DRAFT: January, 2001
FOR: The Nature of Scientific Evidence
M. L. Taper and S. R. Lele, editors
The University of Chicago Press

STATISTICS AND THE SCIENTIFIC METHOD IN ECOLOGY

Brian Dennis
Department of Fish and Wildlife Resources
and
Division of Statistics
University of Idaho
Moscow, ID 83844-1136 USA

phone: 208-885-7423
fax: 208-885-9080
email: brian@uidaho.edu

Abstract. Ecology as a science is under constant political pressure. The science is difficult and progress is slow, due to the variability of natural systems and the high cost of obtaining good data. Ecology however is charged with providing information support for environmental policy decisions with far reaching societal consequences. Demand for quick answers is strong, and demand for answers that agree with a particular point of view is even stronger.

The use of Bayesian statistical analyses has recently been advocated in ecology, supposedly to aid decision makers and enhance the pace of progress. Bayesian statistics provides conclusions in the face of incomplete information. Bayesian statistics, though, represents a much different approach to science than the frequentist statistics studied by most ecologists. The scientific implications of Bayesian statistics are not well understood.

I provide a critical review of the Bayesian approach. I compare, using a simple sampling example, the Bayesian and frequentist analyses. The Bayesian analyses can be "cooked" to produce results consistent with any point of view, because Bayesian analyses quantify prior personal beliefs and mix them with the data. In this, Bayesian statistics is consistent with the postmodern view of science, widely held among nonscientists, in which science is just a system of beliefs that has no particular authority over any other system of beliefs. By contrast, modern empirical science uses the scientific method to identify empirical contradictions in skeptics' beliefs and permit replication and checking of empirical results. Frequentist statistics has become an indispensible part of the scientific method.

I also undertake a critical discussion of statistics education in ecology. Part of the potential appeal of Bayesian statistics is that many ecologists are confused about frequentist statistics, and statistical concepts in general. I identify the source of confusion as arising from ecologists' attempts to learn statistics through a series of precalculus "statistical methods" courses taken in graduate school. I prescribe a radical change in the statistical training of ecological scientists which will greatly increase the level of confidence and facility with statistical thinking.

INTRODUCTION

Science in the crosshairs

Tobacco company scientists argue that there is no evidence that smoking tobacco is harmful. Biblical creation scientists argue that the evidence for evolution is weak. Institutes paid for by industry are staffed by degreed scientists, whose job it is to create in the minds of politicians and the public the illusion of major scientific disagreements on environmental issues.

We are awash in a sea of popular postmodernism. Fact in the postmodern view is just strongly-held belief (Anderson 1990). Native Americans, according to tribal creation stories, did not originally cross over from Asia, but arose independently and originally in the Americas. The U.S. government is concealing dark secrets about the existence of extraterrestrial life, secrets which became partly exposed after a crash near the town of Roswell, New Mexico. Crystals have health and healing properties. One's personality and tendencies are influenced by the positions of solar system objects at the moment of birth. O. J. Simpson did not kill Niccole Brown and Ron Goldman.

The postmodern outlook is not confined to popular culture, but permeates intellectual life as well. Humanities disciplines at universities have abandoned the traditional empirical view of science (Sokal and Bricmont 1998). Feminists claim that the methods and requirements of science are biased against females. According to feminist scholars, if females formed the reigning power structure, science would be more cooperative in occupation and more tolerant of multiple explanations. "Science studies" historians focus on questionable behavior of well-known scientists toward colleagues, in order to expose science as a subjective power struggle. Multicultural philosophers portray science as just another of many legitimate ways of knowing, its successes due primarily to the dominance of European culture. Scientists' writings are deconstructed by literary theorists in order to reveal how the scientists were trapped by the prevailing cultural mental prisons. Political polemics from the intellectual left and right reveal disbelief and disrespect for established scientific knowledge.

Ecology

The questioning of science and the scientific method continues within the science of ecology.

Ecology has become a highly politicized science. Once a quiet backwater of biology, ecology burst into high public profile in the early 1970s (Earth Day, April 1970, was a watershed event) with the emergence of popular concern about environmental issues. After passage of the National Environmental Policy Act, the Endangered Species Act, the National Forest Policy and Management Act, and many other federal and state laws, ecologists and the findings of ecology suddenly had great influence in the lives of people everywhere. The scientific information from ecology, coupled with the environmental laws, forced constraints on peoples' behavior and economic activity.

Many ecological topics, from evolution to conservation biology to global climate change, hit people close to home. As a result, scientific signals are often masked or distorted by political noise. An ecological discovery that has impact on human conduct will often have a debunking campaign mounted against it by monied interests. Government agency scientists are

sometimes muzzled and have their findings reversed by the pen-stroke of a political appointee. Natural resource departments at state universities are pressured by special interest groups. Radio talk show hosts set themselves up as authoritative spokespersons on environmental topics.

Among practicing, credentialed ecologists, the science itself is quite contentious. The topic intrinsically attracts many participants, and the competition for admission to programs, jobs, journal space, grants, and recognition is fierce. Severe scientific and political infighting surfaces during position searches at university departments. Anonymous peer reviews can be ignorant and vindictive. Resources for curiosity-driven research are scarce; ecological research is funded more and more by agencies and companies with particular agendas. Pressures mount from environmental decision-makers for definitive answers. In this postmodern cacaphony, how can a healthy ecological science thrive?

In fact, some ecologists in the past couple of decades have questioned whether the Popperian hypothetico-deductive approach, and the collection of inquiry devices known as the scientific method, are too constraining for ecology. Good empirical data in ecology have often been too slow in coming or too difficult and expensive to collect, and scientific progress in ecology has seemed painfully slow. The calls for relaxed scientific guidelines have come from two main sources. First, some "theoretical ecologists" have sought scientific respect for their pencil, paper, and computer speculations on ecological dynamics. Their mathematical models however, frequently play the role of "concepts" rather than "hypotheses", due to the lack of connections to data and the lack of widely accepted ecological laws with which to build models. As a result, theoretical ecologists have called for judging mathematical models under different criteria than scientific hypotheses would be judged. Articles in the November 1983 issue of *The American Naturalist* debated this question, among others, within the context of community ecology.

Second, applied ecologists and social decision-makers have often viewed the beetles-and-butterflies focus of ecological natural history research to be an unaffordable luxury. Academic ecology research is an exotic world of Galapagos birds, Carribean lizards, jungle orchids, and desert scorpions; it is slow, intellectual, and to some onlookers, produces few useful generalities. Yet, ecology also deals with vital topics within which major social decisions must be made, regardless of the amount of evidence available. Answers, in the form of "best judgements" by experts are needed, fast. For example, will breaching the Columbia watershed dams save salmon, or not?

A partial reading list for a seminar course on the "scientific method in ecology" might include Connor and Simberloff (1979, 1986), Saarinen (1980), *The American Naturalist* (1983 November), Hurlbert (1984), Strong et al. (1984), Hairston (1989), Underwood (1990), Peters (1991), Schrader-Frechette and McCoy (1993), and Dixon and Garrett (1993).

Inevitably bound up in this question about ecological science are concerns about statistical practice. The lack of true replication in many ecological experiments exposed by Hurlburt (1984) was a shocker. The lack of attention to power in many ecological studies has also been criticized (Toft and Shea 1983, Peterman 1990). Distribution-free statistical methods have been advocated (Potvin and Roff 1993), but some advocacy arguments have been challenged (Smith 1995, Johnson 1995). Stewart-Oaten (1995) criticized the tendency for

ecologists to view statistics as a set of procedural rules for data analysis. The widespread misinterpretation of statistical hypothesis testing has inspired much discussion (Simberloff 1990, Underwood 1990, Yoccoz 1991, Johnson 1999). The lack of statistical connections between "nonlinear dynamics" models and ecological data was criticized (Dennis et al. 1995). Specific ecological topics, such as the prevalence of density dependent population regulation, have spawned their own statistical literature (see Dennis and Taper 1994).

Bayesian statistics

According to some (Reckhow 1990, Ellison 1996, Johnson 1999), there is a statistical solution to many of ecology's ills. The touted solution is Bayesian statistics. Bayesian statistics is remarkably different from the variety of statistics called frequentist statistics that most of us learned in college. Bayesian statistics abandons many concepts that most of us struggled (with mixed success) to learn: hypothesis testing, confidence intevals, *P*-values, standard errors, power. Bayesians claim to offer improved methods for assessing the weight of evidence for hypotheses, making predictions, and making decisions in the face of inadequate data. In a cash-strapped science charged with information support in a highly contentious political arena, the Bayesian promises are enticing to ecological researchers and managers alike.

But is there a price to pay? You bet. Bayesians embrace the postmodern view of science. The Bayesian approach abandons notions of science as a quest for "objective" truth and scientists as detached, skeptical observers. Like postmoderns, Bayesians claim that those notions are misleading at best. In the world of Bayesian statistics, truth is personal and is measured by blending data with personal beliefs. Bayesian statistics is a way of explicitly organizing and formulating the blending process.

There is an enormous literature on Bayesian statistics. A glance at the titles in any current statistics journal (say, *Journal of the American Statistical Association*, or *Biometrika*) might convince a casual onlooker that the world of statistics is becoming Bayesian. The Bayesian viewpoint is indeed gaining influence. The burgeoning literature, however, tends to be highly mathematical, and a scientist is right to question whether the attraction is mathematical instead of scientific. Actually, frequentism is alive and well in statistics. Introductory textbooks and courses remain overwhelmingly frequentist, as do canned computer statistics packages available to researchers. Frequentist and Bayesian statisticians waged war for many years, but the conflict quieted down around 1980 or so, and the two camps coexist now in statistics without much interaction.

Ecology, however, represents fertile, uncolonized ground for Bayesian ideas. The Bayesian-frequentist arguments, which many statisticians tired of twenty years ago, have not been considered much by ecologists. A handful of Bayesian papers have appeared in the ecological literature (see the featured group of articles in the November 1996 *Ecological Applications*, vol. 6(4)). Their enthusiastic exposition of Bayesian methods, and portrayal of frequentism as a anachronistic yoke impeding ecological progress, has attracted the attention of natural resource managers (Marmorek 1996).

