2

Designing studies and interpreting population biology data: how do we know what we know?

> *It is not enough to say that we cannot know or judge because all the information is not in. The process of gathering knowledge does not lead to knowing. A child's world spreads only a little beyond his understanding while that of a great scientist thrusts outward immeasurably. An answer is invariably the parent of a great family of new questions. So we draw worlds and fit them like tracings against the world about us, and crumple them when they do not fit and draw new ones. The tree-frog in the high pool in the mountain cleft, had he been endowed with human reason, on finding a cigarette butt in the water might have said, "Here is an impossibility. There is no tobacco hereabouts nor any paper. Here is evidence of fire and there has been no fire. This thing cannot fly nor crawl nor blow in the wind. In fact, this thing cannot be and I will deny it, for if I admit that this thing is here the whole world of frogs is in danger, and from there it is only one step to anti-frogicentricism." And so that frog will for the rest of his life try to forget that something that is, is.*
>
> **John Steinbeck (1960), *Log From the Sea of Cortez***

Introduction

Why learn about study design, data interpretation, and a touch of scientific philosophy and ethics in a book on ecology and conservation of wildlife populations? Simply put, because without reliable knowledge both science and the application of science fall flat. The best scientists, managers, and decision-makers are those who can separate lousy, unreliable, or irrelevant information from important and trustworthy information.

Reliable knowledge is nothing less than the outcome of the quest to judge **truth**. Heavy stuff, for sure, but it has profoundly practical implications. In essence, truth is the correspondence between an idea created in our mind – that is, an $\mathbf{idea}_{\mathbf{mind}}$ – with a referent fact obtained from sensory experience (call this a **fact**). Dr Charles Romesburg – who rattled the field of wildlife science with a classic paper on "Gaining reliable knowledge" in 1981 – has noted that without comparison to a factual referent, an idea of the mind is only an opinion (Box 2.1). Saying it another way, the process of

Box 2.1 Notes on truth, knowledge, and opinion

These notes were modified from unpublished notes by Dr Charles Romesburg, Utah State University.

- There are two kinds of idea:
  1. referent "facts" from sensory experience, and
  2. ideas "from the mind" developed from free creation apart from sensory experience ($idea_{mind}$).
- We say $idea_{mind}$ is **true** when $idea_{mind}$ = fact (what is, is). We say $idea_{mind}$ is **false** when $idea_{mind} \neq$ fact.

Without appeal to a referent humans tend to disagree. That is why there is consensus that a given $idea_{mind}$ that has been tested and found to agree with facts is a rightful candidate to the concept "knowledge", as opposed to a given $idea_{mind}$ that has never been exposed to the factual arbitrator.

Opinions are ideas that have been declared as neither knowledge or falsehood. They are in limbo and will remain so until risked in comparison to the factual referent (either direct or indirect comparison). Note that the tolerance for testing truth is an opinion, making all knowledge depend on opinion.

Reason has always been around. Knowledge made little growth up to the 16th century because reason was the sole basis for knowledge. Then came the scientific revolution that blended reason with facts. Galileo, Kepler, and others changed how science was done by risking the predictions from their reasoning with facts.

moving from opinion to **knowledge** is all about coming up with creative ideas ($idea_{mind}$) and comparing them to reliable facts, then revising the $ideas_{mind}$ if they fail to hold up to the facts[1]. Although Steinbeck's frog (in the quote above) recognizes that his idea of the mind (that cigarettes either originated from local materials, or dispersed) fails to match his facts (there is a cigarette in the pond), he denies the facts instead of recognizing that his $idea_{mind}$ should be revised!

So, the scientific process is really just a process of comparing creative and meaningful $ideas_{mind}$ to reliable facts, and then following an objective and orderly process of modifying the $idea_{mind}$ when the two do not match. That's what this chapter is all about.

[1]Of course, both $idea_{mind}$ and facts are rich fodder for epistemological and philosophical discussion, including subtleties on how the $ideas_{mind}$ affect our ability to recognize facts. I use the terms to make the simple but profoundly important point that reliable knowledge in wildlife population biology requires deep thought as to how nature works (the $ideas_{mind}$) as well as information gained from study design and data collection (the facts).

Obtaining reliable facts through sampling

How do we obtain the facts from the field against which we compare the ideas of our mind? We do it through appropriate sampling and study design. The facts might be estimated effects of treatments, or they might be estimates of **parameters**. Parameters are quantities of the population for a given area and time; for example, the true survival rate of adults. We almost never know what the true parameters are; rather, we estimate parameters from data. By convention, **estimates** have hats ($\hat{}$) over them; for example, an estimate of abundance (N) is denoted $\hat{N}$. In this section I'll discuss some main points of sampling and study design.

