Paths to Statistical Fluency for Ecologists

PATHS TO STATISTICAL FLUENCY FOR ECOLOGISTS

Aaron M. Ellison[1,3] & Brian Dennis[2]

[1]Harvard University, Harvard Forest, 324 North Main Street, Petersham, MA 01366

[2]Department of Fish & Wildlife Resources, and Department of Statistics, University of Idaho, Moscow, ID 83844

[3]For all correspondence: aellison@fas.harvard.edu

ABSTRACT

Twenty-first century ecology requires statistical fluency. Observational and experimental studies routinely gather non-Normal, multivariate data at many spatiotemporal scales. Experimental studies routinely include multiple blocked and nested factors. Ecological theories routinely incorporate both deterministic and stochastic processes. Ecological debates frequently revolve around choices of statistical analyses. Our journals are replete with likelihood and state-space models, Bayesian and frequentist inference, complex multivariate analyses, and papers on statistical theory and methods. We test hypotheses, model data, and forecast future environmental conditions. And many *appropriate* statistical methods are not automated in software packages. It is time for ecologists to understand statistical modeling well enough to construct nonstandard statistical models and apply various types of inference – estimation, hypothesis testing, model selection, and prediction – to our models and scientific questions. In short, ecologists need to move beyond basic statistical literacy and attain statistical fluency.

In a nutshell:

- Ecologists need to use nonstandard statistical models and methods of statistical inference to test models of ecological processes and to address pressing environmental problems.
- Such statistical models of ecological processes include both deterministic and stochastic parts, and statistically-fluent ecologists will need to use probability theory and calculus to fit these models to available data.
- Many ecologists lack appropriate background in probability theory and calculus because there are serious disconnections between the quantitative nature of ecology, the quantitative skills we expect of ourselves and our students, and how we teach and learn quantitative methods.
- A prescription for attaining statistical fluency includes: two semesters of standard calculus; a calculus-based introductory statistics course; a two-course sequence in probability and mathematical statistics; and most importantly, a commitment to using calculus and post-calculus statistics in courses in ecological and environmental-science curricula.

INTRODUCTION

For the better part of a century, ecology has used statistical methods developed mainly for agricultural field trials by statistics luminaries such as Gossett, Fisher, Neyman, Cochran, and Cox (Gotelli and Ellison 2004). Calculation of sums of squares was just within the reach of mechanical (or human) calculators (Fig. 1), and generations of ecologists have spent many hours in their labor of love: caring and curating the results of analysis of variance (ANOVA) models. Basic linear models (ANOVA and regression) continue to be the dominant mode of ecological data analysis; they were used in 75% of all papers published in *Ecology* in 2008 ($N = 344$; 24

papers were excluded from the analysis because they were conceptual overviews, notes, or commentaries that reported no statistics at all). These methods are employed most appropriately to analyze relatively straightforward experiments aimed at estimating the magnitudes of a small number of additive fixed effects or testing simple statistical hypotheses. Although the vast majority of papers published in *Ecology* test statistical hypotheses (75% reported at least one *P*-value) and estimate effect sizes (69%), only 32% provided assessments of uncertainty (*e.g.,* standard errors, confidence intervals, probability distributions) on the estimates of the effect sizes themselves (as distinguished from the common practice of reporting standard errors of observed means).

But these methods do not reflect ecologists' collective statistical needs for the 21st century. How can we use ANOVA and simple linear regression to forecast ecological processes in a rapidly changing world (Clark *et al.* 2001)? Familiar examples or ecological problems that would benefit from sophisticated modeling approaches include: forecasts of crop production; population viability analyses; prediction of the spread of epidemics or invasive species; and predictions of fractionation of isotopes through food webs and ecosystems. Such forecasts, and many others like them, are integral to policy instruments such as the Millennium Ecosystem Assessment (2005) or the IPCC reports (IPCC 2007). Yet such forecasts and similar types of studies are uncommon in top-tier ecological journals. Why? Do ecologists limit their study designs so as to produce data that will fit into classical methods of analysis? Are nonstandard ecological data sometimes mis-analyzed with off-the-shelf statistical techniques (Bolker *et al.* 2009)? In the statistical shoe store, do ecologists sometimes cut the foot to fit the shoe? How do we learn to do more than determine *P*-values associated with mean squared error terms in analysis of variance (Butcher *et al.* 2007)?

The short answer is by studying and using “models”. Statistical analysis is fundamentally a process of building and evaluating stochastic models, but such models were hidden or even forbidden in the agricultural statistics-education tradition that emphasized practical training and de-emphasized calculus. Yet, any ecological process producing variable data can (and should) be described using a stochastic, statistical model (Bolker 2008). Such models may start as a conceptual or “box-and-arrow” diagram, but these should then be turned into more quantitative descriptions of the processes of interest. The building blocks of such quantitative descriptions are deterministic formulations of the hypothesized effects of environmental variables, time, and space, coupled with discrete and continuous probability distributions. These distributions, rarely Normal, are chosen by the investigator to describe how the departures of data from the deterministic sub-model are hypothesized to occur. The Sums of Squares – a surrogate for likelihood in Normal distribution models – is no longer the only statistical currency; likelihood and other such statistical objective functions are the more widely useful coins of the realm.

Alternatives to parametric model-based methods include non-parametric statistics and machine-learning. Classical non-parametric statistics (Conover 1998) have been supplanted by computer simulation and randomization tests (Manly 2006) but the statistical or causal models that they test are rarely apparent to data analysts and users of packaged (especially compiled) software products. Similarly, model-free machine-learning and data-mining methods (Breiman 2001) seek large-scale correlative patterns in data by letting the data “speak for themselves”. Although the adherents of these methods promise that machine-learning and data-mining will make the “standard” approach to scientific understanding – hypothesis → model → test – obsolete (Anderson 2008), the ability of these essentially correlative methods to advance scientific understanding and provide reliable forecasts of future events has yet to be demonstrated. Thus we focus here on the complexities inherent in fitting stochastic statistical models, estimating their parameters, and carrying out statistical inference on the results.