The Bayesian propagule has arrived at the shore. Ecologists need to think long and hard about the consequences of a Bayesian ecology. The Bayesian outlook is a successful competitor, but is it a weed?

I think so. In this paper, I attempt to draw a clear distinction for ecologists between Bayesian and frequentist science. I address a simple environmental sampling problem and discuss the differences between the frequentist and the Bayesian statistical analyses. While I concur with Bayesians regarding critiques of some of the imperfections of frequentism, I am alarmed at the potential for disinformation and abuse that Bayesian statistics would give to environmental pressure groups and biased investigators. At the risk of repeating a lot of basic statistics, I develop the sampling example rather extensively from elementary principles. The aim is to amplify the subtle and not-so-subtle conflicts between the Bayesian and frequentist interpretations of the sampling results.

Readers interested in a more rigorous analysis of the scientific issues in the frequentist/Bayesian debate are urged to consult Mayo's (1996) comprehensive account.

One thing has become painfully clear to me in twenty years of extensive teaching, statistical consulting, reviewing, and interacting in ecology. Ecologists' understanding of statistics in general is abysmally poor. Statistics, which should naturally be a source of strength and confidence to an ecologist, no matter how empirically oriented he/she is, is all too frequently a source of weakness, insecurity, and embarrassment. The crucial concepts of frequentism, let alone Bayesianism, are widely misunderstood. I place the blame squarely on ecologists' statistical educations, which I find all wrong. In a later section of this paper, I offer some solutions to this problem. Ecologists, whether Bayesian or frequentist, will be better served by statistics with a radical revision of university statistics coursework.

WHAT IS FREQUENTISM?

"Nature cannot be fooled." —Richard Feynman

Suppose a reach of a stream is to be sampled for Cu pollution. A total of 10 samples will be collected from the reach in some random fashion, and Cu concentration ($\mu g\, l^{-1}$) will be determined in each sample.

The purpose of the samples is to estimate the average concentration of Cu in the water at the time of sampling. The sampling could be a part of an ongoing monitoring study, an upstream/downstream/before/after study, or similar such study.

Frequentist statistics involves building a probability model for the observations. The modeling aspect of statistics is crucial to its understanding and proper use; however, the pre-calculus statistics methods courses taken by ecologists-in-training tend to emphasize formulas instead of models. I therefore develop the modeling aspect in more detail than is customary in frequentist analyses, so that the approach may be contrasted properly with the Bayesian way.

We might suppose that theobservations of Cu concentration could be modeled as if they arose independently from a normal distribution with mean μ and variance σ^2. Applied statistics texts would word this model as follows: each sample is assumed to be drawn at random from a "population" of such samples, with the population of Cu values having a frequency distribution well-approximated by a normal curve. The population mean is μ, and the population variance is σ^2. Mathematical (i.e. post-calculus) statistics texts would state: the observations $\{X_i\}, i = 1, 2, ..., n$ are assumed to be independent, identically distributed normal(μ, σ^2) random variables. Regardless of the wording and symbology, the important

point is that a probability model is assumed for *how the variability in the data arose*. The analyses are based on the model, and so it will be important to evaluate the model assumptions somehow. If the model is found wanting, then proper analyses will require construction of some other model.

We suppose the observations are drawn; their numerical values are x_1, x_2, ..., x_n ($n = 10$). Symbols are used for the actual values drawn so that the subsequent formulas will be general to other data sets; the lower case notation indicates fixed constants (sample already drawn) instead of upper-case random variables (sample yet to be drawn). The distinction is absolutely crucial in frequentist statistics and is excruciating for teachers and students alike. (Educators in ecology should be aware that pre-calculus basic statistics textbooks, in response to the overwhelming symbol allergies of today's undergraduates, have universally abandoned the big-X little-x notation, and with it all hope that statistics concepts are intended to be understood). Let us suppose that the investigator has dutifully calculated some summary statistics from the samples, in particular the sample mean $\overline{x} = (\sum_{i=1}^{n} x_i)/n$ and the sample variance $s^2 = [\sum_{i=1}^{n} (x_i - \overline{x})^2]/(n-1)$, and that the resulting numerical values are:

$$\overline{x} = 50.6,$$

$$s^2 = 25.0.$$

The probability model for the observations is represented mathematically by the normal distribution, with probability density function (pdf) given by

$$f(x) = (\sigma^2 2\pi)^{-1} \exp[-(x-\mu)^2/(2\sigma^2)], \; -\infty < x < \infty .$$

This is the bell-shaped curve. The cumulative distribution function (cdf) is the area under the curve between $-\infty$ and x and is customarily denoted $F(x)$. The probabilistic meaning of the model is contained in the cdf; it is the probability that a random observation X will take a value less than or equal to some particular constant value x:

$$F(x) = \mathrm{P}[X \le x] = \int_{-\infty}^{x} f(u)du .$$

Again, the lower and upper case X's have different meanings. The constants μ and σ^2 are "parameters." In applied stat texts, μ and σ^2 are interpreted respectively as the mean and variance of the "population" being sampled. Bear in mind that the population here is the collection of all possible samples that could have been selected on that sampling occasion. It is this potential variability of the samples that is being modeled in frequentist statistics.

An essential concept to master for understanding frequentist and Bayesian statistics alike is the likelihood function. The pdf $f(x)$ quantifies the relative frequency with which a single observation takes a value within a tiny interval of x. The whole sample, however, consists of n observations. Under the independence assumption, the product $f(x_1)f(x_2)\ldots f(x_n)$

quantifies the relative frequency with which the whole sample, if repeated, would take values within a tiny interval of x_1, x_2, ... , x_n, the sample actually observed. The product is the "probability of observing what you observed" relative to all other possible samples in the population. Mathematically, the random process consists of n independent random variables X_1, X_2, ..., X_n. The product $f(x_1)f(x_2)\ldots f(x_n)$ is the joint pdf of the process, evaluated at the data values.

The joint pdf of the process, evaluated at the data, is the *likelihood function*. For this normal model, the likelihood function is a function of the parameters μ and σ^2. The relative likelihood of the sample x_1, x_2, ..., x_n depends on the the values of the parameters; if μ were 100 and σ^2 were 1, then the relative chance of observing sample values clustered around 50 would be very small indeed. Written out, the likelihood function for this normal model is

$$L(\mu, \sigma^2) \;=\; f(x_1)f(x_2)\ldots f(x_n)$$

$$=\; \left(\sigma^2 2\pi\right)^{-n/2} \exp\left[\,-\left(2\sigma^2\right)^{-1}\sum_{i=1}^{n}(x_i \;-\; \mu)^2\right].$$

An algebraic trick well-known to statisticians is to add $-\,\overline{x} + \overline{x}$ inside each term $(x_i - \mu)$ in the sum. Squaring the terms and summing expresses the likelihood function in terms of two sample statistics, $\overline{x}$ and s^2:

$$L(\mu, \sigma^2) \;=\; \left(\sigma^2 2\pi\right)^{-n/2} \exp\left\{\,-\left(2\sigma^2\right)^{-1}\left[(n-1)s^2 \;+n\,(\mu - \overline{x})^2\right]\right\}. \qquad (1)$$

Only the numbers $\overline{x}$ and s^2 are needed to calculate the likelihood for any particular values of μ and σ^2; once $\overline{x}$ and s^2 are in hand, the original data values are not required further for estimating the model parameters. The statistics $\overline{x}$ and s^2 are said to be *jointly sufficient* for μ and σ^2.

Because likelihood functions are typically products, algebraic and computational operations are often simplified by working with the log-likelihood function, ln L:

$$\ln L(\mu, \sigma^2) \;=\; -(n/2)\ln(2\pi) \;-\; (n/2)\ln\sigma^2$$

$$-\,(n-1)s^2/(2\sigma^2) \;-\; n(\mu - \overline{x})^2/(2\sigma^2)\,. \qquad (2)$$

Modern frequentist statistics can be said to have been inaugurated in 1922 by R. A. Fisher, who first realized the importance of the likelihood function (Fisher 1922). Fisher noted that the likelihood function offers a way of using data to *estimate* the parameters in a model if the parameter values are unknown. Subsequently, J. Neyman and E. S. Pearson used the likelihood function to construct a general method of statistical hypothesis testing, that is, the use of data to select between two rival statistical models (Neyman and Pearson 1933). More recently, the work of H. Akaike (1973, 1974) launched a class of likelihood-based methods for model selection when there are more than two candidate models from which to choose.

Two cases for inferences about μ will be considered: σ^2 known, and σ^2 unknown. The "σ^2 known" case is obviously of limited practical usefulness in ecological work. However, it allows a simple and clear contrast between the Bayesian and frequentist approaches. The "σ^2 unknown" case highlights the differences in how so-called "nuisance parameters" are handled in the Bayesian and frequentist contexts, and also hints at the numerical computing difficulties attendant with the use of more realistic models. I concentrate on point estimates, hypothesis tests, and confidence intervals.

σ^2 *known*

We assume that σ^2 is a known constant, say, $\sigma^2 = 36$.

Fisher (1922) developed the concept of *maximum likelihood* (ML) estimation. The value of μ, call it $\hat{\mu}$, that maximizes the likelihood function (Eq. 1) is the ML estimate. The ML estimate also maximizes the log-likelihood function (Eq. 2). It is a simple calculus exercise to show that Eq. 2 is maximized by

$$\widehat{\mu} = \overline{x}\,.$$

Thus, a point estimate for μ calculated from the sample is $\hat{\mu} = 50.6$.

ML point estimates were shown by Fisher (1922) and numerous subsequent investigators to have many desirable statistical properties, among them: (a) Asymptotic unbiasedness (statistical distribution of estimate approaches a distribution with the correct mean as n becomes large). (b) Consistency (distribution of the parameter estimate concentrates around the true parameter value as n becomes large). (c) Asymptotic normality (distribution of the parameter estimate approaches a normal distribution, a celebrated central limit theorem-like result). (d) Asymptotic efficiency (the asymptotic variance of the parameter estimate is as small as is theoretically possible). These and other properties are thoroughly covered by Stuart and Ord (1991). Deriving these properties forms the core of a modern Ph. D.-level mathematical statistics course (Lehmann 1983).

