esa ECOSPHERE

Applied statistics in ecology: common pitfalls and simple solutions

E. Ashley Steel,[1],† Maureen C. Kennedy,[2] Patrick G. Cunningham,[3] and John S. Stanovick[4]

[1]*Statistics, Pacific Northwest Research Station, USDA Forest Service, 400 N34th Street, Suite 201, Seattle, Washington 98103 USA*
[2]*Environmental and Forest Sciences, University of Washington, 400 N34th Street, Suite 201, Seattle, Washington 98103 USA*
[3]*Statistics, Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, Oregon 97331 USA*
[4]*Statistics, Northern Research Station, USDA Forest Service, 11 Campus Boulevard, Suite 200,*
Newtown Square, Pennsylvania 19073 USA

Citation: Steel, E. A., M. C. Kennedy, P. G. Cunningham, and J. S. Stanovick. 2013. Applied statistics in ecology: common pitfalls and simple solutions. Ecosphere 4(9):115. http://dx.doi.org/10.1890/ES13-00160.1

Abstract. The most common statistical pitfalls in ecological research are those associated with data exploration, the logic of sampling and design, and the interpretation of statistical results. Although one can find published errors in calculations, the majority of statistical pitfalls result from incorrect logic or interpretation despite correct numerical calculations. There are often simple solutions to avoiding these problems that require only patience, clarity of thinking, probabilistic insight, and a reduced reliance on out-of-the-box approaches. Some of these trouble spots are inherent to all statistical analyses and others are particularly insidious in ecology where true controls or replication are challenging, small sample sizes are common, and correctly linking mathematical constructs and ecological ideas is essential. Here we summarize the most common statistical pitfalls observed over nearly a century of combined consulting and research experience in ecological statistics. We provide short, simple solutions.

Key words: confidence intervals; errors; experimental design; graphing; mistakes; *p* values; statistics education.

Received 1 May 2013; revised 29 July 2013; accepted 30 July 2013; final version received 6 September 2013; **published** 27 September 2013. Corresponding Editor: D. P. C. Peters.

† **E-mail:** asteel@fs.fed.us

Introduction

Statistical errors are common across a wide range of disciplines and even in the most prestigious journals (e.g., Good and Hardin 2003, García-Berthou and Alcaraz 2004, Ionnidis 2005, Strasak et al. 2007, Fernandes-Taylor et al. 2011). One study concluded that approximately 44% of the articles reporting statistics in a high quality medical journal had errors (Glantz 1980). Ecological research isn't immune from statistical mistakes, misconceptions, and miscommunications. Although a few of these pitfalls are related to errors in calculations, the majority of statistical pitfalls are more insidious and result from incorrect logic or interpretation despite correct numerical calculations. Such statistical errors can be dangerous, not because a statistician might come along and hand out a demerit, but because they lead to incorrect conclusions and potentially poor management and policy recommendations.

Over two decades ago, a paper identifying the 10 most common statistical errors from a journal editor's perspective was published in the Bulletin of the Ecological Society of America (Fowler 1990). We revisit and add to the discussion from our own perspective. We have nearly a century of combined experience in statistical consulting and collaboration across a wide range of ecological topics. We note that (1) some errors, identified by

Fowler over 20 years ago, are still prevalent and, in our experience, these weak statistical practices crop up over and over again. Two decades of the same mistakes suggests that the message needs to be repeated. We outline these frequent misapplications of statistics and provide brief, simple solutions. We further contribute by (2) bringing the perspective of consulting and collaborating statisticians who often identify erroneous scientific and statistical logic that infects projects long before the publication stage evaluated by Fowler (1990). Finally, we (3) highlight new sources of opportunity and common confusion that have arisen with some of the key statistical advances over the past 20 years.

Counting up statistical mistakes is both fun and shocking but rarely constructive. Instead, we focus on identifying solutions. We choose in this manuscript to de-emphasize details of the application of statistical tools. We instead focus on statistics fundamentals, probabilistic thinking, and the development of correct statistical intuition. The following is a simple list of what are, in our estimation, the most common and often easily avoidable statistical pitfalls from a non-random, somewhat biased, but extensive sample of projects, questions, and draft manuscripts. The list is loosely organized into four chronological categories: setting up an analysis, experimental design, application of statistics, and interpretation of statistical tests and models. Within each category the pitfalls, presented in no particular order, are intended to serve as a reference or reminder when conducting or reviewing various stages of analysis. We necessarily begin with issues that reflect gaps in statistical thinking and basic statistical practices. These are, by far, the most common and easily avoidable mistakes and are relevant to nearly all statistical analyses. We then describe common errors and solutions for more specific yet also common situations.

Common Statistical Pitfalls in Setting Up an Analysis

1. *Failure to explore the data.*—It is difficult to overestimate the value of plotting data. Data sets have errors from multiple sources, e.g., faulty instrumentation, transcription errors, cut and paste mistakes. Data simply have to be cleaned and the best way to see if data are, in fact, clean is to look at them. As a first quality assurance/quality control (QA/QC) step, we recommend preparing and examining a simple histogram for each variable under consideration for analysis. Are there outliers or unexpected values? Is the distribution mound-shaped? Does the distribution have a different shape that makes sense? Is the distribution centered on what one intuitively expects to be the mean?