Our students and colleagues create or work far less frequently with stochastic statistical models than they use routine ANOVA and its relatives; in 2008, only 23% of papers published in *Ecology* used stochastic models or applied competing statistical models on their data (and about half of these used automated software such as stepwise regression or MARK [White and Burnham 1999] that take much of the testing out of the hands of the user to contrast among models constructed from many possible combinations of parameters). Why? It may be that we (or at least those of us who publish in our leading journals) primarily conduct well designed experiments that test one or two factors at a time and have sufficient sample sizes and balance among treatments to satisfy all the requirements of ANOVA and yield high statistical power. If this is true, the complexity of stochastic models is simply unnecessary. But our data rarely are so forgiving; more frequently our sample sizes are too small, our data are not Normally distributed (or even continuous), our experimental and observational designs include mixtures of fixed and random effects, and we know that process affect our study systems hierarchically. And finally, we want to do more with our data than simply tell a good story. We want to generalize, predict, and forecast. In short, we really *do* need to model our data.

We suggest that there are profound disconnections between the quantitative nature of ecology, the quantitative (mathematical and statistical) skills we expect of ourselves and of our students, and how we teach and learn quantitative methods. We illustrate these disconnections with two motivating examples and suggest a new standard – *statistical fluency* – for quantitative skills that are learned and taught by ecologists. We close by providing a prescription for better connecting (or reconnecting) our teaching with the quantitative expectations we have for our students so that ecological science can progress more rapidly and with more relevance to society at large.

Two Motivating Examples

The first law of population dynamics

Under optimal conditions, populations grow exponentially:

$$N_t = N_0 \, e^{rt} \qquad \text{(Eqn. 1)}$$

In this equation, N_0 is the initial population size, N_t is the population size at time t, r is the instantaneous rate of population growth (units of individuals per infinitesimally small units of time t), and e is the base of the natural logarithm. This simple equation is often referred to as the first law of population dynamics (Turchin 2001) and it is universally presented in undergraduate ecology textbooks. Yet we all know all too well that students in our introductory ecology classes view exponential growth mainly through glazed eyes. Why? Equation 1 is replete with complex mathematical concepts normally encountered in the first semester of calculus: the concept of a function, raising a real number to a real power, and Euler's number e. But the majority of undergraduate ecology courses *do not require* calculus as a prerequisite, thereby insuring that understanding fundamental concepts such as exponential growth is not an expected course outcome. The current financial meltdown associated with the foreclosure of exponentially ballooning sub-prime mortgages illustrates writ large Albert Bartlett's assertion that "the greatest shortcoming of the human race is our inability to understand the exponential function". Surely ecologists can do better.

Instructors of undergraduate ecology courses that do require calculus as a prerequisite often find themselves apologizing to their students that ecology is a quantitative science and go on to provide conceptual or qualitative workarounds that keep course enrollments high and deans happy. Students in the resource management fields – forestry, fisheries, wildlife, *etc*. – suffer even more, as quantitative skills are further de-emphasized in these fields. Yet resource managers need a deeper understanding of exponential growth (and other quantitative concepts) than do academic ecologists; for example, the relationship of exponential growth to economics or its role in the concept of the present value of future revenue. The result in all these cases is the perpetuation of a *culture of quantitative insecurity* among many students.

The actual educational situation with our example of population growth models in ecology is much worse. The exponential growth expression as understood in mathematics is the solution to a differential equation. Differential equations, of course, are a core topic of calculus. Indeed, because so many dynamic phenomena in all scientific disciplines are naturally modeled in terms of instantaneous forces (rates), the topic of differential equations is one of the main reasons for studying calculus in the first place! To avoid introducing differential equations to introductory ecology classes, most ecology textbooks present exponential growth in a discrete-time form: $N_{t+1} = (1 + \text{births} - \text{deaths})\, N_t$ and then miraculously transmogrify this (with little or no explanation) into the continuous time model given by $dN/dt = rN$. The attempts at intuition obscure, for instance, the exact nature of the quantities "births" and "deaths" and how they would be measured, not to mention the assumptions involved in discrete time versus continuous time formulations.

Furthermore, Eqn. 1 provides no insights into how the unknown parameters (r and even N_0 when population size is not known without error) ought to be estimated from ecological data. To convince yourself that it is indeed difficult to estimate unknown parameters from ecological data, consider the following as a first exercise for an undergraduate ecology laboratory: for a given set of demographic data (perhaps collected from headstones in a nearby cemetery), estimate r and N_0 in Eqn. 1 and provide a measure of confidence in the estimates.

Finally, to actually use Eqn. 1 to describe the exponential growth of a real population, one must add stochasticity by modeling departures of observed data from the model itself. There are many different ways of modeling such variability that depend on the specific stochastic forces acting on the observations; each model gives a different likelihood function for the data and thereby prescribes a different way for estimating the growth parameter. In addition, the choices of models for the stochastic components, such as demographic variability, environmental variability, and sampling variability, must be added to (and evaluated along with) the suite of modeling decisions concerning the deterministic core, such as changing exponential growth to some density dependent form or adding a predator. Next, extend these concepts and methods to "simple" Lotka-Volterra models of competition and predation...

The Cumulative Distribution Function for a Normal curve

Our second motivating example deals with a core concept of statistics:

$$\int_a^b \left(\sigma^2 2\pi\right)^{-1/2} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] dy = \Phi(b) - \Phi(a) \qquad \text{(Eqn. 2)}$$

The function $\Phi(y)$ is the cumulative distribution function for the Normal distribution and Eqn. 2 describes the area under a Normal curve (with two parameters: mean = μ and variance = σ^2) between a and b. This quantity is important because the Normal distribution is used as a model assumption for many statistical methods (*e.g.,* linear models, probit analysis), and Normal probabilities can express predicted frequencies of occurrence of observed events (data). Also, many test statistics also have sampling distributions that are approximately Normal. Rejection regions, *P*-values, and confidence intervals all are defined in terms of areas under a Normal curve.

