably large sample to detect a meaningful difference at the specified error rates, then the only options are to adjust those rates or to adjust the definition of a biologically significant effect. It may turn out that the results of the power calculation imply that the study is not worth doing.

Power calculations are also done in a post-test situation if the results of the test fail to yield significant results. This is a common situation in the *Journal* and the *Bulletin*. In either case it is inappropriate for authors to claim that there is no effect if the results of a test yield a nonsignificant *P*-value and low power. In other words, failing to reject the null hypothesis of no effect is not the same thing as accepting the null hypothesis.

For example, Cotter and Gratto (1995:95) stated, "Adult hens did not produce..." and reported the results of a test with a *P*-value of 0.92 and power of 0.17. It is important to understand what this means. They were saying that if a specified biologically meaningful difference existed prior to testing, then they had an estimated 17% chance of detecting it

with their sample size. However, the claim of no effect was not warranted because they had little chance of detecting the effect even if it had existed.

Cotter and Gratto (1995) appeared to have conducted post-test power analyses. Wilson et al. (1995) reported that pre-test power calculations yielded power of 0.27. They then reported 97 *P*-values, most of which led to the conclusion that there was no effect. However, given that the 97 null hypotheses were all false and the specified biologically meaningful effects actually existed, they would have made fewer Type II errors determining significance by flipping a fair coin 97 times and denoting heads significant and tails not significant. This criticism is not meant to imply that Wilson et al. (1995) did not learn something useful from their study. But nonsignificant statistical results from tests with power equal to 0.27 did not shed any light on what they did learn.

Testing assumptions

Statistical tests come with an assortment of assumptions. Most researchers are aware of these. However, it is all too common to find the assumptions ignored. Even when authors do evaluate the adequacy of assumptions, they tend to focus on the wrong ones and ignore the most important one. By way of example, the subsequent discussion deals mostly with the commonly conducted pooled variance 2-sample *t*-test.

The relevant assumptions for the pooled variance 2-sample *t*-test are that the data must be in the form of 2 simple random samples from 2 normally distributed populations with the same variance. It is not uncommon to read that the assumption of normality was checked. This is frequently followed by reports of some kind of transformation or a switch to a nonparametric test. However, if one has a large enough sample to conduct a meaningful test of normality one probably does not need to be worried about nonnormality. The distribution of the test statistic in the pooled *t*-test will follow a Student's *t* distribution if the sampling distribution of the sample means is normal, the variances of the 2 populations are equal, and if the 2 samples are simple random samples from the 2 target populations. The only way to ensure this with small sample sizes is to require that the data be normally distributed, and with small sample sizes investigators generally have to live with this assumption because there is no good way to test it. However, the Central Limit Theorem implies that with large enough sample sizes the sampling distributions of the sample means will be approximately normal regardless of the distribution of the underlying populations. What is large enough? There is no clear-cut answer. It is always dangerous to give prescribed guidelines, but if sample sizes are >40 (for each sample) one typically does not need to worry about nonnormality. The test is remarkably robust for sample sizes ≥15, particularly if the 2 sample sizes are nearly equal (Moore and McCabe 1993). Most normality tests have low power when sample sizes are small enough for nonnormality to be a problem in a *t*-test. If the sample size is large enough for the normality test to have sufficient power to detect nonnormality, then the sample size is probably large enough for the sampling distributions of the sample means to be approximately normal. In short, most of the normality tests reported in the *Journal* and the *Bulletin* are unnecessary. Investigators should check for nonnormality but visual assessment using normal probability plots is almost always adequate. Further, probability plots are more likely to highlight outliers, which tend to be more of a problem than general skewness.

When a suitable transformation cannot be found, researchers often use nonparametric methods. The Mann–Whitney U test is a frequent choice for a nonparametric 2-sample test. One of the assumptions of the test is that both samples come from populations with the same distribution with one shifted to the right of the other. If one distribution is skewed and the other is symmetric, the test is not appropriate. The Mann–Whitney test is more sensitive to violations of its distributional assumption than the 2-sample *t*-test. I have never seen an example of an article in the *Journal* or the *Bulletin*in in which an author or authors attempted to determine if the distributional assumption, or any other assumption (one of which is equal variances), of the Mann–Whitney test was met for their data. Nonparametric implies distribution free, not assumption free. Johnson (1995) discusses this and other problems associated with the use of nonparametric procedures in ecology.

The fear of nonnormality not only leads researchers to unnecessarily transform their data or fall prey to the "statistical siren" of nonparametrics (Johnson 1995). It overshadows the more serious problems caused by violation of the assumption of simple random sampling. This absolutely critical assumption receives scant attention, and examples of inappropriately applied tests due to violations of this assumption are common.

Gorenzel and Salmon (1995) conducted tests to determine if differences existed between urban roost and nonroost trees of crows (*Corvus brachyrhyncos*). They tested for normality (although they had quite reasonable sample sizes of 87 roost trees and 62 nonroost trees) and commented on how they han-

dled unequal variances. But the description of how they sampled the nonroost trees makes it clear that it was not a simple random sample. They gridded their study area, chose grid squares at random, and then randomly selected a nonroost tree in the grid. The essence of simple random sampling is that every sample of a specified size has an equal chance of being selected and this was not true for the nonroost trees in this study. Thus, the standard errors of the difference between the sample means could not be computed using the standard formula, and any *t*-tests conducted using such a standard error calculation were invalid.