Additional plots can be used to understand the structure of the data such as the balance between zero and non-zero observations, presence of collinearity in the potential predictor variables, and likely relationships between predictors and response variables (Zuur et al. 2010). We also recommend preparing a plot that answers the scientific question at hand. Does the expected difference in treatments, or trend over time, appear to occur? For example plots that explore simple linear relationships, see Anscombe (1973). The conclusions you draw from statistical calculations in later stages of analysis should not be surprising after examining these graphs. Note that if in graphing the data you uncover surprising patterns or generate new hypotheses, these patterns and hypotheses cannot be tested later in traditional ways that demand a priori hypotheses. Many errors could be avoided through careful and thoughtful initial data visualization. **Solution:** Plot the data early and often.

2. *Arbitrary thresholds, metrics, and indicators*.—Identifying a threshold, metric, or indicator that is most accurate at a particular site and time and that best reflects the ecological process of interest is not a simple problem. Ecologists frequently assert statements such as, "We considered sites with a value of $x > 85$ to be 'polluted'" and then continue to consider polluted versus non-polluted areas in great detail. Clearly that threshold is part of the analysis because a change in threshold can change the outcome of the analysis. Consider, for example, a dataset in which most values are in the range of 75–95. A threshold of $x > 85$ might lead to a statement that "The majority of rivers are polluted" while a threshold $x > 86$ might lead to the opposite statement.

Magnusson (1997) included "Most ecological categories are arbitrary" as one of the essential concepts that graduates of a basic statistics course should understand. Similar problems arise with summary metrics, e.g., mean of the

7-day maximum temperature or beta-diversity, as well as with weighted indicators, e.g., forest condition index or community resilience index. There are, for example, multiple ways to compute the daily mean temperature when continuous measurements are unavailable: the average of the maximum and minimum of hourly data, the average of the true maximum and minimum from max/min thermometers, or the average of multiple hourly observations. Not only do these methods yield different estimates of the daily mean temperature but differences between these alternatives are larger in some seasons than in others (Ma and Guttorp 2013). Any metric or indicator is a lens on the data that necessarily emphasizes some factors and ignores others. As well, summaries, metrics, and indicators almost always have different sample variances and distributional properties than the original data.

Information is filtered when summaries or aggregations are calculated or when continuous data are broken into categories. It is necessary to consider the effect of that filtering on the inference made from the data. Mallows (1998) defined the zeroth problem in statistics, the most basic problem in theoretical statistics, as "considering the relevance of the observed data, and other data that might be observed, to the substantive problem" (Mallows 1998:2). Considering the relevance of the threshold, metric, or indicator and of other thresholds, metrics, or indicators is an essential part of this most fundamental problem in statistics. The researcher needs to consider (1) whether a chosen threshold, metric, or indicator is, in fact, the most relevant way to summarize information and (2) how the choice of a different threshold, metric, or indicator might have led to different scientific conclusions. It is also valuable to step back and consider whether there is a useful method for the question at hand that doesn't require summary metrics such as time-series analysis or multivariate statistics. For more on defining an ecological research question, see 'Scientific Method for Ecological Research' (Ford 2000). **Solution:** Be explicit in the ecological concept being measured and in the choice and calculation of thresholds, metrics, and indicators. Where there are alternative choices, quantify how those choices might lead to differing conclusions.

3. Assuming that observations are independent.— There are many ways in which two (or more) observations can be dependent and the relationship between these observations depends on the scale of the specific research question at hand. Observations can lack independence through spatial, temporal, or phylogenetic relationships (Magnusson 1997). The two most common forms of dependence in ecology are repeated measures (Gurevitch and Chester 1986) and pseudo-replication (Hurlbert 1984). These problems generally occur when a scientist confuses the unit of observation (sampling unit) and the unit of inference (experimental unit). In forestry research, for example, randomly selected replicate plants are nested within a plot because individual measurements on every seedling or tree at the plot level can't be measured. In this case the replicate plants are the sampling unit and the plot with its associated treatment is the experimental unit.

While it makes good sense to make several observations about one thing for which you want to make inference, be careful. For example, one might want to compare the strength of various classes of wood adhesives and, in so doing, make several observations of the strength of one application of one wood adhesive. Multiple measurements or observations of the same application of adhesive are not independent observations given this research question and therefore cannot be entered into a statistical model as independent data. Doing so would be a clear case of pseudo-replication; the unit of inference is the application of wood adhesive. In another example, one might want to compare growth rates of two subspecies of alpine fir. The unit of inference would be an independent tree, not a measurement of tree size at one point in time. The true sample size of an experiment or observational study is the number of independent observations of that unit of inference. The problem gets trickier with large-scale studies. For example, to quantify effects of forest management, the unit of observation is frequently a treatment unit which might be as large as a whole watershed. One might make observations of multiple trees or even stands within each treatment unit. These are not independent observations of that forest management treatment; they cannot be entered into a statistical model as such. These are non-independent, or repeated measures, of one treatment unit.

Replicate observations can be extremely valuable if understood and handled correctly. Multiple observations of a wood adhesive application or of a particular treatment can be (1) used to assess measurement variability, (2) combined to produce a robust or more informative summary such as mean strength, growth rate, or average growth rate, (3) entered individually into a model which correctly accounts for the dependence between observations, or (4) simply explored for insight into the dependencies and structure of the data. **Solution:** Identify the unit of inference and use it to determine the true sample size.