The meaning, measurement, and teaching of *P*-values continues to bedevil statisticians (*e.g.,* Berger 2003, Hubbard and Byarri 2003, Murdoch *et al.,* 2008), yet ecologists often use and interpret probability and *P*-values uncritically, and few ecologists can clearly describe a confidence interval with any degree of… uh, confidence. To convince yourself that this is a real problem, consider asking any graduate student in ecology (perhaps during their oral comprehensive examination) to explain why $P(10.2 < \mu < 29.8) = 0.95$ is not the correct interpretation of a confidence interval on the parameter μ (original equation from Poole 1974); odds are you will get an impression of someone who is not secure in their statistical understanding. Bayesians should refrain for chortling about the transparency of credible sets. Interpreting Bayesian credible intervals makes equally large conceptual demands (Hill 1968, Lele and Dennis 2009). When pushed, students can *calculate* a confidence interval by hand or with computer software. But interpreting it (Box 1) and generalizing its results is where the difficulty lies.

Three centuries of study of Eqn. 2 by mathematicians and statisticians have not reduced it to any simpler form, and evaluating it for any two real numbers a and b must be done numerically. Alternatively, one can proceed through the mysterious, multi-step table-look-up process, involving the *Z*-tables provided in the back of every basic statistics text. Look-up tables or built-in functions in statistical software may work fine for standard probability distributions such as the Normal or F distribution, but what about non-standard distributions or mixtures of distributions used in many hierarchical models? Numerical integration is a standard topic in calculus classes, and it can be applied to *any* distribution of interest, not just the area under a

Normal curve. Consider the power of understanding: how areas under curves can be calculated for other continuous models besides the Normal distribution; how the probabilities for other distributions sometimes converge to the above form based on the Normal; and how Normal-based probabilities can serve as building blocks for hierarchical models of more complex data (Clark 2007). Such interpretation and generalization is at the heart of statistical fluency.

DEVELOPING STATISTICAL FLUENCY AMONG ECOLOGISTS

Fluency defined

We use the term “fluency” to emphasize that a deep understanding of statistics and statistical concepts differs from “literacy” (Table 1). Statistical literacy is a common goal of introductory statistics courses that presuppose little or no familiarity with basic mathematical concepts introduced in calculus, but it is insufficient for 21st century ecologists. Like fluency in a foreign language, statistical fluency means not only a sufficient understanding of core theoretical concepts (grammar in languages, mathematical underpinnings in statistics) but also the ability to apply statistical principles and adapt statistical analyses for nonstandard problems (Table 1).

We must recognize that calculus is the language of the general principles that underlie probability and statistics. We emphasize that statistics is *not* mathematics; rather, like physics, statistics uses a lot of mathematics (De Veaux and Velleman 2008). And ecology uses a lot of statistics. But the conceptual ideas of statistics are *really hard*. Basic statistics contains abstract notions derived from those in basic calculus, and students who take calculus courses *and use* calculus in their statistics courses have a deeper understanding of statistical concepts and the confidence to apply them in novel situations. In contrast, students who take only calculus-free, cookbook-style statistical methods courses often have a great deal of difficulty adapting the statistics that they know to ecological problems for which those statistics are inappropriate.

For ecologists, the challenge of developing statistical fluency has moved well beyond the relatively simple task of learning and understanding fundamental aspects of contemporary data analysis. The very theories themselves in ecology include stochastic content that can only be interpreted probabilistically and include parameters that can only be estimated using complex statistics. For example, conservation biologists struggle with (and frequently mis-express) the distinctions between demographic and environmental variability in population viability models and must master the intricacies of first passage properties of stochastic growth models. Community ecologists struggle to understand (and figure out how to test) the "neutral" model of community structure (Hubbell 2001), itself related to neutral models in genetics (see Leigh 2007) with which ecological geneticists must struggle. Landscape ecologists must struggle with stochastic dispersal models and spatial processes. Behavioral ecologists must struggle with Markov chain models of behavioral states. All must struggle with huge individual-based simulations and hierarchical (random or latent effects) models. No subfield of ecology, no matter how empirical the tradition, is safe from encroaching stochasticity and the attendant need for the mathematics and statistics to deal with it.

Statistics is a post-calculus subject

What mathematics do we need – to create, parameterize, and use stochastic statistical models of ecological processes? At a minimum, we need calculus. We must recognize that statistics is a post-calculus subject and that calculus is a prerequisite for development of statistical fluency. Expectation, conditional expectation, marginal and joint distributions, independence, likelihood, convergence, bias, consistency, distribution models of counts based on infinite series… are key concepts of statistical modeling that must be understood by practicing ecologist, and these are straightforward calculus concepts. No amount of pre-calculus statistical "methods" courses can make up for this fact. Calculus-free statistical methods courses doom ecologists to a lifetime of insecurity with regard to the ideas of statistics. Such courses are like potato chips: virtually no nutritional value, no matter how many are consumed. Pre-calculus statistics courses are similar to pre-calculus physics courses in that regard; both have reputations for being notorious, unsatisfying parades of mysterious plug-in formulas. Ecologists who have taken and internalized post-calculus statistics courses are ready to grapple with the increasingly stochastic theories at the frontiers of ecology and will be able to rapidly incorporate future statistical advances in their kit of data analysis tools. How do our students achieve statistical fluency?

The prescription

Basic calculus, including an introduction to differential equations, seems to us to be a minimum requirement. Our course prescription includes (1) two semesters of standard calculus and an introductory, calculus-requiring introductory statistics course in college; and (2) a two-semester post-calculus sequence in probability and mathematical statistics in the first or second year of graduate school (Box 2). But it is not enough to simply *take* calculus courses, as calculus already is clearly required (or at least recommended) by virtually all undergraduate science degree programs (Fig. 2). Rather, calculus must be *used*; not only in statistics courses taken by graduate students in ecology but most importantly in undergraduate and graduate courses in ecology (including courses in resource management and environmental science)! If this seems overly daunting, consider that Hutchinson (1978) summarizes “the modicum of infintesimal calculus required for ecological principles” in three and a half pages. Contemporary texts (such as Clark 2007 or Bolker 2008) in ecological statistical modeling use little more than single variable calculus and basic matrix algebra. Like Hutchinson, Bolker (2008) covers the essential calculus and matrix algebra in 4 pages, each half the size of Hutchinson’s! Clark’s (2007) 100-page mathematical refresher is somewhat more expansive, but in all cases the authors illustrate that knowledge of some calculus allows one to advance rapidly on the road to statistical fluency.