I do not wish to criticize the above authors too harshly. Wildlife studies are inherently difficult and getting adequate data is challenging. I know that wildlife researchers worry a great deal about getting a representative sample from the population of interest to them, and they have little if any control over a host of potentially biasing factors. Frankly, if Gorenzel and Salmon (1995) had explicitly noted that they did not have a simple random sample of control trees but were going to act as if they did I would not have chosen their paper as an example. The point I am trying to make is that they invested a great deal of effort evaluating normality and equal variance assumptions and ignored the I assumption that is truly critical for the validity of their results.

Moore and McCabe (1993:472–487) argued that, if statistical inference is the goal, then researchers need to have confidence in a probability model, i.e., a mathematical model that serves as an idealization of the process that produced their data. The assumption of simple random sampling ensures the probability laws apply. Wildlife researchers should be aware that many of the statistical methodologies they use were developed for analysis of data from randomized controlled experiments and applications of those methods to nonrandom data from observational studies must be done with a great deal of care. I agree with Moore and McCabe (1993) that while statistical analysis of such data may be necessary, low *P*-values alone provide little evidence against null hypotheses in these types of studies.

The myth of 0.05

Fixed-level testing is used almost without exception in the *Journal* and the *Bulletin*. The most common level of significance is $\alpha = 0.05$. If the *P*-value lies below this conventional level, then the typical investigator will claim the presence of an effect. Moore and McCabe (1993) argued that choosing a level of significance in advance implies that research is a decision-making process. The goal of research is not to make a decision, but to provide an incremental increase in understanding. Decisions are made when scientists as a group reach a consensus, after evaluating evidence from many studies.

There is nothing sacred about $\alpha = 0.05$ (or any other level). There is no clean demarcation between finding an effect and not finding an effect. A *P*-value provides a measure of the strength of the evidence against a specified null hypothesis. This is a continuous scale of measurement with low *P*-values providing more evidence than large ones. Authors would serve their science better by giving the *P*-value and commenting on what they believe they have learned, leaving it to others to agree or disagree with them. Time and replication will eventually determine who is correct.

The only justification for choosing a fixed level α is the role it plays in power calculations. But specifying a level of α in order to determine if a test has reasonable power does not mean one has to rigidly adhere to that level when evaluating the strength of the evidence. I believe the worst thing about fixed-level testing is that scientists engaged in the incredibly messy business of science (and field ecology is especially messy) abdicate their responsibility to evaluate the significance of a result to a canned, cookbook procedure. Scientists should never, under any circumstances, let a statistical procedure do their thinking for them.

Conclusions

It is easy to criticize. It is harder to criticize constructively, because constructive criticism carries with it the responsibility to offer recommendations for improvement. I offer some recommendations here, recommendations that I hope will prove to be constructive.

My main recommendation is for wildlife researchers to stop taking statistical testing so seriously. I believe that most statisticians would agree with Cox and Snell (1981:39) that testing has "a valuable but limited role to play in the analysis of data." The presence of some 2,400 *P*-values in volume 59 of the *Journal* indicates that the use of testing by wildlife researchers is hardly limited.

I make the following recommendations:

1. Investigators should determine if they are more interested in estimation of effects or testing for the presence or absence of effects. In most studies estimation will be more important than testing. Estimation requires investigators to consider what quantitative characteristics best measure the effects of interest followed by

point estimates, standard errors, and confidence intervals.

2. There is no reason to report the results of tests on hypotheses that are obviously false, true, or biologically insignificant. One does not necessarily need a *P*-value to determine the significance, or lack thereof, of a result.
3. Do not claim to have found no effect after failure to reject the null hypothesis on low power tests. The results of nonsignificant low power tests should not even be explicitly reported.
4. Test normality using graphical procedures. Avoid the uncritical use of nonparametric methods. Be aware of the critical nature of the assumption of simple random sampling and be cautious in interpretation of results when this assumption is violated (and it is always violated).
5. Do not decide to have found or not have found an effect based on which side of 0.05 a *P*-value happens to lie. *P*-values represent a continuous scale of measurement of the strength of evidence against a specified null hypothesis. There is no 1 value to relieve researchers of their responsibility in evaluating that evidence.
6. Learn and make use of the increasingly powerful tools available for initial data analysis. I am not suggesting that researchers go hunting for interesting patterns in their data, and then test hypotheses about those patterns using the data that generated the hypotheses. I am suggesting that if an expected effect does not show up in the initial analysis, then it is probably not worth testing for. Likewise, an expected effect may show up so strongly in the initial analysis that a test is not needed to determine if it is real.
7. Learn (really learn) the basics. The introductory text by Moore and McCabe (1993:472–487) has an excellent section on the use and abuse of tests. Chatfield (1995) is full of basic common sense advice on doing statistics. Cox and Snell (1981) is another good source.
8. The Wildlife Society should explore ways to help authors and reviewers conduct statistical analyses more effectively. The Biometrics Working Group might contribute advice. Editors and associate editors of the *Journal* and the *Bulletin* could help by making it clear that *P*-values are not a prerequisite for publication.