4. Mismatched sampling frame and population.—In order to make inference to a population, the entirety of that population must have had a chance of being included in the sample. Stratified sampling designs and other customized approaches can allow individuals within a population to have different chances of being included in the sample; however, in all forms of probability sampling, all the individuals in the population must be part of the sampling frame. The classic example of a mismatched sampling frame and population is the *Literary Digest* 1936 election poll that incorrectly predicted a landslide for Alfred London. Americans who did not subscribe to the *Literary Digest*, own an automobile, or list their name in the telephone directory were not sampled. There were also a large number of non-responders who were likely quite different from the responders (Squire 1988). Inference made to these unsampled individuals was infamously biased. Despite years of using the *Literary Digest* 1936 election poll as the quintessential example of a statistical pitfall, similar errors persist. In the 2012 presidential election, one election poll analyst, Dean Chambers, predicted a win for Mitt Romney, later admitting he had eliminated poll data that seemed to over-sample Democrats (Bensinger 2012). By not including those data, he could not make inference to those Democrats and his predictions about the population of the USA were therefore inaccurate. Similar types of errors are easily possible when sampling, for example, only accessible areas for presence-absence of non-native invasive species and extrapolating to an entire forest, or when surveying natural resource volunteers to understand their incentives for participation and extrapolating the learned ideas to non-volunteers who likely have a different set of incentives. **Solution:** The sampling frame must include every individual in the population. If that is impossible, the population to which one makes inference must be redefined to match the sampling frame.

Common Pitfalls in Experimental Design

5. Control sites (or reference sites) differ from treatment sites before the treatment occurs.—Generally in designed experiments, researchers use control treatments (also known as control sites or reference sites) to provide a standard against which other treatments are contrasted. This use of control units assumes that they are indistinguishable from treated units except for the application of the experimental treatment. By being indistinguishable from treated units, these control units allow experimenters a degree of confidence that effect estimates are true estimates of the effect being studied, unpolluted by other sources of variation. If controls differ from treated units only in the experimental treatment, observed differences between treatments and controls can be attributed solely to the treatments (and their installation).

It is not uncommon, however, to encounter experiments in ecology in which control units differ systematically from the treated units. At times, control sites are identified not for scientific reasons but in response to logistical constraints, e.g., 'since we're not going to be doing anything with this land, why not leave it as a control site?' In another situation, only the steepest areas are left as controls to avoid building roads in steep areas. These types of decisions ignore the scientific value of an unbiased reference against which other treatments can be compared. Such decisions unintentionally undermine the scientific design of the experiment and often produce effect estimates and hypothesis tests that are incorrect. These incorrect effect estimates lead to misinformation in the scientific literature and frequently to overly optimistic expectations of outcomes when the experimental results are incorporated into operational strategies.

In many cases, true controls are simply impossible and this should eventually be accounted for in the analytical approach. If one is certain that relative differences between controls and treatments are constant before and after treatment, some designs can still be valid. However, if that assumption is false or if measurements are auto-correlated over time, the presence of a control site can actually reduce the power of the design (Roni et al. 2005). Collecting data prior to the experimental phase of the study can help. In this case, analysis of covariance may be used to account for differences in experimental units before treatments are added. However, accounting for pre-treatment differences cannot replace the value of an unbiased reference. Note that as experiments increase in spatial scale to whole ecosystems or watersheds, the very idea of replicated study units becomes an impossibility and careful thought needs to be applied in using traditional statistical tools. **Solution:** Assign treatment and control sites randomly. Where pre-treatment data are available, test for systematic differences between units assigned as controls and treatments (or assigned to different treatment types).

6. Measurement strategies that confound experimental designs.—Designed experiments are powerful tools for acquiring new scientific knowledge. This is as true in ecology as it is in other branches of science. Even within seemingly well-designed experiments, the timing and placement of measurements within experimental units and among treatments are critically important. One must carefully think through all potential sources of variation in the experiment in order to prevent systematic variation from being accidently allowed into the data. For example, some ecological measurements are time-sensitive. That is, the time of day, or time of year can be vitally important. If researchers are looking at, for example, gas exchange in foliage or water conductance in vascular plants, they will want to make sure that measurement times are randomly assigned across treatments to avoid consistently measuring one treatment during peak gas-exchange times or peak water-conductance times. Likewise, if there is the possibility of observer error or measurement drift, measurements must be assigned randomly or systematically across treatments so that these unintended sources of variation or bias are not lumped into the treatment effect. A good experimental or sampling design can only be as good as the measurement strategy used to collect the data. **Solution:** Ensure that measurement strategies account for time- or space-sensitive measurements in designed experiments (or sample surveys).

7. Failure to model covariates at the correct level.—Analysis of covariance is frequently misused in experimental studies, in particular when interactions are present. Unlike observational studies, testing whether the covariate is useful in an experimental setting can save degrees of freedom; however, failure to test covariates can lead to misspecified models and erroneous conclusions. For the most common analysis of covariance model, testing for an effect of treatment amounts to assessing whether parallel regression lines (one for each treatment) are far enough apart to be considered statistically significant. Since the regression lines are parallel, they can be compared at any value of the covariate, including $X = 0$. However, if there is a significant interaction, the standard analysis structure cannot simply be extended to include a separate slope for each treatment. Regression lines are no longer parallel; therefore comparison among the regression lines at different values of the covariate may lead to different conclusions about the presence or absence of a treatment affect. In this case, testing for a treatment effect needs to occur at particular and meaningful values of the covariate in order to determine whether there is a treatment affect and to understand how it might be observed given various levels of the covariate (Littell et al. 1996). **Solution:** When interactions are present between treatment levels and a covariate, look for treatment effects at meaningful values of that covariate.