These statistical properties refer to behavior of the estimate under hypothetical repeated sampling. To illustrate, the whole population of possible samples induces a whole population of possible ML estimates. In our case, to each random sample X_1, X_2, ..., X_n there corresponds a sample mean $\overline{X}$, a random variable. The frequency distribution of the possible estimate values is the *sampling distribution* of the ML estimate. The sampling distribution plays no role in Bayesian inference, but is a cornerstone of frequentist analyses.

When reading statistics papers, one should note that the "hat" notation for estimates (e.g. $\hat{\mu}$) is frequently used interchangeably to denote both the random variable ($\overline{X}$) as well as the realized value ($\overline{x}$). This does not create confusion for statisticians, because the meaning is usually clear from context. However, the distinction can trip up the unwary. A quick test of one's grasp of statistics is to define and contrast μ, $\overline{x}$, $\overline{X}$, and $\hat{\mu}$ (ornery professors looking for curveballs to throw at Ph. D. candidates during oral exams, please take note).

In our example, the sampling distribution of the ML estimate is particularly simple. The exact sampling distribution of $\overline{X}$ is a normal distribution with a mean of μ and a variance of σ^2/n. The independent normal model for the observations is especially convenient because the

sampling distribution of various statistics can be derived mathematically. In other models, such as the multinomial models used in categorical data analysis or the dependent normal models used in time series analysis, the sampling distributions cannot be derived exactly and instead are approximated with asymptotic results (central limit theorem, etc.) or studied with computer simulation.

A *statistical hypothesis test* is a data-driven choice between two statisical models. Consider a fixed reference Cu concentration, μ_0, that has to be maintained or attained, for instance, $\mu_0 = 48$. One position is that the reference concentration prevails in the stream, the other position is that it does not. The positions can be summarized as two statistical hypotheses: H_0: the observations arise from a normal(μ_0, σ^2) distribution, and H_1: the observations arise from a normal(μ, σ^2) distribution, where μ is not restricted to the value μ_0. In beginning statistics texts, these hypothese are often stated as H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$. A decision involves two possible errors, provided the normal distribution portion of the hypotheses is viable. First, H_0 could be true but H_1 is selected (Type I error); second, H_1 could be true but H_0 is selected (Type II error). Both errors have associated *conditional* probabilities: α, the probability of erroneously choosing H_1, given H_0 is true, and β, the probability of erroneously choosing H_0, given H_1 is true. Both of these error probabilities are set by the investigator. One probability, typically α, is set arbitrarily at some low value, for instance 0.05 or 0.01. The corresponding hypothesis assumed true, H_0, is termed the *null* hypothesis. The other probability is controlled by the design of the sample or experiment (sample size, etc.) and the choice of test statistic. The hypothesis assumed true in this case, H_1, is the *alternative* hypothesis.

Several important points about statistical hypothesis tests must be noted. First, α and β are not the probabilities of hypotheses, nor are they the unconditional probabilities of committing the associated errors. In frequentist statistics, the probability that H_0 is true is either 0 or 1 (we just do not know which), and the unconditional probability of committing a Type I error is either α or 0 (we do not know which). In the frequentist view, stating that "H_0 has a 25% chance of being true" is meaningless with regard to inference.

Second, the simpler hypothesis, that is, the statistical model that has fewer parameters and is contained within the other as a special case is usually designated as the null hypothesis, for reasons of mathematical convenience. The sampling distributions of test statistics under the null hypothesis in such situations are often easy to derive or approximate.

Third, statistical theory accords no special distinction between the null and alternative hypotheses, other than the difference by which the probabilities α and β are set. The hypotheses are just two statistical models, and the test procedure partitions the sample space (the collection of all possible samples) into two sets: the set for which the null model is selected, and the set for which the alternative is selected.

Fourth, the concordance of the statistical hypothesis with a scientific hypothesis is not a given, but is part of the craft of scientific investigation. Just because an investigator ran numbers through PROC this-or-that does not mean that the investigator has proved anything to anyone. The statistical hypothesis test can enter into scientific arguments in many different ways, and weaving the statistical results effectively into a body of scientific evidence is a difficult skill to master. Ecologists who have become gun-shy about hypothesis testing after reading a lot of

hand-wringing about the misuse of null hypotheses and significance testing will find the discussions of Underwood (1990) and Mayo (1996) more constructive.

In our stream example, the hypothesis test is constructed as follows. The likelihood function under the null hypothesis is compared to the maximized likelihood function under the alternative hypothesis. The likelihood function under the null hypothesis is Eq. 1 evaluated at $\mu = \mu_0$ ($= 48$) and $\sigma^2 = 36$. The maximized likelihood function under the alternative hypothesis is Eq. 1 evaluated at $\mu = \widehat{\mu} = \overline{x} = 50.6$ and $\sigma^2 = 36$. The likelihood ratio statistic, $L(\mu_0, \sigma^2)/L(\widehat{\mu}, \sigma^2)$, or a monotone function of the likelihood ratio such as

$$G^2 \;=\; -\,2\ln\Big[L(\mu_0, \sigma^2)/L(\widehat{\mu}, \sigma^2)\Big]$$

forms the basis of the test. High values of the test statistic G^2 favor the alternative hypothesis, while low values favor the null. The decision whether to reject the null in favor of the alternative will be based on whether the test statistic exceeds a *critical value* or cutoff point. The critical value is determined by α and the statistical sampling distribution of the test statistic.

A well-known result, first derived by S. S. Wilks (1938), provides the approximate sampling distribution of G^2 for many different statistical models. If the null hypothesis is true, then G^2 has an asymptotic chi-square distribution with 1 degree of freedom, under hypothetical repeated sampling. (The degrees of freedom in Wilks' result is the number of independent parameters in H_1 minus the number of independent parameters estimated in H_0, or $1 - 0 = 1$ in our example.) Using this result, one would reject the null hypothesis if G^2 exceeded $\chi^2_\alpha(1)$, the $100(1-\alpha)$th percentile of a chi-square(1) distribution ($\chi^2_{0.05}(1) \approx 3.843$). Because our example involves observations from the mathematically convenient normal distribution, the sampling distribution can be calculated exactly. Letting $Z = (\overline{X} - \mu_0)/\sqrt{\sigma^2/n}$, the expression for G^2 can be algebraically rearranged (using the $+\,\overline{X} - \overline{X}$ trick again; the upper case $\overline{X}$ reminds us that hypothetical repeated sampling is being considered):

$$G^2 \;=\; Z^2\,.$$

Because Z has a standard normal distribution, the chi-square result for G^2 is exact (square of a standard normal has a chi-square(1) distribution). The decision to reject can be based on the chi-square percentile, or equivalently on whether $|Z|$ exceeds $z_{\alpha/2}$, the $100(1-\alpha/2)$th percentile of the standard normal distribution ($z_{0.025} \approx 1.960$).

For the stream example, the attained value of Z is $z = (50.6 - 48)/\sqrt{36/10} \approx 1.37$. For $\alpha = 0.05$, the critical value of 1.96 is not exceeded. We conclude that the value $\mu_0 = 48$ is a plausible value for μ; there is not convincing evidence otherwise. The *P-value*, or attained significance level, is the probability that Z for a hypothetical sample would be more extreme than the attained value z, under the null model. From the normal distribution, $P[|Z| > 1.37] = P \approx 0.17$. If the test statistic has exceeded the critical value, then also P will be less than α.

Confidence intervals and hypothesis tests are two sides of the same coin. A confidence interval (CI) for μ can be defined in terms of a hypothesis test: it is the set of all values of μ_0 for which the null hypothesis H_0: $\mu = \mu_0$ would not be rejected in favor of the

alternative H_1: $\mu \neq \mu_0$. A CI can be considered a set of plausible values for μ. The sets produced under hypothetical repeated sampling would contain the true value of μ an average of $100(1-a)\%$ of the time. The form of the interval is here

$$\left(\overline{x} - z_{\alpha/2}\sqrt{\sigma^2/n}\,,\, \overline{x} + z_{\alpha/2}\sqrt{\sigma^2/n} \right).$$

The realized CI for the stream example, with $\alpha = 0.05$, is

$$\left(50.6 - 1.96\sqrt{36/10},\ 50.6 + 1.96\sqrt{36/10}\right) = (46.9,\ 54.3).$$

Note that under the frequentist interpretation of the interval, it is not correct to say that $P[46.9 < \mu < 54.3] = 1 - \alpha$. The interval either contains μ or it does not; we do not know which. The concept of a CI can be likened to a playing a game of horseshoes in which you throw the horseshoe over a wall that conceals the stake. Your long-run chance of getting a "ringer" might be 95%, but once an individual horseshoe is thrown, it is either a ringer or it is not (you just do not know which).

There are *one-sided* hypothesis tests, in which the form of the alternative hypothesis might be H_1: $\mu \geq \mu_0$ (or instead $\leq$), and associated one-sided confidence intervals (see Bain and Engelhardt 1992). The one-sided test or CI might be more appropriate for the stream example, if for instance the data are collected to provide warning as to whether an upper level μ_0 of Cu concentration has been exceeded.

σ^2 *unknown*

Realistic modeling studies must confront the problem of additional unknown parameters. Sometimes the whole model is of interest, and no particular parameters are singled out for special attention. Other times, as in the stream example, one or more parameters are the focus, and the remaining unknown parameters ("nuisance parameters") are estimated out of necessity.

The parameter σ^2 in the normal model is the perennial example in the statistics literature of a nuisance parameter. That the estimate of μ becomes more uncertain when σ^2 must also be estimated was first recognized by W. S. Gosset (Student 1908). The problem of nuisance parameters was refined by numerous mathematical/statistical investigators after the likelihood concept became widely known (Cox and Hinkley 1974 is a standard modern reference).

In frequentist statistics, a leading approach is to estimate σ^2 (or any other nuisance parameter) just like one would estimate μ. The approach has the advantage of helping subsidiary studies of the data; for instance, in a monitoring study (such as the stream example), one might have an additional interest in whether or not σ^2 has changed. In this case σ^2 is not really a nuisance, but rather an important component of the real focus of study: the model itself.