We emphasize that nascent ecologists need not take *more* courses to attain statistical fluency; they just need to take courses that are *different* from standard “methods” classes. Current graduate students may need to take refresher courses in calculus and mathematical statistics, but we expect that our prescription (Box 2) will actually reduce the time that future ecology students spend in mathematics and statistics classrooms. Most undergraduate life science students already take calculus and introductory statistics (Fig. 2). The pre-calculus statistical methods courses that are currently required can be swapped out in favor of two semesters of post-calculus probability and statistics. Skills in particular statistical methods can be obtained through self-study or through additional methods courses; a strong background in probability and statistical theory makes self-study a realistic option for rapid learning for motivated students.

Why not just collaborate with professional statisticians?

In the course of speaking about statistics education to audiences of ecologists and natural resource scientists, we often are asked questions such as: “I don’t have to be a mechanic to drive a car, so why do I need to understand statistical theory to be an ecologist? (and why do I have to know calculus to do statistics?)” Our answer, the point of this article, is that the analogy of statistics as a tool or black box increasingly is failing the needs of ecology. Statistics is an essential part of the thinking, the hypotheses, and the very theories of ecology. Ecologists of the future should be prepared to confidently use statistics so that they can make substantial progress at the frontiers of our science.

"But," continues the questioner, "why can't I just enlist the help of a statistician?" Collaborations with statisticians can produce excellent results and should be encouraged wherever and whenever possible, but ecologists will find that their conversations and interactions with professional statisticians will be enhanced if ecologists have done substantial statistical ground work before their conversation begins and if both ecologists and statisticians speak a common language (mathematics!). Collaborations between ecologists and statisticians also can be facilitated by building support for consulting statisticians into grant proposals; academic statisticians rely on grant support as much as academic ecologists do. However, ecologists cannot count on the availability of statistical help whenever it is needed. And, statistical help may be unavailable at many universities. Thus, we believe that ecologists should be self-sufficient and self-assured. We should master our own scientific theories and be able to discuss with confidence how our conclusions are drawn from ecological data. We should be knowledgeable enough to recognize what we do understand and what we do not, learn new methods ourselves, and seek out experts who can help us increase our understanding.

CONCLUSION: MATHEMATICS AS THE LANGUAGE OF ECOLOGICAL NARRATIVES

It is increasingly appreciated that scientific concepts can be communicated to students of all ages through stories and narratives (Fig. 3; see also Molles 2006). We do not disagree with the importance of telling a good story and engaging our students with detailed narratives of how the world works. Nor do we minimize the importance of doing "hands-on" ecology through inquiry-based learning, which is both important and fun. Field trips, field work, and lab work are exciting and entertaining, draw students into ecology, and dramatically enhance ecological literacy. For individuals who pursue careers in fields outside of science, qualitative experiences and an intuitive grasp of the story-line can be sufficient (Cope 2006). But for our students who want the deepest appreciation and joy of how science works – understanding how we know what we know – and for those of us who are in scientific careers and are educating the next generation of scientists, we should use the richest possible language for our narratives of science. And that language is mathematics.

Acknowledgements

This paper is derived from a talk on statistical literacy for ecologists presented by the authors at the 2008 ESA Annual Meeting in the symposium "Why is ecology hard to learn?" We thank Charlene D'Avanzo for organizing the symposium and inviting our participation, the other participants and members of the audience for their thoughtful and provocative comments that helped us to refine our arguments, and Ben Bolker, an anonymous reviewer, and the Associate Editor for useful comments that improved the final version.

References

Anderson C. 2008. The Petabyte age: because more isn't just more – more is different. *Wired Magazine* Issue **16.07** (July 2008):106-120.

Armsworth PR, Gaston KJ, Hanley ND, *et al.* 2009. Contrasting approaches to statistical regression in ecology and economics. *J Appl Ecol* **46**: 265-268.

Berger JO. 2003. Could Fisher, Jeffreys and Neyman have agreed on testing (with comments and rejoinder). *Stat Sci* **18**:1-32.

Bolker B. 2008. Ecological models and data in R. Princeton, NJ: Princeton University Press.

Bolker B, Brooks ME, Clark CJ, *et al.* 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol* **24**:127-135.

Breiman L. 2001. Statistical modeling: the two cultures (with discussion). *Stat. Sci* **16**:199-231.

Butcher JA, Groce JE, Lituma CM, *et al*. 2007. Persistent controversy in statistical approaches in wildlife sciences: a perspective of students. *J Wildlife Manage* **71**:2142-2144.

Clark JS. 2007. Models for ecological data: an introduction. Princeton, NJ: Princeton University Press.

Clark JS, Carpenter SR, Barber M, *et al.* 2001. Ecological forecasts: an emerging imperative. *Science* **293**:657-660.

Conover WJ. 1998. Practical nonparametric statistics, 3rd ed. New York: John Wiley and Sons.

Cope L. 2006. Understanding and using statistics. In: Blum D, Knudson M, and Henig RM (Eds). A field guide for science writers: the official guide of the National Association of Science Writers, 2nd ed. New York: Oxford University Press

delMas RC. 2002. Statistical literacy, reasoning, and learning: a commentary. *J Stat Ed* **10**.

De Veaux RD, and Velleman PF. 2008. Math is music; statistics is literature (or, why are there no six-year-old novelists?). *AmStat News* **September 2008**:54-58.

Desai A, Moorcroft PR, Bolstad PV and Davis KJ. 2007. Regional carbon fluxes from an observationally constrained dynamic ecosystem model: impacts of disturbance, CO_2 fertilization, and heterogeneous land cover. *J. Geophys Res* **112**: G01017

Devore JL. 2007. Probability and statistics for engineering and the sciences, 7th ed. Belmont, CA: Duxbury Press.

Gotelli NJ, and Ellison AM. 2004. A primer of ecological statistics. Sunderland, MA: Sinauer Associates.

Hill BM. 1968. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *J Am Stat Assoc* **63**:677-691.

Hubbard R, and Byarri MJ. 2003. Confusion over measures of evidence (*p*'s) versus errors (α's) in classical statistical testing (with discussion and rejoinder). *Am Stat* **57**:171-182.

Hubbell SP. 2001. The unified neutral theory of biodiversity and biogeography. Princeton, NJ: University Press.