Pitfalls in the Application of Statistics

8. Unnecessary data transformations.—Tools have been relatively recently developed and are now easily available for analyzing data that are not simple, normal observations. For example, it was once recommended that count data be transformed in order to meet the assumptions of Gaussian linear models; however, Gaussian models based on log-transformed data can perform poorly under many circumstances such as zero-inflation, small means, or over-dispersion (O'Hara and Kotze 2010). There is no longer a need to transform count data. Poisson data can be correctly modeled using a generalized linear model for the Poisson or negative binomial family. Similarly, arcsin transformations were routinely recommended for proportion data because we, the scientists, needed to stuff them into a Gaussian linear model. It is now possible to model proportional data correctly using generalized linear models based on the binomial family. These models are interpretable and maintain information in both the numerator and denominator (Warton and Hui 2011). Also remember that transforming the data can change the distribution of the data. **Solution:** Where possible, use a statistical tool that is designed for the distribution and correlation structure inherent in your data. Transform the response as a last resort.

9. Not dealing appropriately with zeros.—Zeros are data, they cannot be eliminated or ignored for convenience, and they are often maddeningly present in ecological data sets, e.g., number of chipmunks per ha, presence/absence of wildfire, or grams of hazardous waste in the waste stream. Inference based on data with a large number of zeros can be downright wrong or inefficient if the zeros are not appropriately understood and modeled. Zeros can arise from multiple mechanisms such as the failure to detect an event or individual that was actually present (false zero) or the absence of an event or individual (true zero). Specialized statistical models, zero-inflated models, exist to account for and model sources of zeros (e.g., Martin et al. 2005).

Often, however, the zeros arise from a different process than the non-zero observations. For example, the zeros might reflect that there was no forest fire at all while the majority of the non-zero observations describe damage from a forest fire. In these cases, a hurdle model can be a useful and simple approach. A hurdle model essentially fits one model to the presence/absence part of the data and a second conditional model (conditional on having leapt the hurdle of presence) to the non-zero observations. In the above example, the first part of the model would predict conditions under which a fire occurred and the second part of the model would predict damage, conditional on the presence of a fire. Hurdle models, though simple and relatively easy to fit, have demonstrated extremely strong performance (correlation between observed and predicted values) when compared to other potential approaches (Potts and Elith 2006). Other choices for modeling data sets with a disproportionately large number of zeros include models based on the negative binomial and quasi-Poisson distributions (Martin et al. 2005). **Solution:** Do not ignore, delete, or average away large numbers of zero observations. Consider the mechanism generating the zeros and choose a modeling approach that properly accounts for that mechanism.

10. Ignoring underlying correlation structure.—Similarly, advances in statistics allow models that incorporate complex underlying correlation structures. Advancements in linear mixed models over the last decade have provided a wealth of covariance structures to model repeated measures over space or time. For example, PROC MIXED in SAS has approximately 35 spatial and temporal covariance structures to choose from (SAS Institute 2011). New approaches are also being developed that include covariance structures based on both ecological networks and 2-dimensioal space (Peterson et al. 2013). AIC/AICc (Akaike information criteria) or BIC (Bayesian Schwartz information criteria) fit statistics can be used to select the most appropriate covariance structure. Avoid testing all available covariance structures as this is a form of multiple testing. As in any model selection protocol, use mechanistic first principles, expert opinion, previously published literature, and model diagnostics to identify subsets of reasonable covariance structures for testing.

In addition, generalized estimating equations (GEE) are fairly robust and give consistent estimators with an unstructured correlation structure (Liang and Zeger 1986), saving the issue of how to model the correlation. Data with correlation structures, imbalances, and dependencies can be attacked with models that account for these correlations, imbalances, and dependencies. While such complex data can sometimes be meaningfully explored through summaries and hierarchical analyses, a common pitfall is to ignore the correlation or to sub-sample data. **Solution:** Where measurements are not likely to be independent, explicitly model the correlation structure in the data.

11. Failure to plot the residuals (or other model *diagnostics)*.—Statistical models explain only part of the story in the data. Residual plots and model diagnostics explore the part of the story that wasn't explained by the model. The modeler's goal is that the model describes the entire signal in that data, leaving only meaningless noise in the residuals. Check by plotting the residuals as a function of the fitted values. If the residuals look like meaningless noise, you're all set. If not, you need to reconsider your model. Other model diagnostics tell you whether you have met the assumptions (usually distributional assumptions and homogeneity of variance) of the statistical model. If there are concerns about model assumptions, be honest and explain how the failure to meet model assumptions might affect conclusions from the analysis. **Solution:** Use model diagnostics such as residual plots to test model assumptions and to identify patterns in the data that are not explained by the model.