Replication and randomization

When you cook a pot of spaghetti and wonder if it is done, how do you decide? Perhaps you slavishly watch the clock, read the directions, and remove the noodles at just the right time. More likely, though, you sample. But how? Typically, you grab one noodle and taste it. If your pot is big enough and the water has boiled vigorously, one noodle from anywhere in the pot should be enough. But if you have a small pot, or it hasn't boiled very well, you would probably sample a few noodles, perhaps at least one from the bottom and one from the top.

So, in a homogeneous noodle population you might draw inferences from as few noodles as one, but as your noodles become heterogeneous – variable in their doneness – you would **replicate** your sampling. There is almost no meaningful question in our field that involves a population as homogeneous as a pot of noodles, so replication is a cornerstone of reliable sampling. Formally, replicates are the multiple members sampled from a population of interest, or the number of units to which a treatment is independently assigned. The **sample size** for a treatment is the number of replicates. Replication keeps us from making a decision based on a single, potentially unusual, sample. Replication also facilitates an estimate of variation, providing a basis for a statistically sound decision as to whether the population in question – given its variable nature – is really different from another population that we care about (Johnson 2002).

There are a couple of rules about appropriate replication. First, the replicates should be sampled at an appropriate scale to capture the relevant variation over time and space. If you are interested in how owl clutch sizes differ between a single logged and a single thinned forest, then 10 reproducing owls in each forest would constitute the number of replicates. However, if instead you were interested in the general question of whether owl clutch size differs in unlogged and thinned forests, then 10 reproducing owls in one forest of each treatment would constitute only one sample (no replication) with 10 **subsamples**. To treat the 10 subsamples as samples would lead to a mismatch between the number of independent samples (one) and the desired inference, leading to **pseudoreplication** (Hurlbert 1984). To avoid pseudoreplication for the question of owls in unlogged and thinned forests, several replicates of different unlogged and thinned forests would be sampled across the landscape of interest, with

one to many owls subsampled within each forest. It makes sense that we would try to avoid pseudoreplication, as it is easy to think of many other extraneous factors other than thinning – including predators, prey, aspect, vegetation, and so on – that would cause the owls in one forest stand to reproduce differently from the owls in another stand even if thinning had nothing to do with it.

A second rule about replicates is that they should be chosen **randomly**; that is, every member of the population should have a chance of being sampled. In a manipulative experiment, the treatment received by each unit should be assigned randomly. Randomization reduces the chance that some ancillary factor will bias our measurements because it makes it less likely that systematic differences other than the treatment could have caused the observed effects[2]. If we picked unlogged forests at high elevation and logged forests at low elevation, and elevational differences led to differences in clutch size, then our inference about the effect of logging would be wrong, even with 100 replicates, because it would really represent an effect of elevation. The bottom line is this: although it may be more convenient to sample nonrandomly – perhaps along a road or a trail – inference becomes limited and much less valuable than with random sampling.

Controls

Controls include a number of ways that scientists ensure that facts are not confounded by some unexpected influence other than the $\text{idea}_{\text{mind}}$ we are evaluating. For example, when someone refers to **controlled conditions** they mean that they are making sure that the desired test conditions are occurring. In field projects, **control sites** or **control treatments** are those whose only consistent difference from the treated sites is the application of the treatment (chosen randomly from available sites). Similarly, control procedures account for effects that might be caused by methods used to apply experimental treatments; for example, if the question is how leg bands affect bird survival, then some birds might be handled but not marked as a control procedure to identify whether the handling may confound the effect of the leg band on survival (Hurlbert 1984).

Finally, when a degree of subjectivity is involved in measuring a factual referent, **blind controls** can minimize the chance that the observed treatment effect does not carry insidious (and often unconscious) influences from wishful thinking by the observer. In a blind control, the person collecting or analyzing the data does not know the treatment group or identity of the sample they are evaluating. For blind controls, impeccable record-keeping and protocols for conducting the test are especially critical.

[2]Although random sampling is usually best, sometimes the nature of the question makes it impossible. In such cases, other approaches, such as systematic, stratified random, cluster, adaptive, sequential, or other sampling methods, may be appropriate (Thompson 2002). The key is to think through how the sampling method may affect estimates of both bias and variance, given the sampled population.

