Overcoming "analysis paralysis"

Benjamin Zuckerberg
Citizen Science Program, Cornell Laboratory of Ornithology, Cornell University, Ithaca, NY (bz73@cornell.edu)

The Chocó Indians of Colombia routinely treat the tips of their blowgun darts with secretions collected from the skin of poison dart frogs. When a dart meets its victim, the result is a prolonged state of paralysis. Although *slightly* less painful, the prospect of data analysis can have a similar effect on the unsuspecting graduate student. "Analysis paralysis" refers to an overload of analysis options that impedes a student's research progress. It usually strikes when a graduate student has reached the final stages of his/her project. I have witnessed student colleagues experience such severe paralysis that they become overwhelmed – not by mosquitoes or rainy days, but by computer spreadsheets – and ultimately abandon their studies altogether.

The ecologist's toolbox has never been so full. Moreover, ecology is an increasingly quantitative undertaking (Kroll 2007). Scanning any recent issue of an ecological journal will reveal terms such as "random-effects models", "phylogeography", "autoregression", "information-theoretic approaches", "Bayesian analyses", "resource-selection functions", "Markov chains", and "variance decomposition". A graduate student's first encounter with these analyses often coincides with an immediate need for implementation. Because learning curves are steep, however, first encounters can lead to bruises and confusion for students who don't know where to begin (or end) their analyses. This is the precise point where "analysis paralysis" sets in. The following represents several ways to avoid this:

Befriend a statistical "guru". Befriending a statistician early in your graduate career is critical. You need to be proactive and prepared. As Sterns (1987) suggests, "If you want to pick somebody's brains, you'll have to go to him or her, because they won't be coming to you". To this I would add, "If you don't know what you are asking, they can't provide any answers". Furthermore, asking a statistician to help you after you have already collected the data is a postmortem exercise. The secret to good data analysis is having good data to analyze in the first place (Johnson *et al.* 2001).

Your "guru" may not be a statistician per se, but could be a professor, a postdoc, or even a fellow graduate student. In fact, a professor or postdoc may be more familiar with the analytical approach most appropriate to your topic than a bona fide statistician. Whatever the case, come equipped with specific questions and don't be afraid to knock on that door. These "gurus" are usually very busy (hence their "guru" status), so your relationship with them will be welcomed all the more if it leads directly to greater independence on your part. This is also a good opportunity to discuss the advantages and disadvantages of various statistical software platforms (eg R, SAS, SPSS, etc).

What's the question?! Identifying the key question is really the main problem behind "analysis paralysis" – if you don't know where you are going, you can't really get there. We are easily seduced by technology and forget about old-fashioned common sense and clear thinking (Kroll 2007). Statistical "gurus" frequently encounter graduate students looking to perform a specific analysis with little justification for *why* they want to perform that particular one. Developing the overriding question(s) begins a process of identifying the population of interest, delineating appropriate sampling units, defining relevant variables, and determining how those variables will be measured (Johnson *et al.* 2001). You may need to spend weeks or even months generating a set of a priori hypotheses (Anderson 2008). The appropriate analytical framework will readily follow. No type or amount of analysis can compensate for poorly defined questions and a lousy study design.

Complexity through simplicity. Sometimes a simple analysis is just as useful as a complex one. In many cases, the results obtained from "simple" approaches are easier to communicate. Complex analyses can give the impression of "perfection" and "sophistication", but the most direct approaches to clearly posed questions are usually best. To determine how to proceed, search the literature on your topic and make a list of commonly used statistical approaches (Johnson *et al.* 2001). Keep in mind that no single approach is universally accepted or preferable. Most

importantly, make up a dummy dataset before you begin your research and work through the analysis before collecting any "real" data. This will save you countless headaches later.