12. Conducting too many tests.—Recent articles have demonstrated that many published scientific findings are false (Ionnidis 2005). This phenomenon may, at least in part, be due to the screening of multiple graphs or tests or summary statistics followed by publication of only the interesting (or apparently 'significant') results. Researchers often run many, many statistical tests searching for a significant result. However, the very definition of a statistically significant result using an alpha-level of 0.05 is that we might find one result like that, by chance alone, for every 20 tests we conduct. If 60 tests are conducted and 3 interesting patterns emerge, that is exactly what we would expect by chance alone. Testing can effectively occur by graphical analysis alone so plotting 100 scatterplots of potential relationships and then only conducting a statistical test on the 2 most interesting plots is, essentially, undermining the system. Stepwise model-fitting procedures are highly susceptible to this statistical pitfall. Each step of the model-fitting procedure can be considered a statistical test and, if there are a large number of potential predictors, a highly significant model is likely to emerge even if the data are uncorrelated random numbers (Freedman 1983).

During data exploration, it is reasonable to conduct many tests and make many graphs but the results of these exercises need to be described as exploratory and the approximate number of tests conducted should be indicated. In many cases, a combined approach is possible in which a researcher tests an a priori hypothesis in a hypothesis-testing framework and identifies whether or not there is a statistically significant result. The researcher might then go on to explore the data in order to document any additional, surprising, or unexpected patterns. Approaches for explicitly correcting for the number of tests conducted, e.g., Bonferroni corrections, or for controlling the false discovery rate are also available (e.g., Benjamini and Hochberg 1995). **Solution:** Be honest about the number of tests performed and incorporate that information when you draw ecological conclusions. Be explicit about when you are testing an a priori hypothesis versus when you are exploring the data.

13. Blind use of a new fancy tool.—New statistical methods are developed, usually tested, and published every day. Often new tools become fashionable and researchers attempt to apply the new tool without deep understanding of how it works, what its limitations are, or even what question, specifically, the tool is answering. Sometimes, within a discipline, use of a new tool or statistical approach becomes seemingly necessary to get published. Take, for example, information-theoretic approaches (Burnham and Anderson 2002). These methods are incredibly valuable in particular contexts and applications, but not all. For many years, it was difficult to publish a paper in wildlife sciences that did not use these methods, leading researchers to apply them even when they were not the best choice. It took an article from the editor of the Journal of Wildlife Sciences to stem the mis-understanding. The article pointed out that "one size does not fit all" and re-focused the discussion on stating objectives, reporting a priori hypotheses, careful distinction between exploratory and confirmatory analyses, and clearly understanding the statistical framework one chooses to employ (Thompson 2010).

More recently, there has been a proliferation of articles and methods papers on applications of data-mining and machine-learning approaches in Ecology (Breiman 2001, He and Garcia 2009). Yet, these approaches are often incorrectly applied as statistical tools (He and Garcia 2009). Machine-learning represents a different paradigm than traditional hypothesis testing or model building and comes with new assumptions, limitations, and possibilities. Identification of best practices for their application may differ from those in traditional statistical analysis. For example, Barbet-Massin et al. (2012) conclude that best practices for generating pseudo-absences differ for regression techniques versus classification and machine-learning techniques. **Solution:** Choose a statistical tool based on the research question at hand and the design under which the data were collected rather than statistical-fashion. Understand that tool, its paradigm, limitations, potential biases, and assumptions.

Pitfalls in the Interpretation of Statistical Tests and Models

14. Misinterpretation of a non-significant p value.—Statisticians can enjoy hardy debate about how best to employ *p* values and critical values (alpha levels) and about what statistical framework is best for hypothesis testing. However, there is generally agreement about what constitutes misinterpretation of a *p* value. A *p* value, originally introduced by Fisher (1926), was intended as a "rough numerical guide of the strength of evidence" (Goodman 2008:135). The *p* value describes the proportion of experiments expected to get data like this or more extreme than this under the null hypothesis (usually) of no difference. 'Data like this or more extreme than this' are most often summarized in a test statistic. The 'no difference' part of the definition could reflect no difference between treatments, between types of observations, between treatments and controls, or between parameter estimates and a reference value, usually zero.

There are a host of ways that the *p* value can be misinterpreted (Goodman 2008) and misinterpretations persist despite decades of articles trying to correct them. The most dangerous of these misconceptions is that a *p* value can be used to make a determination of no difference in a traditional hypothesis testing context. A high *p* value can result when the null hypothesis is quite false but data are simply too sparse or too variable to reject it. Statistical tests are generally set up by assuming that the null hypothesis is true and collecting observations to disprove it. If adequate data are not collected to reject the null hypothesis then, without careful a priori power calculations, little can be said about the likelihood of the null hypothesis actually being true. For example, if the analysis of an experiment designed to compare the effects of riparian forest management techniques on mollusk communities is summarized with a *p* value of 0.211, this does not indicate that there is no difference between the effects of alternative riparian management treatments on mollusk communities. Absence of enough evidence to disprove the null hypothesis is not evidence of the null hypothesis (Altman and Bland 1995). Particularly with high variability and/or small sample sizes, you can only conclude that there was insufficient power to detect a difference; you cannot proclaim that the effect doesn't exist. **Solution:** Do not use *p* values to claim "no difference" or that treatments are "the same" unless you have very specifically set up your analyses to test for similarity.

15. Inappropriate comparisons of p values.—A related temptation is to compare *p* values across sample sizes or populations. Researchers, understandably, want to reduce statistical analyses to a single number such as a *p* value or R^2 in order to ease comparison of results and to allow for concise interpretations. But, in so doing, they may ignore important elements of the context of the analysis. A common pitfall is to observe a *p* value of 0.043 for treatment 1 versus treatment 2 ($n = 186$) and a *p* value of 0.36 for treatment 2 versus treatment 3 ($n = 12$) and then to conclude that treatments 1 and 2 are "more different" than treatments 2 and 3. Comparisons across populations that are particularly homogenous versus particularly variable have similar limitations.