Accuracy, error, and variation

Accuracy and **error** are two sides of the same coin, describing how well the mean estimated from sample observations corresponds to the true mean. Accuracy, or its flip side, error, is made up of two components that can be quite unrelated to each other: bias and precision. Consider an estimate of a population parameter, say abundance, survival, weight, or sprint speed. **Bias** refers to systematic deviation of the estimate from the true parameter of interest. **Precision** refers to the amount of scatter, or repeatability of the estimate when made many times; it is quantified by the variance, with high variance indicating low precision (Box 2.2). Either a large bias or low precision (high scatter) will result in low accuracy[3], as portrayed in the classic bulls-eye diagram shown in Fig. 2.1.

Let's consider in more detail the error that comes from low precision. Lack of precision arises from both process variance and sample variance. **Process variance** is genuine biological variance that arises because conditions vary (e.g. temperature, moisture, diseases; Thompson et al. 1998, Mills and Lindberg 2002). Specifically,

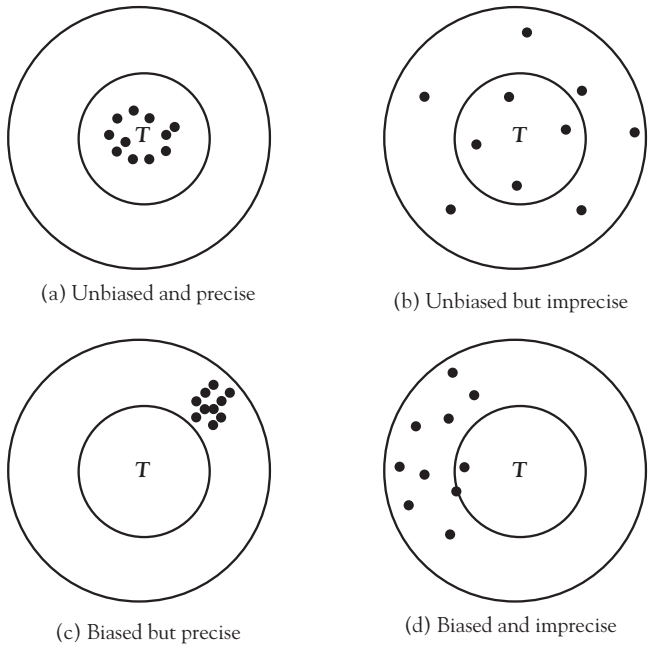

Fig. 2.1 An accurate estimator is one that is unbiased and precise. The center of the circle (the bulls-eye) is truth, denoted by T (for example, a survival rate of 0.85 or an abundance of 220). The dots show sample estimates of the parameter T. Estimates in (a) are accurate, being unbiased and having high precision; all others are inaccurate. Modified from White et al. (1982).

[3]For the statistically inclined, overall error is captured by the mean squared error (MSE) for an estimate X: $\text{MSE}(X) = \text{variance}(X) + \text{bias}(X)^2$ (Williams et al. 2002).

Box 2.2 A primer on variance, standard deviation, and standard error

Variance gives an indication of the **spread** of what you measured in your population. Correcting for finite sampling (as we always must when we sample just part of a population) with n observations, the equation for sample variance is the average squared deviations of all of the x_i measurements from the mean ($\bar{x}$):

$$s^2 = \frac{\sum_{i=i}^{n} (x_i - \bar{x})^2}{n-1}$$

Standard deviation (SD) is the square root of the variance. This is useful because it describes the spread of your variable in terms of what you measured (say, animals per hectare), instead of the non-intuitive "squared animals per hectare" from calculating variance. For data with a bell-shaped (normal) distribution, the mean plus or minus two SDs will contain about 95% of the population.

So far we have only talked spread of measurements from a population. But what if we are interested particularly in how well the mean characterizes the sample? The **standard error** (SE) **of the mean** is the estimate, from a single sample, of the SDs of the distribution of means expected if we collected many samples of size n and calculated a mean for each sample. In practical terms, SE quantifies how confident we are that our estimated mean is close to the true mean.

The SE is estimated from our one sample as $SD/\sqrt{n}$. (In some cases, such as for most mark–recapture estimates of vital rates, the SEs are simply the square root of the variance).

It makes sense that a larger sample size (n) will decrease the SE, because more sampling should increase our confidence that the true population mean will be close to the estimated mean. (In contrast, increasing n will not change the estimate of variance or SD). Notice that the word error in standard error is not a statement of mistakes or bad judgment.

An ***X*% confidence interval** is obtained by adding and subtracting from the mean a value weighted by the SE (for example 1.96 * SE for a 95% confidence interval assuming a normal distribution); informally this is thought of as providing the range in which we suspect the unknown true mean should be found. More formally, that confidence arises from the fact that if we were to repeat the study many times, X% of the confidence intervals constructed in this way would include the true mean. Both the SE of the mean and the confidence interval indicate how precisely the mean is estimated.