Hutchinson G.E. 1978. An introduction to population ecology. New Haven, CT: Yale University Press.

IPCC. 2007. Climate Change 2007: Synthesis report. Contribution of working groups I, II, and III to the fourth assessment report of the intergovernmental panel on climate change. Geneva: IPCC.

Larson RJ, and Marx ML. 2005. An introduction to mathematical statistics and its applications, 4th ed. Upper Saddle River, NJ: Prentice-Hall.

Leigh, EG Jr. 2007. Neutral theory: a historical perspective. *J Evol Biol* **20**:2075-2091.

Lele SR, and Dennis B. 2009. Bayesian methods for hierarchical models: are ecologists making a Faustian bargain? *Ecol Appl* **19**:581-584.

Manly BJF. 2006. Randomization, bootstrap and Monte Carlo methods in biology, 3rd ed. Boca Raton, FL: CRC Press.

Matross DM *et al.* 2006. Estimating regional carbon exchange in New England and Quebec by combining atmospheric, ground-based and satellite data. *Tellus* **58B**: 344-358.

Millennium Ecosystem Assessment. 2005. Ecosystems and human well-being: synthesis. Washington, DC: Island Press.

Molles M. C. 2006. Ecology: concepts and applications, 4th edition. New York: McGraw-Hill.

Murdoch DJ., Tsai Y-L, and Adcock J. 2008. *p*-values are random variables. *Am Stat* **62**:242-245.

National Science Foundation. 1996. *Undergraduate origins of recent (1991-95) science and engineering doctorate recipients,* detailed statistical tables, NSF 96-334. Arlington, Virginia: National Science Foundation

Poole RW. 1974. An introduction to quantitative ecology. New York: McGraw-Hill.

Rice JA. 2006. Mathematical statistics and data analysis, 3rd edition. Belmont, CA: Duxbury Press.

Stoll EL. 1983. Mark I. In: Ralston A and Reilly ED (Eds.) Encyclopedia of computer science and engineering, 2nd ed. New York: Van Nostrand Reinhold.

Turchin P. 2001. Does population ecology have general laws? *Oikos* **94**:17-26.

Wackerly D, Mendenhall W, and Scheaffer RL. 2007. Mathematical statistics with applications, 7th edition. Belmont, CA: Duxbury Press.

White GC and Burnham KP. 1999. Program MARK: Survival estimation from populations of marked animals. *Bird Study* **46 (Supp.)**: 120-138.

Table 1. The different components and stages of statistical literacy.* "Process" refers to a statistical concept (such as a *P*-value or confidence interval) or method.

Basic literacy	**Ability to reason statistically**	**Fluency in statistical thinking**
Identify the process	Explain the process	Apply the process to new situations
Describe it	Why does it work?	Critique it
Rephrase it	How does it work?	Evaluate it
Translate it		Generalize from it
Interpret it		

* modified from delMas 2002

Figure Legends

Figure 1 – *Milestones in statistical computing.* **A**. *Women (ca. 1920) in the Computing Division of the U.S. Department of the Treasury (or the Veterans' Bureau) determining the bonuses to be distributed to veterans of World War I. Photograph from the Library of Congress Lot 12356-2, negative LC-USZ62-101229*. **B.** *Professor (and Commander) Howard Aiken, Lieutenant (and later Rear Admiral) Grace Hopper, and Ensign Campbell in front of a portion of the Mark I Computer. The Mark I was designed by Aiken, built by IBM, fit in a steel frame 16 m long × 2.5 m high, weighed approximately 4,500 kg, and included 800 km of wire, It was used to solve integrals required by the U.S. Navy Bureau of Ships during World War II, and physics problems associated with magnetic fields, radar, and the implosion of early atomic weapons. Grace Hopper was the lead programmer of the Mark I. Her experience developing its programs led her to develop the first compiler for a computer programming language (which subsequently evolved into COBOL), and she developed early standards for both the FORTRAN and COBOL programming languages. The Mark I was programmed using punched paper tape and was the first automatic digital computer in the U.S. Its calculating units were mechanically synchronized by an ~ 15-m long drive shaft connected to a 4 kW (5 horsepower) electric motor. The Mark I is considered to be the first universal calculator (Stoll 1983). Photograph from the Harvard University Office of News and Public Affairs, Harvard University Archives call number HUPSF Computers (2), and reproduced with permission of the Harvard University Archives.* **C.** *A ca. 2007 screen-shot of the open-source* R *statistical package running on a personal computer. The small, notebook computers that on which we run* R *and other statistical software every day have central processors that execute10,000 – 100,000 MIPS (million instructions per second). In contrast, the earliest commercial computers executed 0.06-1.0 KIPS (thousand instructions per*

second), and Harvard's Mark I computer took approximately 6 seconds to simply multiply two numbers together; computing a single logarithm took more than a minute. (Image from http://www.r-project.org, copyright the R Foundation, and used with permission).

Figure 2 - *Total number of quantitative courses, calculus courses, and statistics courses required at the 25 liberal-arts colleges and universities that produce the majority of students who go on to receive Ph.D.s in the life sciences. Institutions surveyed are based on data from the National Science Foundation (1996). Data collected from department web sites and college or university course catalogs, July 2008.*

Figure 3 – *Telling a compelling ecological story requires quantitative data. Here, Harvard Forest researcher Julian Hadley describes monthly cycles of carbon storage in hemlock and hardwood stands. The data are collected at 10-20 Hz from three eddy-covariance towers, analyzed and summarized with time-series modeling, and incorporated into regional estimates (*e.g., *Matross* et al. *2006) and forecasts (*e.g., *Desai* et al. *2007), and used to determine regional and national carbon emissions targets and policies. Photograph by David Foster, and used with permission of the Harvard Forest Archives.*