Take, for example, data on wildfire risk factors such as drought and the prevalence of insect damage. In a particular region, one might observe that drought significantly increases wildfire risk ($p = 0.01$). In a separate t-test, the effect of the prevalence of insect damage might be non-significant ($p = 0.3$). The superficial interpretation of these results is that drought has an effect on wildfire risk but that the prevalence of insect damage does not. Looking more thoughtfully at the data, we find that this conclusion is incorrect and could lead to poor decision-making and risk assessments. Imagine that, on average, plots with higher drought ratings ($n = 1157$) were much more likely to have a wildfire with an increase in the odds of wildfire of 0.04 for every unit increase in drought rating. While, on average, plots with higher prevalence of insect damage (n = 204) were much more likely to have a wildfire with an increase of 0.12 in the odds of wildfire for every unit increase in insect damage prevalence. What is the correct interpretation? We can say with confidence that drought tends to increase wildfire risk and that these results are unlikely due to chance alone. We also see that the estimated effect size (Grissom and Kim 2005) for the prevalence of insect damage is much larger than that of drought but because few stands had insect damage it was not possible to conclude that these results are unlikely due to chance alone. **Solution:** Compare effect sizes and report your certainty about whether the observed effects are due to chance alone. Do not make conclusions by comparing *p* values where sample sizes or variability differ across elements being compared.

16. Implying ecological significance from statistical significance where there are very large sample sizes.—In some data sets, sample size is so large that even miniscule effects yield very small *p* values and what could be declared to be highly significant results. The word 'significance' is frequently misinterpreted to mean 'important' or 'meaningful'. The problem has become so intractable that some scientists are calling for a ban on the use of the word 'significant' (Higgs 2013). With very large sample sizes, you essentially have a much bigger magnifying glass for exploring the data and you can therefore detect a much smaller signal. What if 100,000 subjects were studied with equal numbers of people randomly selected from the East Coast and from the West Coast, and a significant difference of 0.001 grams is found in the birth weights of the two coasts? A statistically correct but biologically misleading headline could read: "Birth outcomes differ significantly by coast!" Although the effect is statistically significant, is the effect size (Grissom and Kim 2005) biologically relevant? Clearly the context of statistical results is as important as the results themselves, and this context should be included in any interpretation of statistical analyses. **Solution:** Particularly with large sample sizes, effects sizes should be considered. Meaningful results need to be both statistically significant and ecologically important.

17. Misinterpretation of coefficients in multiple regression models.—Coefficients in a multiple regression model cannot be interpreted in isolation and without considering the observed range of values of the predictor variables. Imagine that through some model selection procedure the following final multiple regression model is chosen:

$$Y = \beta_0 + \beta_1 \times \text{Canopy cover} + \beta_2 \times \text{Elevation}$$

where Y is any response variable. Can this model be used to state whether canopy cover or elevation has a larger effect on the response? An extremely common interpretation if β_1 is larger than β_2, is that β_1 has a greater effect than β_2 on the response. This is incorrect and sloppy.

Correct interpretations are not mathematically challenging. The correct interpretation of the intercept (β_0) is that it is the expected value of the response when both canopy cover and elevation are zero. In the absence of interactions, each coefficient describes the expected change in in the response for a 1-unit change in the predictor, keeping all other predictor variables constant. If $\beta_1 = 2$, we can conclude that for every 1% increase in canopy cover (assuming no changes in elevation), we expect that the response increases by 2. A coefficient of 0.5 for elevation indicates that for every meter change in elevation (assuming no change in canopy cover), the response variable is expected to increase by 0.5. The correct interpretation is that it would take a change in elevation of 4 meters ($4 \times 0.5 = 2$) to equal the effect of a 1% change in canopy cover, assuming all other variables are kept constant. The math is not the trickiest part. The challenge of determining which variable has a greater effect on the response is in the definition of 'greater'. Note that the ecological and mathematical interpretation of model coefficients should also consider their standard errors, the potential presence of interactions, and any underlying correlations between predictors. **Solution:** Interpret the coefficients of a linear model slowly, carefully, and in the context of the explanatory and response variables.

18. Extrapolation.—Extrapolation is tempting. Scientists, managers, and even statisticians long to make inference about one set of conditions from data collected under a different set of conditions. Predictions about new places, times, situations, and species for which we have no data will always be needed. However, these extrapolations need to be presented honestly along with the host of necessary caveats (Chatfield 2006). The most common form of extrapolation is to make predictions for new conditions which are outside of the range of the data used to build the model, including the correlation structure of the data used to build the model. The usefulness of these predictions depends greatly on whether the model form was correctly specified and whether that same model form holds for the new conditions.

The traditional example is temperature. Increases in water temperature increase fish growth rates in many experiments. This relationship is often assumed to be linear and, within the range of many experiments, it is linear. However, if one were to extrapolate to, say, 100°C, a prediction of extremely fast growth rates based on the linear model would be comical and potentially tragic. The relationship is not linear between the observed and new conditions and therefore cannot be extrapolated. In many landscape models, several moderately correlated variables are applied to make a prediction. For example, one might predict stream habitat from a linear combination of the percent of agriculture and road density in the watershed. Such a model might work well in Area A where agriculture tends to be poorly correlated with roads but strongly correlated with percent alluvium. In nearby Area B (or even in Area A measured at a finer scale), differing underlying correlations between agriculture, roads, and alluvium might lead to extremely inaccurate predictions (Lucero et al. 2011). **Solution:** Be honest when predictions are extrapolated to new places, times, conditions, or correlation structures and provide a thoughtful list of assumptions.