The **coefficient of variation** (CV) expresses variation relative to the mean:

$$\text{CV} = \text{SD/mean}$$

CV is useful in the many cases where variance (and SD) increase with the mean, so we want to know whether some measure (say, offspring production) is relatively more variable for a group with a high mean (say, coyotes) compared to one with a low mean (say, bears). It is often expressed as a percentage.

Box 2.3 An insidious form of error due to being human

Although researchers work hard to avoid bias and increase precision in measurements – for example by calibrating instruments, and by proper replication and randomization – unconscious errors (or lack of accuracy) can sneak up when the study requires a subjective decision. Social psychologists, medical researchers, and educators have worried about this a lot, and regularly use so-called double-blind setups where the person being interviewed does not know the purpose of the study, and the interviewer does not know the identity of the interviewee.

During a study of coyotes in Northern Utah (Mills & Knowlton 1989), Fred Knowlton and I wondered if radiotelemetry error tests really captured the error inherent in estimating azimuths to animals. Field assistants were working under grueling conditions, squashed in a little fixed-station box on the back of a truck for 8 hours or more through the night. Might the error in their telemetry bearings be different in those conditions than in a known telemetry test where they were aware that they were being tested?

To evaluate this possibility, we told field assistants that we had captured some new coyotes over the weekend, when actually the new transmitters were test transmitters placed at known locations. After several nights of collecting data (with known error because we knew the exact location of the fake coyotes), we conducted a traditional telemetry accuracy check where assistants knew they were being tested.

When observers knew they were being tested their precision was quite a bit better than the blind test. Knowing that their diligence and accuracy were being examined, they worked harder and longer to obtain azimuths. Our recommendation to obtain a better estimate of telemetry accuracy is to conduct telemetry accuracy tests without field helpers knowing they are being tested.

spatial process variance arises from changes in species present, habitat quality, and habitat heterogeneity over the landscape, which in turn may be related to environmental conditions such as aspect, slope, precipitation, and successional-stage differences. Temporal process variance is often driven by weather, as well as interactions with competition, predation, disease, and human impacts. Even if you know a parameter (say, survival) exactly, it will fluctuate over time and space due to process variance.

In contrast to process variation, which acts directly on organisms, sampling variation is a product of incomplete information from the act of sampling a larger population. One component of **sample variance** comes from human error, as when observers make subjective decisions (Box 2.3), but more fundamentally it is an inevitable result of estimating something by sampling from a population. Although sample variance is present and real in nearly every measurement of a fact, it is a nuisance that inflates total variance artificially. Thus process variance, the actual biological variation inherent in the thing that we measure, has to be teased out of the total variance measured[4]:

[4]We'll be returning to process versus sampling variance as components of total variance in later chapters.

Real process variance = total variance measured – sample variance

Process variance is often as important to quantify as the average or mean. As an example, here's the folksy plea of a waterfowl biologist named Johnny Lynch, frustrated by how duck population goals were being set based on a national average that ignored variation (quoted by Ankney 1996:41):

> *Did someone say that the "average" numerical standing of the North American mallard population over some period of years would be a good standard for management to try to maintain? Don't let the Old Forecaster hear such talk. Not long ago, he got involved in certain philosophical deliberations regarding the "average" condition of dynamite. Which commodity, in its quiescent state, was a small cylinder having a volume of a few cubic inches, yet at its peak of explosion occupied hundreds of cubic yards. Seemingly, the "average" condition of dynamite could be determined by adding together measurements taken at various levels between these two extremes, and dividing their sum by the number of measurements. The forecaster took one look at the results of all this arithmetic, turned slightly purple, and then decided ruefully that dynamite in its "Average condition" must be one helluva thing to crate, ship, and otherwise handle. He has assiduously avoided "averages" ever since that unfortunate experience.*

Without a doubt, variance and extreme events are critical for understanding wildlife dynamics. In fact, throughout the book we will see that the resolution to many applied issues – ranging from expected time to extinction of endangered species, to the number of animals in an area over time, to the consequences of a particular harvest rate – are driven by the extremes, as in Lynch's dynamite example. Variance may be of interest in its own right, so we may compare (for example) a coefficient of variation (Box 2.2) in clutch size among species, or in relative humidity among logging treatments. In other cases the variance of estimated treatment means can warn us to be humble about concluding differences. For example, if density of snowshoe hares in logging treatment A averages 1.8/ha and density in another logging treatment B averages 1.3/ha, can we say that density is higher in A? It depends on the variation around the means. So, embrace uncertainty by describing variation.