Figure 1A

Figure 1B

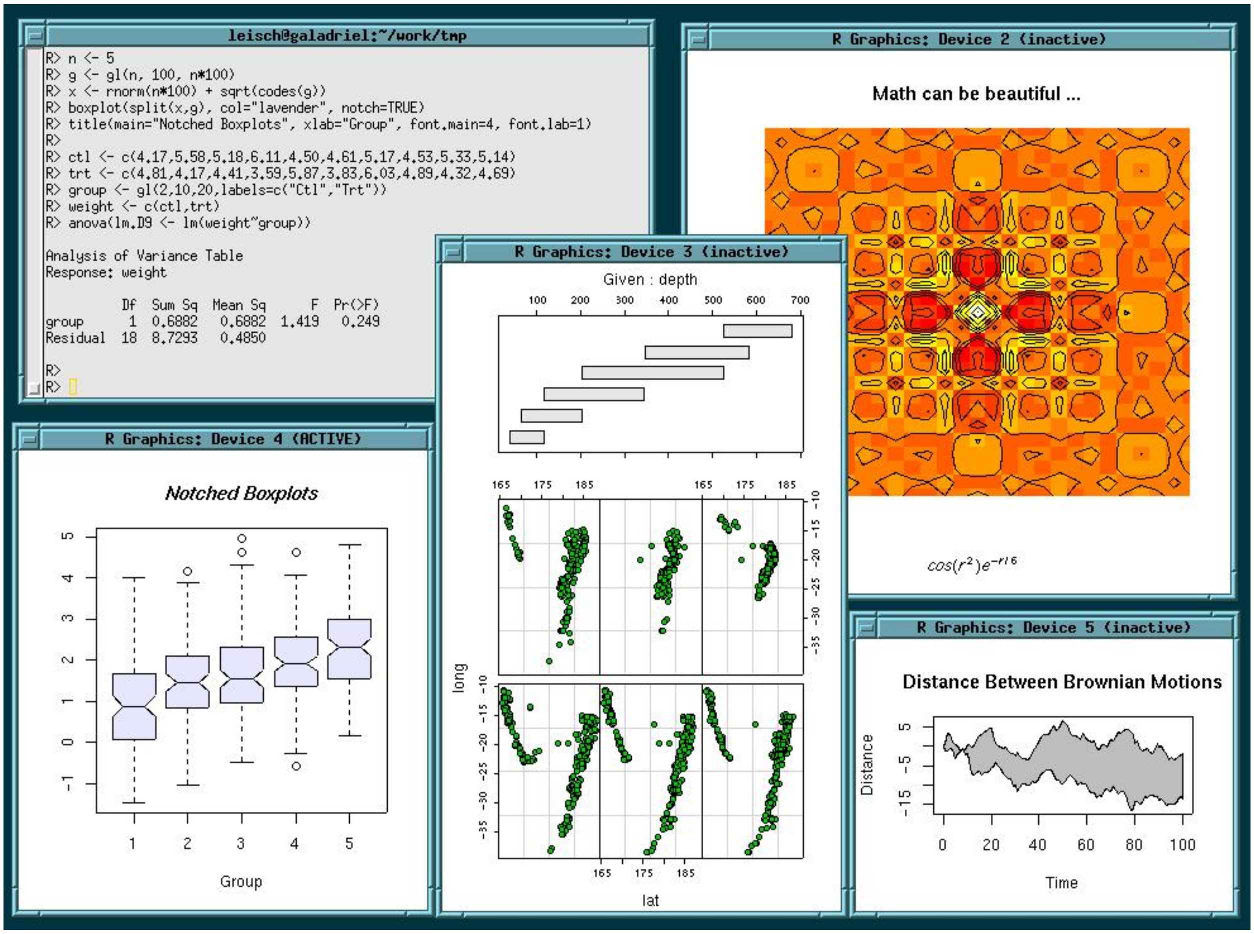
leisch@galadriel:~/work/tmp
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))
Analysis of Variance Table
Response: weight
Df Sum Sq Mean Sq F Pr(>F)
group 1 0.6882 0.6882 1.419 0.249
Residual 18 8.7293 0.4850
R>
R>
R Graphics: Device 2 (inactive)
Math can be beautiful ...
$cos(r^2)e^{-r/6}$
R Graphics: Device 3 (inactive)
Given : depth
long
lat
R Graphics: Device 4 (ACTIVE)
Notched Boxplots
Group
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
Time

Figure 1C

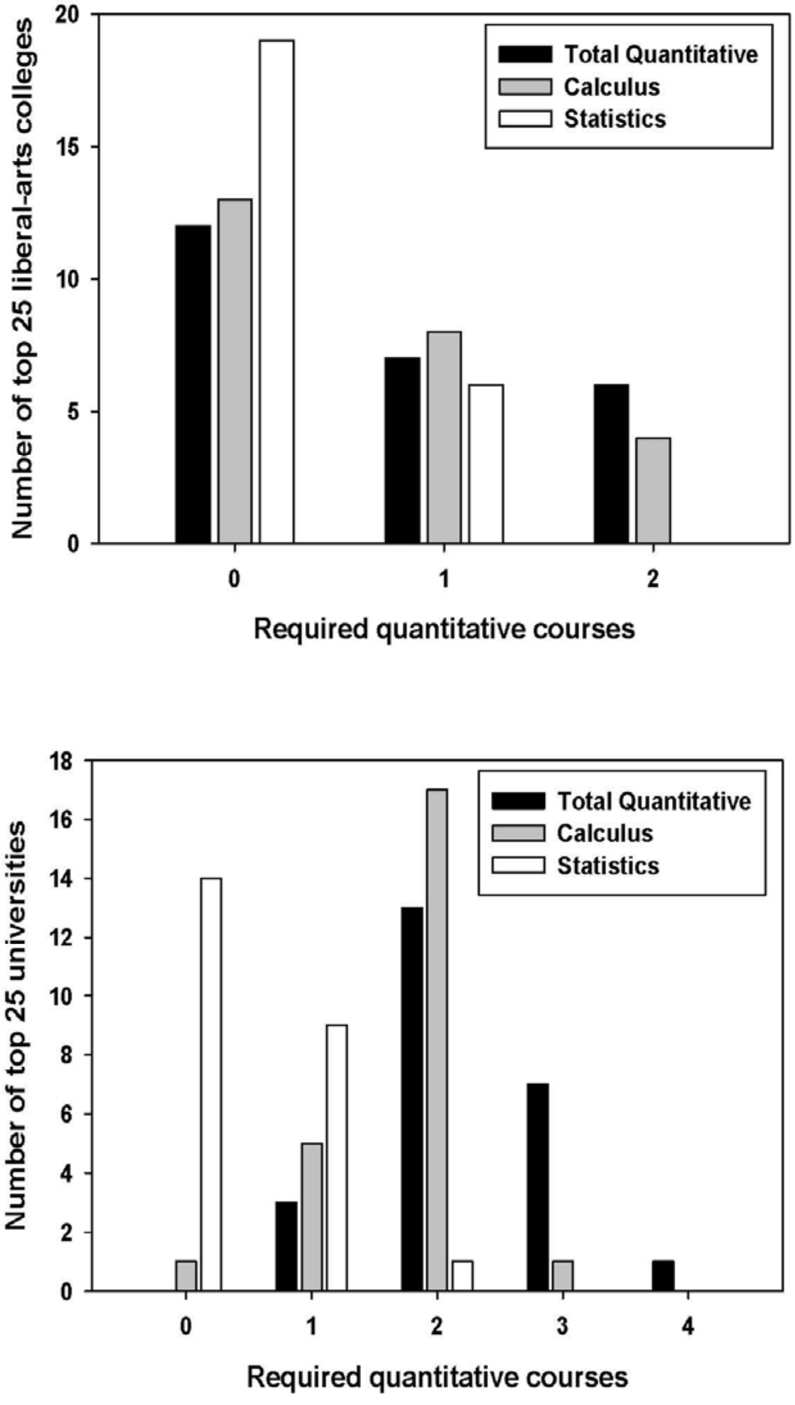
Number of top 25 liberal-arts colleges
Total Quantitative
Calculus
Statistics
Required quantitative courses
Number of top 25 universities
Total Quantitative
Calculus
Statistics
Required quantitative courses