Conclusions

There has been a great deal of discussion about the need for statistical thinking (e.g., Snee 1990, Wild and Pfannkuch 1999) which was defined by Snee (1990:118) as "thought processes, which recognize that variation is all around us..." A 1993 American Statistical Association working group defined statistical thinking as "(a) the appreciation of uncertainty and data variability and their impact on decision-making; (b) the use of the scientific method in approaching issues and problems" (Mallows 1998:3). Some basics of statistical intuition were also outlined in an article entitled "Ecology 101" by Magnusson (1997). He identified 11 concepts that students should understand at the end of a statistics course and before collecting data. Sixteen years later, we frequently observe that recent graduates who have taken several statistics courses are not familiar with these basic ideas. That errors in the basics of statistical thought and analysis remain embedded in ecological training suggests a need to reconsider how we incorporate high quality statistics in ecological education and research.

Statistical inference is an integral component of ecological research. We encourage all ecologists to both shore up their own statistical training and to work collaboratively with someone trained in applied statistics from the earliest stages of research planning. Statistical training can be gained most easily in graduate school. Encouraging students in all branches of ecology to add an extra statistics class (or two or three) to their curriculum plan will surely benefit the discipline in the long run. Statistics is evolving rapidly however, perhaps more rapidly than most other scientific disciplines because of the massive increases in both computing power and data availability. Reading statistics papers (and we would encourage statisticians to remember the applied and even non-statistical specialist audience), attending lectures, taking on-line courses, and focused self-study are all options for expanding one's statistical know-how long after graduate school. Statistical intuition and statistical thinking often develop with experience and collaborations beyond classroom learning. Collaboration with trained and experienced applied statisticians is an excellent way to improve a study plan or experimental design, "significantly" reduce the occurrence of pitfalls as described here, and, perhaps, gain access to tools and approaches that you might not have otherwise considered.

Often the best road is to slow down and think. As one anonymous Associate Editor from *Northwest Science* wrote in response to a manuscript that came across one of the co-authors' consulting desk, "I would love to see this in print, but

Table 1. Simple solutions for avoiding common statistical pitfalls.

Statistical pitfall	Simple solution
Common statistical pitfalls in setting up an analysis	
Failure to explore the data	Plot the data early and often.
Arbitrary thresholds, metrics, and indicators	Be explicit in the choice and calculation of thresholds, metrics, and indicators.
Assuming that observations are independent	Identify the unit of inference and use it to determine the true sample size.
Mismatched sampling frame and population	Ensure that the sampling frame includes every individual in the population.
Common pitfalls in experimental design	
Control sites (or reference sites) differ from treatment sites before the treatment occurs	Assign treatment and control sites randomly; test for systematic differences between units assigned as controls and treatments.
Measurement strategies that confound experimental designs	Ensure that measurement strategies account for time- or space-sensitive measurements.
Failure to model covariates at the correct level	When interactions are present between treatment levels and a covariate, look for treatment effects at meaningful values of that covariate.
Pitfalls in the application of statistics	
Unnecessary data transformations	Transform the response as a last resort.
Ignoring underlying correlation structure	Where measurements are not likely to be independent, explicitly model the correlation structure in the data.
Failure to plot the residuals (or other model diagnostics)	Plot the residuals to identify patterns in the data that are not explained by the model.
Conducting too many tests	Be honest about the number of tests performed and explicit about whether you are testing an a priori hypothesis versus exploring the data.
Not dealing appropriately with zeros	Do not ignore, delete, or average away large numbers of zero observations.
Blind use of a new fancy tool	Choose a statistical tool based on the research question at hand not statistical fashion.
Pitfalls in the interpretation of statistical tests and models	
Misinterpretation of a non-significant *p* value	Do not use *p* values to claim "no difference" or that treatments are "the same" unless you have very specifically set up your analyses to test for similarity.
Inappropriate comparisons of *p* values	Compare effect sizes and report your certainty about whether the observed effects are due to chance alone.
Implying ecological significance from statistical significance where there are very large sample sizes.	Particularly with large sample sizes, effects sizes should be considered. Meaningful results need to be both statistically significant and ecologically important.
Misinterpretation of coefficients in multiple regression models	Interpret the coefficients of a linear model slowly, carefully, and in the context of the explanatory and response variables.
Extrapolation	Be honest when predictions are extrapolated to new places, times, conditions, or correlation structures.

you really need to take it back to the drawing board, drop all the hypothesis tests, and do some good old fashioned thinking about what the data are telling you, practically and biologically." Impressive calculations and complex modeling should not come at the cost of deeply understanding how the math and the ecology interact to create new knowledge.

Clearly, avoiding pitfalls often takes time. Clean, careful quantitative analysis usually takes more time than sloppy statistics. Clean, careful quantitative analyses, however, are an indispensable tool for making inference from observations in a world full of natural variability. Statistical intuition, thoughtful application of even simple statistics, deep understanding of the statistical paradigm being applied, and honesty can go a long way in preventing the misuse of statistics (Table 1). "Why, given the extensive resources and time it takes to collect the data, do some people expect to be able to do the analysis in an afternoon? Why would they want to?" (Clayton 2007:303).