Linking observed facts to ideas$_{mind}$ leads to understanding

So far we have thought a bit about how to obtain field observations, or facts, that we can count on. This is critically important, to be sure, but with only a fact in hand we come face to face with the question, "So what?" For a fact to be converted into meaningful knowledge or understanding it must be linked to an idea of the mind that helps to explain a phenomenon. I will touch on some ways to do that via frameworks to evaluate hypotheses using frequentist statistics and *P* values, information-theoretic criteria, and Bayesian analysis.

The hypothetico-deductive approach

Instead of belaboring philosophy of science or the **scientific method**, I will talk a bit about the hypothetico-deductive method as a workhorse for gaining reliable knowledge. The essential steps of the hypothetico-deductive method are:

- identify the question or phenomenon of interest,
- develop hypotheses that explain the phenomenon[5],
- deduce diagnostic predictions that follow from each hypothesis, and
- gather observations (facts) to test the predictions.

Do not underestimate the importance and difficulty of the first steps of defining an important question and deriving meaningful and testable hypotheses *a priori* (before the study or data analysis begins). One important basis for hypothesis generation is descriptive work, including natural history studies and the accumulation of insights from field observations. I think there is ample reason to mourn the shattered stature of natural history work, both because it helps frame our questions and because our souls are filled when we can be a sponge in the field, absorbing all that we can see, hear, smell, and touch. However, descriptive studies by themselves are not sufficient to quickly advance knowledge in any branch of science. Therefore, while we should be earnest students of natural history, filling our field notebooks with observations and thoughts, realize that this is just a first step in a process that leads to rapid accumulation of reliable knowledge.

Similarly, **induction** – the process of forming general conclusions based on associations in a collection of observations – serves an excellent role in hypothesis generation, but is not an efficient scientific process by itself. Correlations between two variables are a classic form of induction, and the truism that correlation does not equal causation applies to all walks of life. For example, I have a newspaper article that makes a serious attempt to use a rather sketchy correlation to imply causation between values of stocks in the USA and women's skirt lengths! Correlations and associations are good for generating hypotheses to be tested, but weak for concluding mechanisms.

Once a hypothesis is developed, it must be strongly connected to predictions. First, the prediction(s) must logically follow from the hypothesis. In other words, if the prediction is falsified we need to have confidence that the hypothesis is also false; a prediction that may or may not follow from the hypothesis will not reject the hypothesis as false. Second, confounding factors may cause the predictions to be supported even if the hypothesis is false. Avoiding the insidious problem of concluding a hypothesis is supported, when really a confounding factor supported the prediction, is at the heart of most study design and statistical analysis.

[5]The distinction between **hypothesis** and **theory** is a topic of much discussion among philosophers of science; typically a theory is a broader conceptual framework from which specific, testable hypotheses are derived. Also, here I'm using hypothesis in the sense of a **research hypothesis** into how nature works, as opposed to **null** or **statistical hypotheses** that investigate specific questions that may or may not be causal (Steidl et al. 2000).

Hypothesis tests can be accomplished either by making observations about the world as it is or by manipulating something and observing what happens. The latter approach is by far the most powerful, because by manipulating the system you make it less likely that the observed results came from something other than your treatment (Romesburg 1981, Johnson 2002). The process of multiple alternative hypotheses being exposed to critical experiments to efficiently weed out non-viable alternatives is known as **strong inference** (Platt 1964).

Obviously, manipulation (and replication and randomization for that matter) can be difficult or impossible for many important questions, especially those on big scales (say, community-level effects of removal of sea otters) or those that happen in just one place (effects of an oil spill, or a dam). Still, such questions can be formally evaluated, perhaps through careful collection of data before and after treatment, and by minimizing extraneous confounding factors. Also, whole studies can and should be repeated (**metareplication**; Johnson 2002) and analyzed as a group (**meta-analysis**). If an $\text{idea}_{\text{mind}}$ is supported in a study repeated in different years, at different sites, with different methodologies, or by different investigators, you are much more likely to believe that it is real.

P values, power, and biologically important differences

In the most common form of hypothesis testing, the final step is to determine a *P* value, otherwise known as a test for **statistical significance**. By convention, if $P < 0.05$ (or sometimes $P < 0.10$) the null hypothesis is rejected and a **significant** effect is declared[6]. The validity and misuse of null hypothesis significance testing has been vigorously debated in the last decade (Yoccoz 1991, Anderson et al. 2000, Robinson & Wainer 2002). Here are some main points. First, no heavenly dictum supports a magic threshold of $P < 0.05$. Far too often biological sense has been thrown out of the window with a $P = 0.05$ dichotomy leading to $P = 0.04$ being trumpeted as significant – with all sorts of implications for ecology and management – while the same test with $P = 0.06$ is panned as meaningless and insignificant. We should "stop treating statistical testing as an all-or-nothing procedure and instead use appropriate wording to describe degrees of uncertainty" (Robinson and Wainer 2002:269). For example, differences might lean in a certain direction, or indicate a hint about the true direction, or even indicate simply that differences could not be determined; thus the study needs to be repeated before reliable inference can be made.