For estimation, the likelihood function (Eq. 1) is regarded as a joint function of the two unknowns, μ and σ^2. The ML estimates of μ and σ^2 are those values which jointly maximize the likelihood (Eq. 1) or log-likelihood (Eq. 2). A simple calculus exercise sets partial derivatives of $\ln L(\mu, \sigma^2)$ with respect to μ and σ^2 simultaneously equal to zero. The resulting ML estimates are:

$$\widehat{\mu} \;=\; \overline{x}\,,$$

$$\widehat{\sigma}^2 \;=\; \frac{(n-1)}{n}\, s^2\,.$$

Note that the ML estimate of σ^2 is not the sample variance ($n-1$ in denominator). An estimate that adjusts for a small-sample bias is

$$\widetilde{\sigma}^{\,2} \;=\; s^2\,.$$

The ML estimate of σ^2, however, has smaller mean-squared error in small samples; $\widehat{\sigma}^2$ and $\widetilde{\sigma}^{\,2}$ are virtually identical in large samples.

Hypothesis tests and confidence intervals for μ again revolve around the likelihood ratio statistic. With additional unknown parameters in the model, the statistic compares the the likelihood function maximized (over the remaining parameters) under the null hypothesis, H_0: $\mu = \mu_0$, with the likelihood maximized (over all the parameters including μ) under the alternative hypothesis H_1: $\mu \neq \mu_0$. When $\mu = \mu_0$, the value of σ^2 that maximizes ln $L(\mu_0, \sigma^2)$ (Eq. 2) is

$$\widehat{\sigma}_0{}^2 \;=\; \frac{(n-1)}{n}\, s^2 \;+\; (\mu_0 - \overline{x})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_0)^2\,.$$

The (log-) likelihood ratio statistic is

$$G^2 \;=\; -\,2\ln\left[\frac{L(\mu_0, \widehat{\sigma}_0{}^2)}{L(\widehat{\mu}, \widehat{\sigma}^2)}\right],$$

where in the brackets is the ratio of the null and alternative likelihoods, evaluated at the ML estimates. With some algebraic rearrangement (the $+\,\overline{x} - \overline{x}$ trick again), G^2 becomes

$$G^2 \;=\; -\,n\ln\left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0{}^2}\right) \;=\; -\,n\ln\left[\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2}\right]$$

$$= -\,n\ln\left(\frac{1}{1 + \frac{t^2}{n-1}}\right),$$

where

$$t \;=\; \frac{\overline{x} - \mu_0}{\sqrt{s^2/n}}$$

is recognized as Student's t-statistic. The hypothesis test can be based on the asymptotic chi-square(1) sampling distribution of G^2, or better yet, on the known exact distribution of $T = (\overline{X} - \mu_0)/\sqrt{S^2/n}$. One rejects $H_0: \mu = \mu_0$ in favor of $H_1: \mu \neq \mu_0$ if $|T|$ exceeds $t_{\alpha/2,\, n-1}$, the 100[1 − (α/2)]th percentile of the Student's t-distribution with $n - 1$ degrees of freedom.

For the stream example with $\mu_0 = 48$, the attained value of T is $t = (50.6 - 48)/\sqrt{25/10} \approx 1.64$. For $\alpha = 0.05$, the critical value of $t_{0.025,\, 9} \approx 2.262$ is not exceeded by $|t|$, and we conclude that the value $\mu_0 = 48$ is a plausible value for μ. The P-value for the test is obtained from Student's t-distribution (9 degrees of freedom): $P[|T| > 1.64] = P \approx 0.14$.

Confidence intervals, as before, can be defined by inverting the hypothesis test. The values of μ_0 for which H_0 is not rejected, that is for which $|T| \leq t_{\alpha/2,\, n-1}$, make up the interval

$$\left(\overline{x} - t_{\alpha/2,\, n-1}\sqrt{s^2/n}\,,\; \overline{x} + t_{\alpha/2,\, n-1}\sqrt{s^2/n} \right).$$

This constitutes a $100(1 - \alpha)\%$ CI for μ. The interval also represents a *profile likelihood* CI. For a range of fixed values of μ_0, $L(\mu_0, \sigma^2)$ is maximized (over σ^2 values) and compared to $L(\widehat{\mu}, \widehat{\sigma}^2)$ (the maximized likelihood under the model H_1). The set of μ_0 values for which $G^2 \leq k$, where k is some fixed constant, is a profile likelihood CI. In the above interval, $k = n \ln \left[1 + \frac{\left(t_{\alpha/2,\, n-1}\right)^2}{n-1} \right]$, the critical value of the likelihood ratio test using the exact Student's t-distribution. In non-normal models, k is typically a percentile of the chi-square distribution used to approximate the sampling distribution of G^2. Frequently for such models, repeated numerical maximizations are necessary for calculating profile likelihood intervals.

The stream example gives a 95% CI of

$$\left(50.6 - 2.262\sqrt{25/10},\; 50.6 + 2.262\sqrt{25/10}\,\right) \;=\; (47.0,\, 54.2)\,.$$

This CI represents a range of plausible values for μ, taking into account the uncertainty of estimation of σ^2.

Much of standard introductory statistics, in the form of t-tests, tests of independence in contingency tables, analysis of variance, and regression, can be understood in the context of the above concepts. In particular, normal linear models (analysis of variance and regression) are formed by allowing μ to be reparameterized as

$$\mu \;=\; \beta_0 \;+\; \beta_1 r_1 \;+\; \beta_2 r_2 \;+\; \ldots \;+\; \beta_m r_m\,,$$

where β_0, β_1, ..., β_m are unknown parameters and r_1, r_2, ..., r_m are values of covariates (indicator variables or predictor variables).

Model evaluation

Once the point estimates are calculated, tests are performed, and confidence intervals are reported, the job is not done. The estimates and tests have valid sample space statistical properties only if the model is a reasonable approximation of how the original data arose. *Diagnostics* are routine checks of model adequacy. Diagnostics include examining residuals (in this case, $x_i - \widehat{\mu}$) for approximate normality via normal quantile-quantile plots or tests, tests for outliers or influential values, and graphical plotting of model and data. The model implicit in the statistical analysis is to be questioned, and if found wanting, some other model might be necessary.

Such model checking, it must be noted, involves sample space properties of the model. If the correct model is being used, observations of the process are expected to be in *control*, that is, within the usual boundaries of model-predicted variability. A process out-of-control is indicated by wayward observations and calls for further investigation. This is a standard principle of *quality control*, which involves the systematic and routine use of statistical models to monitor variability and is used in virtually all modern manufacturing processes (Vardeman 1994).

WHAT IS BAYESIANISM?

"What he and I are arguing about is different interpretations of data."
—Duane Gish, in an evolution/creation debate

The concepts of frequentism revolve around hypothetical repetitions of a random process. The probabilities in a frequentist problem are probabilities on a sample space. The quantity α, for instance, is the probability that the sample will land in a particular region of sample space, given that a particular model describes the process.

In Bayesian statistics, sample space probabilities are not used. Instead, probability has a different meaning. Probability in Bayesian statistics is an investigator's personal measure of the *degree of belief* about the value of an unknown quantity such as a parameter.

Let us again turn to the example problem. We have a sample of 10 observations of Cu concentration in a stream. We assume that these observations can be modeled as if they arose independently from a normal distribution with a mean of μ and a variance of σ^2. Again, we treat separately the cases of σ^2 known and σ^2 unknown.

σ^2 known

There is only one unknown parameter, μ. The Bayesian formulates his/her beliefs about the value of μ into a *prior probability distribution*. The prior distribution has pdf denoted by

$g(\mu)$ and cdf given by

$$G(\mu) = \int_{-\infty}^{\mu} g(v)dv \, .$$

The form of $g(\mu)$ must be specified completely by the investigator. There are various ways to do this. One way is to "elicit" such a distribution by determining the odds the investigator would give for betting on various values of μ.

The subsequent formulas work out algebraically if we assume that the form of $g(\mu)$ is a normal pdf with a mean of θ and a variance of τ^2, with the values of θ and τ^2 to be elicited. However, more complicated distributional forms nowadays are possible to implement in practice. Skewed gamma-type distributions or curves fitted to the investigator's odds declarations can be used.

Our investigator in this example works for the mining company upstream. This investigator would give one-to-three odds that the Cu concentration is below 18.65, and three-to-one odds that the Cu concentration is below 21.35. If the 25th and the 75th percentiles of a normal distribution are set at 18.65 and 21.35 respectively, then solving

$$G(18.65) = 0.25$$

$$G(21.35) = 0.75$$

simultaneously gives

$$\theta \approx 20,$$

$$\tau^2 \approx 4 \, .$$

It is worth pausing a moment to reflect on the prior. It is not claimed that μ is a random variable. Indeed, μ is a fixed quantity, and the objective is to estimate its value. Rather, μ is an *unknown* quantity, and *personal beliefs about μ can be represented as if they follow the laws of probability.* This is because the odds that the investigator would give for the value of μ increase smoothly from 0 for values of $\mu < -\infty$, to $+\infty$ for values of $\mu < +\infty$. Such increase and range are the precise properties of a cdf written in terms of odds: $G(\mu)/[1 - G(\mu)]$.

Data, in the Bayesian view, modify beliefs. The data enter the inference through the likelihood function. The likelihood function is as central to Bayesian inference as it is to frequentist inference. However, its interpretation is different under the two outlooks.

In Bayesian statistics, the likelihood arises as a conditional probability model. It is the joint pdf of the process, evaluated at the data, *given values of the unknown parameters*. In other words, the set of beliefs about all possible values of μ and all possible outcomes of the data-production process are contained in a joint pdf, say $h(x_1, x_2, \ldots, x_n, \mu)$. The likelihood function (Eq. 1) is the conditional pdf of $x_1, x_2, \ldots, x_n$ given μ:

$$L(\mu, \sigma^2) \;=\; \left(\sigma^2 2\pi\right)^{-n/2} \exp\left\{ -\left(2\sigma^2\right)^{-1} \left[(n-1)s^2 \;+ n\,(\mu - \overline{x})^2\right]\right\}$$

$$= \; h(x_1, x_2, \ldots, x_n \big| \mu)\,.$$

The frequentist simply regards the likelihood as a function of possible values of μ, with no underlying probability attached to the μ values.