Some may feel that I have overstated the extent of the problem. But these are problems of a fundamental nature that are, or should be, topics of discussion in introductory statistics courses. The failure to understand the basics of statistical practice, and consequent misuse of statistics, diminishes the value of otherwise good studies.

Acknowledgments. J. B. Borkowski, R. A. Garrott, M. A. Hamilton, R. R. Rossi, J. R. Rotella, and W. Thompson read and commented on earlier drafts of the manuscript. The opinions expressed are entirely my own.

Literature cited

Aebischer, N. J., and P. A. Robertson. 1993. Compositional analysis of habitat use from animal radio tracking data. Ecology 74:1313–1325.

Beyers, C. R., R. K. Steinhorst, and P. R. Krausman. 1984. Clarification of a technique for analysis of utilization-availability data. Journal of Wildlife Management 48:1050–1053.

Cherry, S. 1996. A comparison of confidence interval methods for habitat use–availability studies. Journal of Wildlife Management 60:653–658.

Chatfield, C. 1995. Problem solving: a statistician's guide. Chapman and Hall, London, England.

Cohen, J. 1990. Things I have learned (so far). American Psychologist 45: 1304–1312.

Cohen, J. 1995. The earth is round ($P < 0.05$). American Psychologist 49: 997–1003.

Cotter, R. C., and C. J. Gratto. 1995. Effects of nest and brood visits and radio transmitters on rock ptarmigan. Journal of Wildlife Management 59:93–98.

Cox, D. R. 1977. The role of significance tests. Scandinavian Journal of Statistics 4:49–70.

Cox, D. R., and E. J. Snell. 1981. Applied statistics: principles and examples. Chapman and Hall, London, England.

Franklin, W. L., and W. E. Johnson. 1994. Hand capture of newborn open-habitat ungulates: The South American guanaco. Wildlife Society Bulletin 22:253–259.

Goodman, S. N. 1992. A comment on replication, *p*-values and evidence. Statistics in Medicine 11: 875–879.

Goodman, S. N. 1993. *P* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. American Journal of Epidemiology 137:485–496.

Gorenzel, W. P., and T. P. Salmon. 1995. Characteristics of American crow urban roosts in California. Journal of Wildlife Management 59:638–645.

Hochberg, Y., and A. J. Tamhane. 1987. Multiple comparison procedures. John Wiley and Sons, New York, New York.

Hyvärinen, H., and T. Nygrén. 1993. Accumulation of copper in the liver of moose in Finland. Journal of Wildlife Management 57:469–474.

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

Leif, A. P. 1994. Survival and reproduction of wild and pen-reared ring-necked pheasant hens. Journal of Wildlife Management 58:501–506.

Moore, D. R., and G. P. McCabe. 1993. Introduction to the practice of statistics, Second edition. W. H. Freeman and Company, New York, New York.

Millspaugh, J. J., and G. C. Brundige. 1996. Estimating elk weight from chest girth. Wildlife Society Bulletin 24:58–61.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. Applied linear statistical models, Fourth edition. Irwin Press, Burr Ridge, Illinois.

Neu, C. W., C. R. Byers, and J. M. Peek. 1974. A technique for

analysis of habitat utilization–availability data. Journal of Wildlife Management 38:541–545.

Rosnow, R. L., and R. Rosenthal. 1989. Statistical procedures and the justification of knowledge in psychological science. American Psychologist 44:1276–1284.

Rothman, K. J. 1986. Significance questing. Annals of Internal Medicine 105:445–447.

Salsburg, D. S. 1985. The religion of statistics as practiced in medical journals. American Statistician 39:220–223.

Simon, R. 1986. Confidence intervals for reporting results of clinical trials. Annals of Internal Medicine105:429–435.

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

Tukey, J. W. 1991. The philosophy of multiple comparisons. Statistical Science 6:100–116.

Wilson, C. W., R. E. Masters, and G. A. Bukenhofer. 1995. Breeding bird response to pine-grassland community restoration for red-cockaded woodpeckers. Journal of Wildlife Management 59:56–67.

Yates, F. 1951. The influence of statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association 46:19–34.

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

Steve Cherry is currently associate professor of mathematics in the Department of Mathematics at Montana Tech of The University of Montana. He received his B.S. in applied mathematics at North Carolina State University and his M.S. in ecology at the University of Tennessee. He received an M.S. and a Ph.D. in statistics from Montana State University. He worked for several years as a Conservation Officer with the Utah Division of Wildlife Resources. He was a visiting postdoctoral researcher at the National Center for Atmospheric Research in Boulder, Colorado, and an adjunct assistant professor of statistics at Montana State University for 2 years. His research interests are in the areas of environmental and ecological statistics. He also enjoys teaching the science of statistics.