Fig. 2

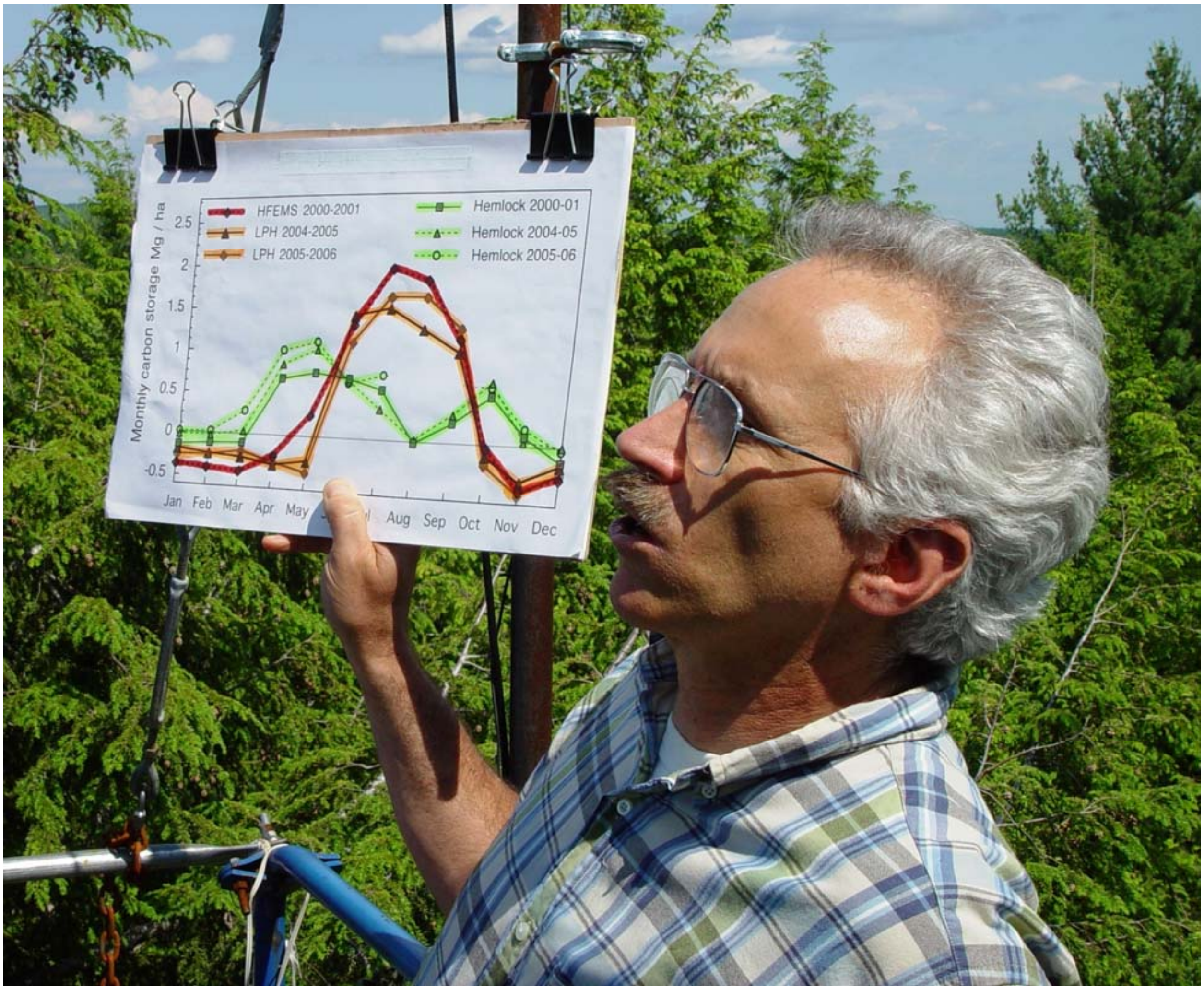
HFEMS 2000-2001
LPH 2004-2005
LPH 2005-2006
Hemlock 2000-01
Hemlock 2004-05
Hemlock 2005-06
Monthly carbon storage Mg / ha
2.5
2
1.5
1
0.5
0
-0.5
Jan
Feb
Mar
Apr
May
Aug
Sep
Oct
Nov
Dec

Figure 3.

Box 1. Why "$P(10.2 < \mu < 29.8) = 0.95$" is not a correct interpretation of confidence interval, and what are confidence intervals, anyway?