What is sought in Bayesian statistics is the probability distribution of beliefs after such beliefs have been modified by data. This distribution is known as the *posterior* distribution and is the distribution of μ given the data, $x_1, x_2, \ldots, x_n$. Bayes' theorem in probability is a mathematical result about joint and conditional probability distributions that is not in dispute between frequentists and Bayesians. In the present context, the theorem is used to write the pdf of μ given $x_1, x_2, \ldots, x_n$, denoted $g(\mu \big| x_1, x_2, \ldots, x_n)$, in terms of the the likelihood function and the prior distribution:

$$g(\mu \big| x_1, x_2, \ldots, x_n) \;=\; C h(x_1, x_2, \ldots, x_n \big| \mu) g(\mu)\,.$$

The quantity C is a *normalization constant* that causes the area under $g(\mu \big| x_1, x_2, \ldots, x_n)$ to be equal to 1. It is:

$$C \;=\; \frac{1}{\int_{-\infty}^{+\infty} h(x_1, x_2, \ldots, x_n \big| \mu) g(\mu)\, d\mu}\,.$$

The calculation of C is the mathematical and computational crux of Bayesian methods. Obtaining C is algebraically straightforward for the forms of the prior and the likelihood selected in our stream example. The quantity μ appears quadratically in the exponential function in the product $h(x_1, x_2, \ldots, x_n \big| \mu) g(\mu)$, and so C is related to the integral of a normal distribution. The end result is that the posterior pdf is that of a normal distribution:

$$g(\mu \big| x_1, x_2, \ldots, x_n) \;=\; \left({\tau_1}^2 2\pi\right)^{-\frac{1}{2}} \exp\left[-(\mu - \theta_1)^2 / \left(2{\tau_1}^2\right)\right],$$

where the mean θ_1 is

$$\theta_1 \;=\; \left(\frac{\sigma^2}{\sigma^2 + \tau^2 n}\right)\theta \;+\; \left(\frac{\tau^2 n}{\sigma^2 + \tau^2 n}\right)\overline{x}\,,$$

and the variance ${\tau_1}^2$ is

$${\tau_1}^2 \;=\; \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2 n}\,.$$

The mean θ_1 is a weighted combination of the prior mean, θ, and the sample mean of the data, $\overline{x}$. As the sample size increases, the weight on the prior mean decreases, approaching zero in the limit as $n \to \infty$.

The point estimate of μ in Bayesian statistics is usually taken to be the expected value of the posterior distribution: θ_1. This estimate can be regarded as a *prediction* of the value of μ. The posterior distribution, like the prior, represents degree of belief. The prior prediction was θ, and the posterior prediction θ_1 quantifies how the prior prediction has been modified by the advent of the data. In our stream example, the ten data points changed the prior prediction of $\theta = 20$ into the posterior prediction of $\theta_1 \approx 36.1$. The variance of the posterior is ${\tau_1}^2 \approx 1.89$.

One should note that the Bayesian point estimate of μ is *biased* in the frequentist sense. If hypothetical repetitions of the sampling process are imagined (for the same Bayesian with the same prior), the frequency distribution of the Bayesian's estimates would be off-center from μ. If we denote by Θ_1 the sample-space random version of the point estimate θ_1, then the expected value of Θ_1 over the sample space is

$$\mathrm{E}(\Theta_1) \;=\; \left(\frac{\sigma^2}{\sigma^2 + \tau^2 n}\right)\theta \;+\; \left(\frac{\tau^2 n}{\sigma^2 + \tau^2 n}\right)\mathrm{E}(\overline{X})$$

$$= \;\mu \;+\; \left(\frac{\sigma^2}{\sigma^2 + \tau^2 n}\right)(\theta - \mu)\,.$$

The amount of bias is seen to be the difference $\theta - \mu$ (bias in the prior) multiplied by the Bayesian weight. In the stream example, if the null hypothesis ($\mu = 48$) were true, the amount of bias in the Bayesian estimate is about -13.3; this Bayesian's long-run frequency distribution of estimates would be centered at a distance more than twice the standard deviation ($\sigma = 6$) below the true value of μ.

The posterior distribution also yields to the Bayesian information about the uncertainty with which to regard the prediction. One way to summarize this uncertainty is the Bayesian belief interval, formed by taking an interval containing $100(1-\alpha)$% of the probability in the posterior density. The smallest such interval is the *highest probability region* (HPR). The HPR is analogous to the confidence interval of frequentist statistics, but has a much different interpretation. The Bayesian asserts that there is a 95% chance that μ is within a given 95% HPR, because probability represents belief on a parameter space (all possible values of μ). The frequentist cannot assert that there is a 95% chance that μ is within a given 95% CI, because probability to a frequentist represents long-run frequency on a sample space (all possible outcomes of the sample). With our normal model, the HPR region is the interval centered at the posterior mean, θ_1, containing $100(1-\alpha)$% of the area under the posterior density:

$$\left(\theta_1 - z_{\alpha/2}\sqrt{{\tau_1}^2}\,,\, \theta_1 + z_{\alpha/2}\sqrt{{\tau_1}^2}\right).$$

In the stream example, the 95% HPR is $(33.4, 38.8)$, which is quite different from the 95% confidence interval (46.9, 54.3) obtained under the frequentist approach. However, as the

sample size becomes large and the data swamp the prior beliefs, the HPR in this normal-based example converges rapidly to the confidence interval. In other words, the Bayesian and the frequentist will report essentially the same interval estimate for μ if good data are available. While this asymptotic behavior of HPRs is typical for standard statistical models, it is somehow not a comforting point of agreement for Bayesians and frequentists, in that the interpretation of the two intervals is so different. Also, for some models and circumstances, the rate of convergence of the Bayesian HPR to the frequentist CI is alarmingly slow (see σ^2 *unknown*, below).

A key aspect of Bayesianism is adherence to the *likelihood principle*. The principle states that sample space probabilities are irrelevant to inferences about unknown parameters. The data only influence the inferences through the likelihood function. This principle is embodied in the posterior density, $g(\mu|x_1, x_2, \ldots, x_n)$. All inferences about μ are contained in the posterior density and are phrased in terms of probabilities on parameter space. Only the data actually observed appear in the posterior (via the likelihood function); no hypothetical data, such as a critical value for $\overline{x}$, or probabilities of hypothetical data, such as P-values or Type I & II error probabilities, are considered in the conclusions about μ.

Bayesians are adamant on this point (Lindley 1990, Berger and Berry 1988). Type I & II error probabilities and P-values are probabilities of "data that didn't happen," and Bayesians question what relevance such quantities could possibly have for conclusions about a parameter.

The use of sample space probabilities in frequentist statistics has surprising, and to Bayesians, undesirable consequences. Foremost is the dependence of the statistical conclusions on the *stopping rule* of the experiment. For instance, were the stream samples drawn sequentially, one by one, until some threshold high or low value of $\overline{x}$ was attained? Or, were simply ten samples drawn? Or, did the investigator actually draw 11 samples, but drop one jar accidentally?

σ^2 *unknown*

Bayesians claim that the treatment of nuisance parameters within the Bayesian framework is one of the key advantages of their approach. Let us examine how this claim operates in practice.

With σ^2 unknown, the concept behind the Bayesian analysis is straightforward. The posterior distribution for μ, represented by the pdf $g(\mu|x_1, x_2, \ldots, x_n)$, is still sought. First, though, beliefs about μ and σ^2 must be summarized in a joint prior distribution for μ and σ^2. So-represented, the beliefs are entered into the mix with the likelihood function (Eq. 1), and the posterior distribution for μ is then obtained (at least in principle) with Bayes' theorem.

The joint prior pdf for μ and σ^2, denoted $g(\mu, \sigma^2)$, is that of a bivariate continuous distribution. The distribution would presumably be defined for positive real values of σ^2, and real (or positive real) values of μ. A joint distribution in general would contain some correlation between μ and σ^2. Rarely, however, can any dependence of beliefs about μ on those about σ^2 be acknowledged or elicited. Consequently, the form often proposed for the joint pdf is a

product of univariate prior pdfs for μ and σ^2:

$$g(\mu, \sigma^2) \;=\; g_1(\mu) g_2(\sigma^2)\,.$$

Here $g_1(\mu)$ is a pdf for μ (such as the normal pdf in the σ^2-known case above), and $g_2(\sigma^2)$ is a pdf for σ^2. The form of $g_2(\sigma^2)$ selected by the investigator could be a gamma, reciprocal gamma, lognormal, or other distribution on the positive real line. The product form of $g(\mu, \sigma^2)$ assumes (or implies) that the prior "information" about μ is independent of that of σ^2.

With the elicited joint prior in hand, the analysis proceeds via Bayes' theorem. The joint posterior pdf for μ and σ^2 is proportional to the product of the prior pdf and the likelihood function:

$$g(\mu, \sigma^2 \big| x_1, x_2, ..., x_n) \;=\; C_1 h(x_1, x_2, ..., x_n \big| \mu, \sigma^2) g_1(\mu) g_2(\sigma^2)\,.$$

The likelihood function is again written as $h(x_1, x_2, ..., x_n \big| \mu, \sigma^2)$ to emphasize its role as a conditional pdf. The constant C_1 is the normalization constant given by

$$C_1 \;=\; \frac{1}{\int\int h(x_1, x_2, ..., x_n \big| \mu, \sigma^2) g_1(\mu) g_2(\sigma^2)\, d\mu\, d\sigma^2}\,,$$

where the integrals are over the ranges of μ and σ^2 in the prior pdfs. Some remarks about the daunting process of obtaining C_1 are given below. In principle, the joint posterior pdf $g(\mu, \sigma^2 \big| x_1, x_2, ..., x_n)$ contains all the beliefs about μ and σ^2, updated by the data $x_1, x_2, ..., x_n$. Moreover, the nuisance parameter σ^2 is vanquished by integrating it out of the joint posterior to get the posterior marginal distribution for μ:

$$g(\mu \big| x_1, x_2, ..., x_n) \;=\; \int_0^\infty g(\mu, \sigma^2 \big| x_1, x_2, ..., x_n)\, d\sigma^2\,.$$

This posterior pdf for μ reflects all beliefs about μ after the advent of the data. The pdf could be used, for instance, to obtain an HPR for μ, just as was done in the σ^2-known case above.