Second, the *P* value does not give the probability that the null hypothesis is true, and smaller *P* values alone do not necessarily mean a more false null hypothesis. Rather, *P* values tell you the probability of observing data as extreme (or more extreme) as the observed data given that the null hypothesis is true, with repeated sampling of

[6]To be precise, Type I error or α is set at 0.05 as a cutoff, and the *P* value for the test is compared to the preset α; if *P* is less than $\alpha = 0.05$, then the test is deemed statistically significant (Type I error is discussed later in this section).

Table 2.1 Possible outcomes of decisions based on frequentist statistics and *P* values. (a) A medical example where the null hypothesis is that a patient does not have a fatal disease. (b) An example from common endangered species monitoring where the null hypothesis is that a population is stationary (no downward trend in numbers). λ represents the population growth rate. Statistical power is represented by the lower right-hand cell in both panels. Values in parentheses give the probabilities associated with each decision.

(a) Medical example

	Patient actually . . .	
***Your* decision**	**does not have the disease**	**does have the disease**
Fail to detect disease	Correct: state no disease ($1 - \alpha$)	Incorrect: Type II error (β)
Detect disease	Incorrect: Type I error (α)	Correct: state that disease is present ($1 - \beta$)

(b) Endangered species monitoring example

	Population actually . . .	
***Your* decision**	**is stationary, $\lambda = 1.0$**	**is declining, $\lambda < 1.0$**
Population is stationary, $\lambda = 1.0$	Correct: state no decline ($1 - \alpha$)	Incorrect: Type II error (β)
Population is decreasing, $\lambda < 1.0$	Incorrect: Type I error (α)	Correct: decline detected ($1 - \beta$)

the data[7]. A small *P* value does not necessarily indicate that the effect or treatment was large because small *P* values can also arise with a small effect size if sample sizes are large or variability small.

When testing a hypothesis using *P* values, the inference either correctly matches the true state of nature, or is wrong. If wrong, the inference can either conclude a difference between treatments when really there is none, or conclude no difference when really there is one (Table 2.1). The first error, falsely concluding a difference, has traditionally received the most attention and for historical reasons is called a Type I error (symbolized by α); a *P* value is deemed statistically significant if it is less than α. The second error, concluding no difference when really there is one (or, in null hypothesis jargon, concluding that the null hypothesis of no change is supported when really the null is false) is Type II or β.

We are ingrained, often through statistics classes, to focus primarily on minimizing Type I error to decrease the probability of saying there is a difference or effect when

[7]The fact that the statistics are conceptually based on long-run frequencies under repeated sampling explains the moniker of **frequentist** statistics.

really there is not one. That's why the arbitrary threshold of $\alpha = 0.05$ (5%) is so ingrained in our field. But is Type I error always worse than Type II? Consider a medical analogy (Table 2.1a). The null hypothesis is that a person does not have a fatal but treatable disease. If the person really does not have the disease, then we could either correctly detect no disease, or incorrectly detect the disease, thereby falsely rejecting the null hypothesis and committing a Type I error (called a false positive in medical research). On the other hand, if the person actually has the disease, we could either correctly detect the disease or commit a Type II error by failing to detect it (a false negative). Which of these mistakes is worse? A Type I error would upset the patient, and potentially cause them to question the credibility of the test. But follow-up tests would surely occur, indicating no disease. If, however, a Type II error occurred, the disease would not be detected and the doomed patient would head back out in the world, soon to die, a victim of Type II error.

By analogy, an assertion that an endangered species is decreasing when it is really stationary (Type I error) may initiate unnecessary restrictions, leading to some loss of credibility (Table 2.1b). However, a declaration that the population is stationary when really it is declining (a Type II error) means that the population is heading toward extinction while nothing is done! Similarly, when evaluating the effects of exploitation, the consequences of Type II errors (failing to detect a real negative effect) may be of more concern than Type I error (Nichols et al. 1995).