This statement says that the probability that the true population mean μ lies in the interval (10.2, 29.8) equals 0.95. But μ is a fixed (but unknown) constant: it is either in the interval (10.2, 29.8) or it is not. The probability that μ is in the interval is zero or one; we just do not know which. A confidence interval actually asserts that 95% of the confidence intervals resulting from hypothetical repeated samples (taken under the same random sampling protocol used for the single sample) will contain μ in the long run. Think of a game of horseshoes in which you have to throw the horseshoe over a curtain positioned so that you cannot see the stake. You throw a horseshoe and it lands (thud!); the probability is zero or one that it is a ringer, but you do not know which. The confidence interval arising from a single sample is the horseshoe on the ground, and μ is the stake. If you had the throwing motion practiced so that the long run proportion of successful ringers was 0.95, then your horseshoe game process would have the probabilistic properties claimed by 95% confidence intervals. You do not know the outcome (whether or not μ is in the interval) on any given sample, but you have constructed the sampling process so as to be assured that 95% of such samples in the long run would produce confidence intervals that are ringers. The distinction is clearer when we write the probabilistic expression for a 95% confidence interval:

$$P(L \le \mu \le U) = 0.95$$

What this equation is telling us is that the true (but unknown) population mean μ will be found 95% of the time in an interval bracketed by *L* at the lower end and *U* at the upper end, *where L* *and U vary randomly from sample to sample*. Once the sample is drawn, the lower and upper bounds of the interval are fixed (the horseshoe has landed), and μ (the stake) is either contained in the interval or it is not.

Many standard statistical methods construct confidence intervals symmetrically in the form of a "point estimate" plus or minus a "margin of error". For instance, a $100(1-\alpha)\%$ confidence interval for μ when sampling from a Normal distribution is constructed based on the following probabilistic property:

$$P(\overline{Y} - t_{\alpha/2}\sqrt{S^2/n} \leq \mu \leq \overline{Y} + t_{\alpha/2}\sqrt{S^2/n}) = (1-\alpha) \ .$$

Here $t_{\alpha/2}$ is the percentile of a t-distribution with $n - 1$ degrees of freedom such that there is an area of $\alpha/2$ under the t-distribution to the right of $t_{\alpha/2}$, and $\overline{Y}$ and S^2 are respectively the sample mean and sample variance of the observations. The quantities $\overline{Y}$ and S^2 vary randomly from sample to sample, making the lower and upper bounds of the interval vary as well. The confidence interval itself becomes $\overline{y} \pm t_{\alpha/2}\sqrt{s^2/n}$, in which the lowercase $\overline{y}$ and s^2 are the actual numerical values of sample mean and variance resulting from a single sample. In general, modern-day confidence intervals for parameters in non-Normal models arising from computationally intensive methods such as bootstrapping and profile likelihood are not necessarily symmetric around the point estimates of those parameters.

Box 2. A prescription for statistical fluency.

The problem of how to use calculus in the context of developing statistical fluency can be solved *easily* and *well* by rearranging courses and substituting different statistics courses (those hitherto rarely taken by ecologists) for many of the statistical methods courses now taken in college and graduate school. The suggested courses are standard ones, with standard textbooks, and already exist at most universities. Our prescription is as follows.

For *undergraduate majors* in the ecological sciences (including "integrative biology", ecology, evolutionary biology), along with students bound for scientific careers in resource management fields such as wildlife, fisheries, and forestry:

1. **At least two semesters of standard calculus**. "Standard" means real calculus, the courses taken by students in physical sciences and engineering. Those physics and engineering students go on to take a third (multivariable calculus) and a fourth semester (differential equations) of calculus, but these latter courses are not absolutely necessary for ecologists. Only a small amount of the material in those additional courses is used in subsequent statistics or ecology courses and can be introduced in those courses or acquired through self-study. Most population models must be solved numerically, methods for which can be covered in the population ecology courses themselves. (Please note we do not wish to discourage additional calculus for those students interested in excelling in ecological theory; our prescription, rather, should be regarded as minimum core for those who will ultimately have Ph.Ds in the ecological sciences, broadly defined.)

2. **An introductory statistics course which lists calculus as a prerequisite**. This course is standard everywhere; it is the course that engineering and physical science students take, usually as juniors. A typical textbook is Devore (2007).

3. **A commitment to using calculus and post-calculus statistics in courses in life-science curricula** must go hand-in-hand with course requirements in calculus and post-calculus statistics. Courses in the physical sciences for physical science majors use the language of science – mathematics – and its derived tool – statistics – unapologetically, starting in beginning courses. Why don't ecologists or other life scientists do the same? The basic ecology course for majors should include calculus as a prerequisite and must use calculus so that students see its relevance.

For *graduate students* in ecology (*sensu lato*):

1. **A standard two-course sequence in probability and mathematical statistics.** This sequence is usually offered for undergraduate seniors and can be taken for graduate credit. Typical textbooks are Rice (2006), Larson and Marx (2005), or Wackerly *et al.* (2007). The courses usually require two semesters of calculus as prerequisites.

2. **Any additional graduate-level course(s) in statistical methods**, according to interests and research needs. After a two-semester post-calculus probability and statistics sequence, the material covered in many statistical methods courses also is amenable to self-study.

3. M**ost ecologists will want to acquire some linear algebra** somewhere along the line, because matrix formulations are used heavily in ecological and statistical theory alike. Linear algebra could be taken either in college or graduate school. Linear algebra is often reviewed extensively in courses such as multivariate statistical methods and population

ecology, and necessary additional material can be acquired through self-study. Those ecologists whose research is centered on quantitative topics should consider formal coursework in linear algebra.

The benefit of following this prescription is a rapid attainment of statistical fluency. Whether students in ecology are focused more on theoretical ecology or on field methods, conservation biology, or the interface between ecology and the social sciences, a firm grounding in quantitative skills will make for better teachers, better researchers, and better interdisciplinary communicators (for good examples see Armsworth *et al.* 2009 and other papers in the associated special feature on "Integrating ecology and the social sciences" in the April 2009 issue of the *Journal of Applied Ecology*). Since our prescription replaces courses rather than adds new ones, the primary cost to swallowing this pill is either to recall and use calculus taken long ago or to take a calculus refresher course.