The technical difficulties with the analysis reside in evaluating the multiple integrals for C_1 and in integrating out σ^2 to get the marginal posterior for μ. For nearly all forms of prior distributions $g_1(\mu)$ and $g_2(\sigma^2)$, the integrals must be performed numerically. Up until the middle 1980s, the lack of symbolic results for the integrals were the death knell for Baysianism, because methods for reliably evaluating multi-dimensional integrals were poorly developed. However, clever simulation methods were devised for these integrals; the methods exploit the fact that the integrals are essentially expected values of functions with respect to the prior distributions. Papers on Bayesian analyses in the statistics literature subsequently exploded in number, starting in the late 1980s. The simulation methods have been for a decade a part of the hidden culture of statisticians, described tersely or implicitly in dense mathematical terms in the statistics literature, but are now receiving excellent expositions for broader scientific audiences (for instance, Robert and Casella 1999). Investigators must be warned, however, that the numerical methods at present involve heavy computer programming efforts, post-calculus

statistics knowledge, and sometimes days of computer time; the methods are not ready yet for routine use by busy laboratory or field scientists.

What if the investigator does not really have, or is unwilling to admit, any prior beliefs about σ^2? Bayesian writers have proposed "uninformative priors" for such situations. Use of these priors has also been advocated for situations in which investigators disagree about the prior information and require a relatively "neutral" prior for mediation (Lee 1989). However, there are different approaches to specifying neutral priors. One is the *maximum entropy* approach (Jaynes 1968). The investigator in the maximum entropy approach specifies only numerical summaries of the prior distribution, such as the mean, or the mean and variance both. The prior is then the distribution that maximizes the "entropy content" (expected value of $-\ln g_2(\sigma^2)$) of the prior while retaining the numerical summaries. If the mean of the prior for σ^2 is fixed at ϕ, for example (and the range is taken to be the positive real line), the maximum entropy criterion yields an exponential distribution for σ^2 with pdf $g_2(\sigma^2) = (1/\phi)\exp(-\sigma^2/\phi)$. Another approach is that of the *uniform prior*: the beliefs about σ^2 are taken to have a uniform distribution. This type of prior is sometimes called an *improper prior* because it is not integrable over the entire range of the parameter (here, the positive real line). Actually, the uniform distribution for σ^2 is taken to range properly from 0 to, say, some large unspecified number γ. The prior pdf (a constant, $1/\gamma$) is then integrable, and the value of γ, if large, turns out to affect the calculations about μ only negligibly.

A conceptual problem with uninformative priors is that ignorance about σ^2, expressed in an uninformative prior distribution for σ^2, does not translate into ignorance about a function of σ^2, say $\ln\sigma^2$. For instance, if σ^2 has a uniform distribution on the interval from 0 to γ, then the distribution of $\ln\sigma^2$ is non-uniform. A uniform prior distribution for σ^2 leads to a different posterior distribution for μ than when a uniform prior for $\ln(\sigma^2)$ is used. This disparity has motivated some Bayesians to investigate how to choose the scale upon which their ignorance is to be expressed. Textbook discussions of such investigations gravitate to scales which allow convenient algebra, i.e., scales for which the problematic integrals noted above can be evaluated symbolically (e.g. Lee 1989).

So that some numerical results might be displayed for the stream example without having to refer to a workstation, let us employ such a scale for σ^2. Suppose the prior distribution for $\ln(\sigma^2)$ is taken to be a uniform distribution on some large, unspecified interval of the real line. Then, from the transformation rule for distributions (Rice 1995), the prior distribution for σ^2 has a pdf of the form $g_2(\sigma^2) = C_2/\sigma^2$, where C_2 is a constant. This prior is improper on the entire positive real line, but again it will be thought of as ranging from 0 to some large but unspecified upper value. The joint prior distribution for μ and σ^2 becomes, assuming independence of beliefs about the two parameters, the product of marginal prior pdfs:

$$g(\mu, \sigma^2) = C_2\left(\sigma^2\sqrt{\tau^2 2\pi}\right)^{-1}\exp\left[-(\mu-\theta)^2/(2\tau^2)\right].$$

Substituting this joint prior into the expression for the posterior distribution for μ and σ^2 above, and using the normal likelihood function (Eq. 1) for $h(x_1, x_2, ..., x_n|\mu, \sigma^2)$, one obtains

$$g(\mu, \sigma^2 | x_1, x_2, ..., x_n) \;=\; C_1 \left(\tau^2\right)^{-\frac{1}{2}} (2\pi)^{-\frac{n+1}{2}} \left(\sigma^2\right)^{-\left(\frac{n}{2}+1\right)} \exp\{ -\frac{1}{2\sigma^2}(n-1)s^2$$

$$-\frac{1}{2\sigma^2} n(\mu - \overline{x})^2 - \frac{1}{2\tau^2}(\mu - \theta)^2 \} .$$

This posterior pdf for μ and σ^2 is a dome-shaped function reflecting the joint beliefs about the two parameters after the advent of the data. The nuisance parameter is now eliminated in an act which is to Bayesians conceptually as well as algebraically symbolic. The terms in the posterior involving σ^2 are (thanks to our selection of prior) in the form $\left(\sigma^2\right)^{-a} \exp\left(-b/\sigma^2\right)$. The form is like a reciprocal gamma pdf and yields $b^{-(a-1)}\Gamma(a-1)$ when integrated over the positive real line. Thus:

$$g(\mu | x_1, x_2, ..., x_n) \;=\; \int_0^\infty g(\mu, \sigma^2 | x_1, x_2, ..., x_n)\, d\sigma^2$$

$$= C_3 \left[\frac{(n-1)s^2}{2} \;+\; \frac{n(\mu - \overline{x})^2}{2} \right]^{-n/2} \exp\left[-\frac{(\mu - \theta)^2}{2\tau^2} \right]$$

$$= C_3 \left[1 \;+\; \frac{t^2}{(n-1)} \right]^{-n/2} \exp\left[-\frac{(\mu - \theta)^2}{2\tau^2} \right] .$$

Here $t = \sqrt{n}(\overline{x} - \mu)/s$ is the t-statistic that would be used by frequentists to test a particular value of μ as a null hypothesis. The posterior distribution for μ is seen to be the prior normal pdf for μ weighted by the pdf of a Student's t-distribution. Rather awkwardly, the normalization constant C_3 cannot be evaluated symbolically, and so to obtain a point estimate or HPR the workstation must be booted up even for this simple illustrative example. Fortunately, the numerical integration for one dimension is straightforward.

The 95% HPR resulting for the stream example (using, as will be recalled, $\overline{x} = 50.6$, $s^2 = 25.0$, $n = 10$, $\theta = 20.0$, $\tau^2 = 4.0$) is approximately (17.3, 25.3). Recall that the frequentist 95% confidence interval based on the Student's t-distribution was (47.0, 54.2). In the previous case in which σ^2 was taken as a known constant ($\sigma^2 = 36$), the HPR (33.4, 38.8) was considerably closer to the corresponding frequentist confidence interval (46.9, 54.3). Apparently, when σ^2 is unknown the preponderance of weight remains in this example on the prior; the HPR is remarkably insensitive to the data. While the frequentist results take at face value the estimate, s^2, of σ^2, the Bayesian results in the face of lack of knowledge about σ^2 contain an inherent preference for beliefs about μ over data. Indeed, one can ask just how much evidence is necessary for the Bayesian here to start noticing the data. What if the values of $\overline{x}$ and s^2 from the stream had resulted from larger sample sizes? Suppose the value of n is increased in the posterior pdf for μ above, while keeping all the quantities fixed at the same values. At $n = 40$, prior beliefs are still heavily weighted: the HPR is (21.8, 30.0). A strangely

sudden change of heart occurs between 60 and 65 observations. At $n = 60$, the posterior pdf for μ has developed a prominent "shoulder" near $\mu = 50$, and the upper end of the HPR has started to reach upward; the HPR is (26.1, 40.1). The frequentist 95% confidence interval for $n = 60$ is by contrast only two units wide: (49.3, 51.9). At $n = 63$, the posterior pdf for μ is bimodal, one peak influenced by the prior, and one peak influenced by the data; the HPR is (28.8, 47.1). By $n = 65$, the data-peak has grown taller than the prior peak, and the HPR is (31.2, 48.3). By $n = 70$, the laggard lower end of the HPR has finally entered the 40's; the HPR is (42.7, 49.0).

If $\mu = 48$ were cause for alarm, the frequentist scientist would have detected this state of affairs with as few as 10 observations. It would take at least 65 observations before our Bayesian, to whom beliefs are evidence on equal footing with data, would sound a Cu pollution warning.

DISCUSSION

Beliefs

It should be evident from the above example that Bayesian and frequentist statistics arise from different views about science. In Bayesian statistics, beliefs are the currency traded among investigators. Beliefs are evidence. Data are used to modify beliefs.

"Beliefs" is not necessarily a four-letter word. In Bayesian statistics, nothing precludes the prior density being "rationally" constructed based on common sense information. Indeed, investigators frequently encounter situations in which a parameter is not completely unknown. In our stream example, it is known for a fact that the mean Cu concentration μ is not a negative quantity. Is it not possible to account for such knowledge in the analysis?

Frequentists account for such information by building more realistic statistical models. The normal distribution is a mathematical approximation. Real Cu concentrations cannot be negative. In the stream example, if the normal approximation is adequate, negative Cu concentrations are wildly improbable, and a negative μ value would result in an extremely bad model. However, a different distribution model might be required, if, for instance, some concentrations are bunched near zero. Even a new, more realistic model is always subject to questioning, via model evaluation procedures.

Frequentists also use prior information in designing surveys and experiments. For instance, the optimal allocation allocation of samples in a stratified sampling scheme depends on knowing the variances within each stratum (Scheaffer et al. 1996). In addition, selecting a sample size in experimental design depends on the desired power,
which in turn depends on the effect size and the variance (Ott 1993). The investigator must have some prior information about the effect size and variance for the design to achieve the desired power.

To the frequentist, though, such prior information is regarded as *suspect*. This is a major departure point from the Bayesians. Frequentists build on prior information tentatively, using sample space variability properties constantly for checking the reliability of the knowledge. Prior information is placed in the likelihood function itself, and is thereby vulnerable to empirical challenge.

Bayesians will cry foul at my rough handling of the Bayesian analysis in the stream example. The selection of the prior mean of $\theta = 20$ by the mining company scientist seems lopsided and biased, cynically calculated to come to a pre-defined conclusion. The scientist is supposed to formulate the prior distribution based on real costs and consequences of being wrong. What if the scientist were forced to deliver on the bet implied by the prior?