Statistical power is the probability of not making a Type II error, or the probability of correctly rejecting a false null hypothesis. In practical terms, statistical power is the probability of detecting a pre-specified difference or effect that is really there. Power is positively related to α, sample size, and size of effect (e.g. the difference among treatments or the steepness of trend lines), and negatively related to variation. Thus, if sample sizes, effect sizes, or the specified Type I error rate are very small, or if variation is large, then power may be too low to detect a difference that is real. In fact, because studies of trends in endangered species almost always involve subtle changes, small sample sizes and/or large variance, there will be low power to detect real declines (Taylor & Gerrodette 1993, Gibbs et al. 1998)[8]. Beware of the stacked deck of cards presented to biologists assessing declines or other effects in a null hypothesis framework: it may be very difficult to document a real decline when sample sizes are small or variance is large.

Power analysis should be used in the planning stages of an experiment or management action to determine necessary sampling designs and sample sizes, and whether the study is even possible given logistical and financial constraints that limit sample size or lead to power-busting high variation[9]. Such *a priori* analysis greatly improves

[8]Sometimes people will increase α to, say, 0.1 instead of 0.05, as a way to increase power.

[9]Power tests typically should not be used to interpret results after a *P* value has been obtained (so-called *a posteriori*, retrospective, or *post hoc* power analysis). Although many statistical packages offer these *post hoc* power tests, the power estimates are redundant with the *P* value of the finished study, and are unreliable (Steidl et al. 1997, Gerard et al. 1998). Once you've done your study, just present effect sizes and confidence intervals and let the reader interpret how sample size, variance, α, and effect size affect the interpretation of biological significance.

sampling design because it helps determine: (i) the sample size needed to detect an effect that you feel is biologically meaningful; (ii) the detectable effect for a given sample size, power, variance, and α level; and (iii) the power to detect a certain effect if we were to initiate a study with an expected effect, sample size, variance, and α. Ideally one should consider a range of values, which is relatively easy given the power-analysis modules in data-analysis software (Thomas & Krebs 1997).

The misinterpretations of *P* values and problems with power in the null hypothesis framework have led some to suggest that the *P* value framework be abandoned in favor of only presenting effect sizes and confidence intervals (Johnson 1999, Anderson et al. 2000). Certainly presenting a naked *P* value without the effect size and precision is almost always a bad idea. Presenting the effect (e.g. histograms of means) and precision (e.g. SE) helps clarify the distinction between biological and statistical significance because it lets the reader judge the implications of the observed effect and how well the parameter of interest was estimated (Yoccoz 1991).

Recently, much attention has been paid to alternative frameworks for hypothesis testing that avoid the issues of *P* values and power entirely. These include information-theoretic methods and Bayesian methods, briefly described below.

Model selection based on information-theoretic methods

Instead of using *P* values, null hypotheses, significance testing, and α or β errors, a model-selection framework selects among models conceived by the researcher to identify biologically realistic sources of variation (Burnham & Anderson 2002, Johnson & Omland 2004). By determining which models in the candidate set best approximate the data, hypotheses about biological processes (alternative models) can be tested, and parameters can be estimated.

The first step is to build models from biological intuition, knowledge of the system, and previous studies. With these *a priori* models in hand, data are collected and used to test the fit of each model to the data. The comparison among models requires an objective criterion, bringing us to **Akaike's information criterion** (AIC), one of the most famous analytical buzzwords roaring onto the scene of 21st century population ecology. AIC in particular and information-theoretic methods in general operate on the simple principle that $\text{ideas}_{\text{mind}}$ should be rewarded when they fit the data with the least number of parameters, or pieces. Under the principle of **parsimony** the best model (or models) are those with the highest likelihood[10], given the data, but also those that are the simplest, with the fewest parameters. If a model has too few parameters, or the wrong ones, it will not fit the data well and will lead to biased estimates; adding parameters will almost always improve fit to the data – thereby decreasing bias – but the additional pieces make variance balloon and make spurious factors more likely to

[10]In case the word **likelihood** is foreign to you, the key to distinguishing between likelihood and the more familiar probability is "with probability the hypothesis is known and the data are unknown, whereas with likelihood the data are known and the hypotheses unknown" (Hilborn & Mangel 1997:133).

be deemed important. So AIC is calculated for each model; the lower the AIC value the better the fit to data without extra parameters. Formally,

$$\text{AIC} = -2\ln\left[L\left(\hat{\theta}\right)|\,data\right] + 2K \tag{2.1}$$

Where $\ln\left[L\left(\hat{\theta}\right)|\,\text{data}\right]$ is the value of the maximized log-likelihood parameter θ given the data and the model, and K is the number of parameters estimated in that model. All models are ranked, beginning with the model with the lowest AIC considered to be the best one for that set of empirical data[11]. The **Akaike weight** of each model provides a relative weight (w_i) of evidence for each model, interpretable as the probability that model i is the best for the observed data, given the candidate set of models[12].