Misery loves company, but so does creativity. A few years ago, a graduate student was told by a professor that random-effects models represented the new frontier in resource selection modeling. Needing guidance on this new approach, several graduate students (with limited formal training in statistics) came together, identified a test dataset, performed analyses, and eventually published a paper on the method (Gillies *et al.* 2006). Graduate students are often frustrated by limited options in statistics courses, inconsistencies in past studies, and ongoing debates among scientists about various statistical approaches (Butcher *et al.* 2007). A group of graduate students can create a safe haven to explore these issues and resolve the ambiguities. Don't overlook the value of inviting the leader in the field with respect to a particular procedure to come and lead a workshop.

For those graduate students dreading analysis, the thing you least want to do is that which you should do first: think critically about what you are trying to do in the research process ahead. Procrastination and fuzzy hypotheses are the key ingredients of analysis paralysis. Although technology and statistics have widened our circle of investigation, advances in ecology often start with good questions. Graduate students should stay abreast of developments in the field, talk to each other, identify good statistical "gurus", and, most of all, become well-versed skeptics.

Faculty response

James P Gibbs
Professor, College of Environmental Science and Forestry,
State University of New York,
Syracuse, NY (jpgibbs@esf.edu)

Europeans have long enjoyed the edible frog, *Rana esculenta*, whose name is derived from the Latin term for "delicious frog". When prepared with butter, garlic, and parsley sauce, consumption can create a prolonged state of pleasure. Such is the case with statistical analyses in ecology – you can choose to be poisoned or pleasured, depending on which frog you select.

To Ben's excellent insights I would add just one other: that students need to know the basics and prepare well. I too often see graduate students trying, with great uncertainty, to defend a thesis that presents complex statistics summarizing vast amounts of data. When asked to explain calculation of a sample mean and its variance, they are often unable to respond. Do you know the formula for the standard deviation of a population sample? It is unwise to skip such essentials; one way or another, they underpin even the most sophisticated statistical analyses.

For example, the simple contingency table and humble confidence interval remain among the most powerful tools in the proverbial ecologist's toolbox. Develop a "gut" sense for how these and other basic statistical tests work. As Ben mentions in a related comment, the best way to do this is through simulating data to examine interactions among effect sizes, significance levels, power, replication, and so forth. These days, you should also fiddle around with concepts of maximum likelihood and model fitting criteria. You don't need to be a programmer – even MS Excel will suffice as a vehicle. A good starting point is the Vermont Cooperative Fish and Wildlife Research Unit's Spreadsheet Project (www.uvm.edu/envnr/vtcfwru/spreadsheets/ – I happen to like the hypothesis-testing exercise for getting at the essentials).

By "messing around" with lots of data, you start to develop an intuitive sense of the essentials of making inferences under uncertainty – the essence of statistical analysis. You then start to take ownership of the enterprise, rather than perceiving statistical analyses merely as a series of obstacles others have set before you, just to trip you up. Who knows? One day, you may find that you actually enjoy statistics. Those "gurus" Ben mentions became that way for a reason – stats can be both fun and interesting, compelling one slowly toward "guru" status. Whatever the case, mastering the basics is key to ensuring that tackling sophisticated analyses will induce more of a sense of pleasure than of paralysis.

References

Anderson DR. 2008. Model based inference in the life sciences: a primer on evidence. New York, NY: Springer.

Butcher JA, Groce JE, Litunia CM, *et al.* 2007. Persistent controversy in statistical approaches in wildlife sciences: a perspective of students. *J Wildlife Manage* **71**: 2142–44.

Gillies CS, Hebblewhite M, Nielsen SE, *et al.* 2006. Application of random effects to the study of resource selection by animals. *J Anim Ecol* **75**: 887–98.

Johnson DH, Shaffer TL, and Newton WE. 2001. Statistics for wildlifers: how much and what kind? *Wildlife Soc B* **29**: 1055–60.

Kroll AJ. 2007. Integrating professional skills in wildlife student education. *J Wildlife Manage* **71**: 226–30.

Sterns SC. 1987. Some modest advice for graduate students. *Bull Ecol Soc Am* **68**: 145–53.