It must be remembered, however, that the scientist is employed by the mining company.

First, the real cost to the scientist comes not from being far from the truth, but rather from defeat. The two are not necessarily concordant. The scientist has a real financial stake in defending the mining company's view. If this scientist is not willing, some other scientist will gladly step in and cook the numbers.

Second, the scientist can be wrong, or even deceived. It is quite possible that the scientist's beliefs were genuine. During previous monitoring of the stream, say, Cu concentrations might indeed have hovered around 20 μg l^{-1}, and the scientist had no reason to believe that today's samples would be different. In this hypothetical scenario, the company had a pollution event, and failed to inform their monitoring group.

Either way, the scientist's beliefs do nothing but contaminate the data analysis. They add no legitimate information to the estimate of Cu concentration.

In fact, for scientists in general there is often a conflict of interest between a scientist's beliefs and the truth. In an ideal world, the scientist who is most successful in discovering truth will be the most successful in building a scientific career. While this circumstance is fortunately common, the real world of scientific careers admits additional complexities (Sindermann 1982). Scientists gain reputations for being advocates of certain theories. Laboratories and research programs grow from mining particular techniques or points of view. A scientist's career is measured in the form of socially warranted visibility: jobs, papers, research grants, citations, and seminar invitations. Young graduate students know one career syndrome well: stubborn adherence of older scientists to old-fashioned explanations and quick dismissal by such scientists of newer ideas before even understanding them. Senior scientists know another career syndrome well: rapid study and advocacy by younger scientists of fashionable new hypotheses that contradict established doctrine and are beyond the frontiers of available data. Everyone in science knows of investigators that took wrong turns toward untenable hypotheses and then spent whole careers defending the hypotheses with contrived arguments. To an individual scientist with a career to build, maintain, and defend, victory, rather than truth, is often the objective.

Scientific method

Is science just a postmodern “way of knowing” after all? At the level of the individual scientist, it would certainly seem so, given all the explicit and implicit social pressures. Science in the postmodern view is a belief system, and scientists achieve success only by participating in a socially warranted system of thought and action, which changes from place to place and year to year.

Bayesianism, through the incorporation of personal beliefs into statistical analyses, accepts the postmodern view of science. A scientist's acceptance or rejection of a hypothesis is a decision made in light of beliefs influenced by costs or utilities. To the Bayesian, science is

improved by explicitly stating, organizing, and acting on beliefs. A scientist summarizes his/her prior beliefs into a probability distribution and modifies those beliefs in a controlled and systematic way with data. Observers are free to quibble with the scientist's prior, or use their own priors and come to their own conclusions. Consensus of beliefs will supposedly emerge as data become more available and priors become diluted. However, the process for this emergence is not clear, for, human nature being what it is, priors will inevitably become more opinionated in the face of growing data. Fundamentally, at the heart of it all, the interpretation of results is in terms of beliefs. In Bayesianism, beliefs are sanctioned, not repudiated.

Modern science, though, has been wildly successful despite the imperfect humans that make up the ranks of scientists, and, incidentally, despite the almost complete absence of Bayesianism in day-to-day scientific life. The postmodern claim that science is socially constructed reality is an intellectual fraud (Sokal and Bricmont 1998). Hydrogen atoms, and the speed of light, are the same in India, Alaska, and the Andromeda galaxy. True, scientists, and groups of scientists, often come to the wrong conclusions. It is the *process* that is responsible for the enormous gains in understanding we have attained, in ecology and in other disciplines. Our understanding does not just jump from one fashionable paradigm to another; it *improves*. Science is like a river that flows mostly forward, but with slow pools and backcurrents here and there. It is the collective process of empirical investigation, involving weeding out of untenable notions and careful checking of working hypotheses, that makes progress possible. The invisible empirical hand of Galileo, the Adam Smith of science, promotes the emergence of reliable knowledge.

Bayesians and postmoderns alike miss the fundamental idea of science. Science is not about beliefs; science is about skepticism.

Science is not about prediction, estimation, making decisions, data collection, or data interpretation. Scientists engage in these activities, but these activities do not constitute science. Science, rather, is about constructing convincing explanations and acquiring reliable knowledge. "Convincing" means a reasoned skeptic is forced, by logic and evidence, to accept the explanation as, at least, a serious contender for the true explanation. "Reliable" means that others can reproduce the results and rely on them for building further explanations. Scientific arguments are aimed at reasoned skeptics. "Reasoned" means open to acknowledging evidence that might contradict prior points of view.

The *scientific method* is a series of logical devices for eliminating or reducing points of reasoned skepticism. One premise of the scientific method is that human judgment is inherently flawed. This is because reasoned skeptics might validly argue in any situation that a scientist's personal beliefs are suspect. Successful scientists seek to counter that criticism by adopting investigative methods that eliminate conscious or unconscious biases. Frequentist analyses are an important tool in the scientific method.

Frequentism accepts only a portion of the postmodern critique. To the frequentist, the actions and behaviors of individual investigators are indeed mired in beliefs. However, to the frequentist, the methods of statistical analysis are set up to discount those beliefs as much as possible. The assumption that a scientist's judgments are not to be trusted has a long history in frequentist statistics, and is expressed in the concepts of design-based sampling, replication,

randomization, experimental design, unbiased estimation, model diagnostics, and explicit stopping rules.

Frequentist statistics adheres to the principles of the scientific method. Experimental subjects are selected at random. Observations are sampled at random. Variability of the process under study is carefully controlled and modeled, so that future investigators can replicate and check the work. In frequentist hypothesis testing, the skeptic's null hypothesis is *assumed* to be true, but the unlike the Bayesian's prior, the assumption is just an argumentative device. The assumption is then found to be tenable or untenable under the data. The statistical models used by the investigator are suspect and must have demonstrated reliability and usefulness for future investigators. By the continual modeling of and referral to sample space variability of a data production process, frequentism can not only show that some hypotheses are untenable in a classic Popperian way, but can also establish that other hypotheses are operationally reliable and can serve as the bases for future studies (Mayo 1996).

Evidence

Bayesians claim that scientists long for numerical measures of evidence. If only one could attach, in some reliable way, a number to a hypothesis, indicating the relative weight of evidence for that hypothesis as opposed to others, then scientific conclusions would be clearer and more helpful to policy decisions. Why must we avoid doing what seems natural, that is, stating that the chance hypothesis A is true is Q%?

The answer is that the number is scientifically meaningless, and the price is too high. In Bayesian analyses, the evidentiary number cannot exist except in the personal belief system of the investigator. Neither priors nor likelihood functions can be empirically challenged in the Bayesian scheme, and so personal beliefs are always present to some degree in conclusions. With the postmodern foot in the door, the way is opened for limitless political pressure to influence the weight of evidence. Bayesian statistics might seem like a shot in the arm for a stalled science, but Bayesian science unfortunately fails to convince.

The evidentiary number in Bayesian statistics is likelihood, modified by beliefs. Is it possible to eliminate the belief considerations, while retaining likelihood? Various investigators through the years have proposed statistical approaches which accept the likelihood principle but reject the use of priors (Edwards 1972, Royall 1997). The relative weight of evidence for hypothesis A over hypothesis B is determined by comparing their likelihoods under these schemes.

The likelihood principle eliminates consideration of any sample space events, other than the actual data outcome, as evidence. But likelihood has little absolute meaning by itself, without appeal to sample space properties. One cannot determine from the statement, "$\ln L_A = -43.7$", whether A is a viable model or not. One cannot determine from the statement, "$\ln L_A - \ln L_B = 5.8$", whether model A is unquestionably better than model B. The viability of a model depends on a variety of things, for instance, on whether it *fits*. A difference in log-likelihoods as big as 5.8 might easily be within the range of variability expected by chance. Without analyses based on hypothetical sample space events, these possibilities cannot be addressed.

Also, the likelihood principle eliminates consideration of stopping rules. Whether the sample size was sequentially determined or fixed, the evidence is the same under the likelihood principle. Unfortunately, we cannot then determine whether or not the investigator's results are unusual under a particular experimental protocol, and consequently we cannot question the likelihood upon which the investigator's conclusions are based. The stopping rule dependence exists because we do not trust the scientist: we insist upon the option of repeating the study to as close a degree as possible.

Finally, the current likelihood-only analyses are developed only for simple hypotheses, that is, for statistical models with no estimated parameters (Royall 1997). Nuisance parameters are simply not treated (but it is a topic under active study and new developments are emerging: see Royall 2000). The practical realities of real scientific problems strongly suggest that likelihood-only methods are not yet ready for prime time.

It should be clear by now that evidence in science is not and should not be a single number. Evidence is a structure of arguments, in which each structural piece survives continual and clever empirical challenges.

Tobacco company science

There is a class of scientific-appearing people that I call unreasoned skeptics. Unreasoned skeptics do not accept the tenets of the scientific method. They view science as an activity of data interpretation either in light of prior beliefs or to maximize certain utilities. Money, power, and influence are the objects of the scientific game. In this they have a decidedly postmodern outlook. Unreasoned skeptics include tobacco company scientists and Biblical creation scientists. It is perhaps fortunate that these professional debaters tend not to know much about statistics, for I fear that they would find Bayesian statistics well-suited for their sponsored disinformation campaigns.

EDUCATION

"I have taken 18 credits of statistics classes, but I still do not understand statistics."
—Ph.D. student in wildlife

The distressed wildlife student confessed the above to me, toward the end of a long, rigorous program of graduate study. The student had taken a succession of graduate statistics "methods" courses, such as regression, analysis of variance, experimental design, nonparametric statistics, and multivariate statistics, virtually the entire "service" offering of the university statistics program, and had worked hard and recieved near-perfect grades. Yet, the student felt that the subject was still a mystery. My impression, based on twenty years of teaching, research, and consulting in ecological statistics, is that this student's confusion about statistics is not an isolated case, but rather represents the norm in the life sciences. What this student's case illustrates is the sad fact that *the "applied" courses insisted upon for their students by life science educators are designed to perpetuate the confusion*. One can take statistics "methods" courses until the cows come home, and be no nearer to understanding statistics than one is to understanding quantum mechanics.