Acknowledgments

We would like to thank the students, ecologists, and statisticians who have presented interesting problems to us over the years; colleagues who have listened to our experiences, shared similar experiences, and provided encouragement to write this manuscript; as well as two anonymous reviewers whose comments have dramatically improved the manuscript.

Literature Cited

Altman, D. G., and J. M. Bland. 1995. Absence of evidence is not evidence of absence. British Medical Journal 311:485.

Anscombe, F. J. 1973. Graphs in statistical analysis. American Statistician 27:17–21.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecology and Evolution 3:327–338.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57:289–300.

Bensinger, K. 2012. Number crunchers defeat pundits in election predictions. Seattle Times November 8 .

Breiman, L. 2001. Statistical modeling: The two cultures. Statistical Science 16:199–231.

Burnham, K. P. and D. R. Anderson. 2002. Model selection and multi-model inference: a practical information-theoretic approach. Springer-Verlag, New York, New York, USA.

Chatfield, C. 2006. Model uncertainty: Encyclopedia of environmetrics. John Wiley and Sons, Hoboken, New Jersey, USA.

Clayton, M. K. 2007. How should we achieve high-quality reporting of statistics in scientific journals? A commentary on "Guidelines for reporting statistics in journals published by the American Physiological Society". Advances in Physiology Education 31:302–304.

Fowler, N. 1990. The 10 most common statistical errors. Bulletin of the Ecological Society of America 71:161–164.

Fernandes-Taylor, S., J. K. Hyun, R. N. Reeder, and A. H. S. Harris. 2011. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. BMC Research Notes 4:304.

Ford, E. D. 2000. Scientific method for ecological research. Cambridge University Press, Cambridge, UK.

Freedman, D. A. 1983. A note on screening regression equations. American Statistician 37:152–155.

García-Berthou, E., and C. Alcaraz. 2004. Incongruence between test statistics and P values in medical papers. BMC Medical Research Methodology 4:13.

Glantz, S. A. 1980. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation 61:1–7.

Good, P. I., and J. W. Hardin. 2003. Common errors in statistics (and how to avoid them). John Wiley and Sons, Hoboken, New Jersey, USA.

Goodman, S. 2008. A dirty dozen: Twelve p-value misconceptions. Seminars in Hematology 45:135–140.

Grissom, R. J., and J. J. Kim. 2005. Effect sizes for research: univariate and multivariate applications. Routledge, New York, New York, USA.

Gurevitch, J., and S. T. Chester. 1986. Analysis of repeated measures experiments. Ecology 67:251–255.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21:1263–1284.

Higgs, M. D. 2013. Do we really need the s-word? American Scientist 101:6–9.

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

Ionnidis, J. P. A. 2005. Why most published research findings are false. PLoS Med 2(8).

Liang, K., and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. Biometrika 73:13–22.

Littell, R. C., G. A. Milliken, W. W. Stroup, and R. D. Wolfinger. 1996. SAS system for mixed models. SAS Institute, Cary, North Carolina, USA.

Lucero, Y., E. A. Steel, K. M. Burnett, and K. Christiansen. 2011. Untangling human development and natural gradients: implications of underlying correlation structure for linking landscapes and riverine ecosystems. 19:207–224.

Ma, Y., and P. Guttorp. 2013. Estimating daily mean temperature from synoptic climate observations. International Journal of Climatology 33:1264–1269.

Magnusson, W. E. 1997. Teaching experimental design in ecology, or how to do statistics without a bikini. Bulletin of the Ecological Society of America 78:205–209.

Mallows, C. 1998. 1997 Fisher Memorial Lecture: The zeroth problem. American Statistician 52:1–9.

Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters 8:1235–1246.

O'Hara, R. B., and D. J. Kotze. 2010. Do not log-transform count data. Methods in Ecology and Evolution 1:118–122.

Peterson, E. E. et al. 2013. Modelling dendritic ecological networks in space: an integrated network perspective. Ecology Letters 16:707–719.

Potts, J. M., and J. Elith. 2006. Comparing species abundance models. Ecological Modeling 199:153–163.

Roni, P., M. C. Liermann, C. Jordan, and E. A. Steel. 2005. Steps for designing a monitoring and evaluation program for aquatic restoration. Pages 13–34 *in* P. Roni, editor. Monitoring stream and watershed restoration. American Fisheries Society, Bethesda, Maryland, USA.

SAS Institute Inc. 2011. SAS/STAT 9.3 user's guide. SAS Institute, Cary, North Carolina, USA.

Snee, R. D. 1990. Statistical thinking and its contribution to total quality. American Statistician 44:116–121.

Squire, P. 1988. Why the 1936 Literary Digest poll failed. Public Opinion Quarterly 52:125–133.

Strasak, A. M., Q. Zaman, G. Marinell, K. P. Pfeiffer, and H. Ulmer. 2007. The use of statistics in medical research. American Statistician 61:47–55.

Thompson, III, F. R. 2010. Editor's Message: Application and presentation of statistics. Journal of Wildlife Management 74:617–619.

Warton, D., and F. K. C. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3–10.

Wild, C. J., and M. Pfannkuch. 1999. Statistical thinking in empirical inquiry. International Statistical Review 67:223–265.

Zuur, A. F., E. N. leno, and C. S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1:3–14.