In this last section of my essay, I offer a prescription for change. It might seem like a digression, but I contend that it is a crucial part of the Bayesian/frequentist problem. Many ecologists have never really been comfortable with statistical concepts (e.g. thinking that a *P*-value is the probability of a hypothesis, etc.), and this discomfort can be exploited by polemicists.

Ecologists are ill-served by their statistics education. For a science in which statistics is so vital, it is paradoxical that statistics is such a source of insecurity and confusion. It is as if the the subject of statistics is a big secret. Ecologists are given glimpses and previews of the subject in their "methods" courses, but the subject itself is never revealed. Shouldn't ecologists instead be trained to wield statistical arguments with strength and confidence?

The topic of statistics could hardly be more important to a science than it is in ecology.

First, ecologists are routinely confronted by nonstandard data. The random mechanisms and sampling schemes encountered in ecology often are not well described by the statistical models underlying "off-the-shelf" statistical methods. I find it ironic that ecologists spend a fair amount of time and journal space arguing about statistics; quantitatively-oriented ecologists even teach statistics and attempt to invent new statistical methods. These are tasks for which ecologists (without the education I discuss below) are by and large untrained. With "methods" courses, one never learns the foundational principles from which statistics methods arise; one merely learns the methods that have already arisen. No amount of "methods" courses and no amount of familiarity with computer packages can compensate for this gap in understanding. In particular, jury-rigged attempts to transfer off-the-shelf analyses to nonstandard situations can result in embarrassment and frequently is the subject of useless controversy.

Second, ecological systems are stochastic. Stochastic models are rapidly becoming an integral part of the very theories and concepts of ecology. Yet, confusion about stochastic models has often marred published ecological discourse. For instance, the density dependence vs. density independence debates, a staple in the ecological literature since the '50s, continue to feature mathematically incorrect statements about persistence, autocorrelation, and statistical tests (see discussion by Dennis and Taper 1994).

I propose that ecologists take less statistics courses. Yes, that is not a typo.

The core of an ecology graduate student's statistical training should be a one-year course sequence in mathematical statistics. The standard "math-stat" course offered at most colleges and universities is an upper division undergraduate course (usually can be taken for graduate credit). It is where the secret is revealed, and by the way is where statisticians commence training their own students. With this course sequence, statistics will be a source of strength and confidence for any ecologist. Though the math-stat sequence is a tough challenge, the ecologist will be rewarded by needing far fewer methods courses in their educations. The usual math-stat sequence, incidentally, gives balanced coverage to both the frequentist and the Bayesian approaches without developing the scientific issues to any great degree (Bain and Engelhardt 1992).

Proper preparation for math-stat is essential. Statistics is a post-calculus subject, and that is the heart of the educational problem. There is no way around this fact. The reduced number of methods courses during the training of a Ph. D. will have to be partially compensated

by thorough undergraduate calculus preparation. This does not mean "business calculus" or even "calculus for life sciences." This means genuine calculus, taken by scientists and engineers. Genuine calculus should be considered a deficiency in graduate admissions; the one-year math-stat sequence can be undertaken in the first or second year of graduate training.

It will help if math-stat is not the first exposure to statistics. The student who has had the usual undergraduate basic statistics course will be continually amazed in math-stat to discover how unified and powerful are the concepts of statistics.

This basic curriculum, calculus, intro statistics, and math-stat, can be followed by selected "methods" courses of particular interest to the student.

One pleasant surprise to a student taking this curriculum will be that the methods courses are quite easy for someone with math-stat training. Textbooks on many statistics topics become readily accessible for self-study. Imagine reading statistics like one reads biology! Another, perhaps more disconcerting surprise will be that the student will be sought out *constantly* in his/her department by faculty and other grad students as a source of statistical help. The demand for statistics consulting can be overwhelming. But that hints at a third surprise, a very pleasant one indeed: command of statistics greatly enhances employability.

Ecologists-in-training who are particularly interested in theoretical ecology, or who are simply interested in strong prospects for gainful employment as scientists, might consider acquiring a graduate degree in statistics. Our M.S. statistics graduates from the University of Idaho wave at their faculty advisors from their lofty tax brackets.

Ecologists struggling with the Bayesian/frequentist issues will ultimately benefit from greater statistical understanding in general. Statisticians are not going to solve the scientific issues of ecological evidence; ecologists must do that to their own satisfaction. But at their current command level of statistics, ecologists are not ready for the task.

CONCLUDING REMARKS

Ecology is a difficult science because of the large variability in ecological systems and the large cost of obtaining good information. It is difficult also because of the sensitive political nature of its scientific questions and the pressing demand for quick answers. However, if ecologists yield to the Bayesians' call for watering down the standards of evidence, the end result will be tobacco company science, not science.

LITERATURE CITED

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, eds. Second international symposium on information theory. Akademiai Kiado, Budapest, Hungary.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716-723.

Anderson, W. T. 1990. Reality isn't what it used to be: theatrical politics, ready-to-wear religion, global myths, primitive chic, and other wonders of the post modern world. Harper & Row, New York, Nesw York, USA.

Bain, L. J., and M. Engelhardt. 1992. Introduction to probability and mathematical statistics. Second edition. Wadsworth Publishing Company, Belmont, California, USA.

Berger, J. O., and D. A. Berry. 1988. Statistical analysis and the illusion of objectivity. American Scientist 76:159-165.

Connor, E. F., and D. Simberloff. 1979. You can't falsify ecological hypotheses without data. Bulletin of the Ecological Society of America 60:154-155.

Connor, E. F., and D. Simberloff 1986. Competition, scientific method, and null models in ecology. American Scientist 75:155-162.

Cox, D. R., and D. V. Hinkley. 1974. Theoretical statistics. Chapman and Hall, London, UK.

Dennis, B., R. A. Desharnais, J. M. Cushing, and R. F. Costantino. 1995. Nonlinear demographic dynamics: mathematical models, and biological experiments. Ecological Monographs 65:261-281.

Dennis, B., and M. L. Taper. 1994. Density dependence in time series observations of natural populations: estimation and testing. Ecological Monographs 64:205-224.

Dixon, P. M., and K. A. Garrett. 1993. Statistical issues for field experimenters. In R. J. Kendall and T. E. Lacher, eds. Wildlife toxicology and population modeling. Lewis Publishers, Boca Raton, Florida, USA.

Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge, UK.

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

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London A 222:309-368.

Hairston, N. G. Sr. 1989. Ecological experiments: purpose, design, and execution. Cambridge University Press, Cambridge, UK.

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

Jaynes, E. T. 1968. Prior probabilities. IEEE Transactions on systems science and cybernetics SSC 4:227-241.

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

Johnson, D. H. 1999. The insignificance of statistical hypothesis testing. Journal of Wildlife Management 63:763-772.

Lee, P. M. 1989. Bayesian statistics: an introduction. Oxford University Press, New York, New York, USA.

Lehmann, E. L. 1983. Theory of point estimation. John Wiley & Sons, New York, New York, USA.

Lindley, D. V. 1990. The 1988 Wald memorial lectures: the present position in Bayesian statistics (with discussion). Statistical Science 5:44-89

Marmorek, D. R., ed. 1996. Plan for analyzing and testing hypotheses (PATH). ESSA Technologies, Vancouver, British Columbia, Canada.

Mayo, D. G. 1996. Error and the growth of experimental knowledge. The University of Chicago Press, Chicago, Illinios, USA.

Neyman, J., and E. S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London A 231:289-337.

Ott, R. L. 1993. An introduction to statistical methods and data analysis. Fourth edition. Brooks/Cole Publishing Company, Pacific Grove, California, USA.

Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences 47:2-15.

Peters, R. H. 1991. A critique for ecology. Cambridge University Press, Cambridge, UK.

Potvin, C., and D. A. Roff. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? Ecology 74:1617-1628.

Reckhow, K. H. 1990. Bayesian inference in non-replicated ecological studies. Ecology 71:2053-2059.

Rice, J. A. 1995. Mathematical statistics and data analysis. Wadsworth Publishing Company, Belmont, California, USA.

Robert, C. P. and G. Casella. 1999. Monte Carlo statistical methods. Springer, New York, New York, USA.

Royall, R. 1997. Statistical evidence: a likelihood paradigm. Chapman & Hall, London, UK.

Royall, R. 2000. On the probability of observing misleading statistical evidence (with discussion). Journal of the American Statistics Association 95:760-768.

Saarinen, E., ed. 1980. Conceptual issues in ecology. Reidel, Dordrecht, Netherlands.

Scheaffer, R. L., W. Mendenhall, and L. Ott. 1996. Elementary survey sampling. Fifth edition. Brooks/Cole Publishing Company, Pacific Grove, California, USA.

Schrader-Frechette, K. S., and and E. D. McCoy. 1993. Method in ecology: strategies for conservation. Cambridge University Press, Cambridge, UK.

Simberloff, D. 1990. Hypotheses, errors, and statistical assumptions. Herpetologica 46:351-357.

Sindermann, C. J. 1982. Winning the games scientists play. Plenum Press, New York, New York, USA.

Smith, S. M. 1995. Distribution-free and robust statistical methods: viable alternatives to parametric statistics? Ecology 76:1997-1998.

Sokal, A. and J. Bricmont. 1998. Fashionable nonsense: postmodern intellectuals' abuse of science. Picador, New York, New York, USA.

Stewart-Oaten, A. 1995. Rules and judgments in statistics: three examples. Ecology 76:2001-2009.

Strong, D. R., D. Simberloff, L. G. Abele, and A. B. Thistle, eds. 1984. Ecological communities: conceptual issues and the evidence. Princeton University Press, Princeton, New Jersey, USA.

Stuart, A., and J. K. Ord. 1991. Kendall's advanced theory of statistics. Volume 2: classical inference and relationships. Fifth edition. Griffin, London, UK.

Student. 1908. The probable error of a mean. Biometrika 6:1-25.

Toft, C. A., and P. J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. The American Naturalist 122:618-625.

Underwood, A. J. 1990. Experiments in ecology and management: their logics, functions, and interpretations. Australian journal of ecology 15:365-389.

Vardeman, S. B. 1994. Statistics for engineering problem solving. PWS Publishing, Boston, Massachusetts, USA.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. Annals of Mathematical Statistics 9:60-62.

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