If a single model is clearly the best approximating of the set – say seven or more AIC units smaller than the next best model – then there is little uncertainty in model selection. In many cases, however, there is not clear support for any single model, and indeed a commonly used convention is to consider any models within two AIC units of the best approximating model as having substantial support. In such cases, one can account for **model-selection uncertainty** (within the candidate models considered) by summing Akaike weights for all the models that contain particular parameters or predictor variables. Similarly, if you are interested in estimating a parameter such as survival, and no single model has overwhelming support (say, with an Akaike weight, w_i, of the best model <0.9), then you could calculate a weighted average of ($\hat{\theta}$) across all R models:

$$\hat{\theta} = \sum_{i=1}^{R} w_i \hat{\theta}_i \tag{2.2}$$

(for variance and confidence intervals see Burnham & Anderson 2002).

There are several important caveats. First, AIC is not the only game in town; alternative model-selection criteria exist (Taper 2004), including the Bayesian approaches

[11] Actually, in virtually all applications in our field you should use the small-sample unbiased version, AIC_c. If the data are overdispersed – with inflated variance – a variant called QAIC_c is used. Overall, realize that although I just refer to AIC in the text, you will need to go to more advanced sources to learn the subtleties of which version of AIC to use.

[12] The dirty details: first you calculate a ΔAIC for each model i as the difference between that model's AIC and that of the best model (the one with the lowest AIC):

$$\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$$

Because the likelihood of each model i given the data is $\exp(-1/2\Delta_i)$, the Akaike weight normalizes the likelihoods across all R models in the set, so they sum to 1:

$$\text{Akaike weight} = w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^{R}\exp\left(-\frac{1}{2}\Delta_r\right)}$$

described next. Second, the concept of sufficient sample sizes and statistical power continues to be relevant here as much as with *P* values, because appropriate model selection depends on the data available.

Third, remember that under this model-selection framework all inferences are dependent on (or, in statistics lingo, are conditional on) both the data and the set of models considered[13]. You should avoid **overfitting** a shotgun blast of arbitrary and potentially spurious models that lead to wrong estimates and indicate support for variables that fit the particular data-set but that are not biologically relevant. Perhaps more serious, however, is **underfitting**, or leaving out important models. In an interesting discussion of the consequences of underfitting in AIC analyses, Beissinger and Snyder (2002) argued that biological inferences – and subsequent management recommendations for snail kite recovery – were fundamentally compromised when a study left out key models when testing for effects of water level on nesting success (see the response by Dreitz et al. 2002).

To minimize the risk of a poor model set, the most parameterized **global model** (the one that includes all potentially relevant effects and causal mechanisms considered) can be tested for **goodness of fit**. Model-fit statistics (e.g. regression residuals, R^2, or formal chi-square tests) assess whether the most complex model adequately describes the variation in the data. In general, the dependence of inferences on *a priori* models developed by the researcher reinforces the points made above about the importance of the critical development of biological hypotheses early in the study (preferably prior to data collection, and certainly before data analysis!).

There are certainly critics of information-theoretic approaches based on AIC (Guthery et al. 2005). However, model-selection approaches are here to stay as facilitators of the estimation of population parameters, and as a complement – or in some cases an alternative – to null hypothesis testing using *P* values.

Bayesian statistics: updating knowledge with probability distributions

Bayesian statistics are another alternative to traditional null hypothesis testing using *P* values, and in fact have much in common with model selection using an information-theoretic framework (Hilborn & Mangel 1997). In essence, Bayesian methods formally incorporate information that has already been acquired (or presumed) to establish a **prior probability** that an $\text{idea}_{\text{mind}}$ is true; new data update those prior probabilities, leading to a **posterior** probability that a model is true, given the data. This becomes the revised current opinion, to be again modified as more data are collected.

Bayesian statistics trace back to a short memoir published posthumously in 1763 by a preacher and hobby-mathematician named Thomas Bayes (Bayes 1763). Bayes' Theorem quantifies how new evidence changes the probability that the existing belief is correct (see Ellison 1996, Hilborn & Mangel 1997, Wade 2000). The pieces are as follows.

[13] A quotable quote from Burnham and Anderson (2002:64): " 'Truth' is elusive; model selection tells us what inferences the data support, not what full reality might be."

- $P(H|D)$ The posterior probability of the hypothesis being true (or of obtaining the specified parameter, such as survival or abundance), given the data at hand.
- $P(H)$ The prior probability before the experiment is conducted or data collected. This is your initial estimate of the weight of evidence in favor of the hypothesis.
- $P(D|H)$ The likelihood function for the data, given that the hypothesis is correct. This is the same as the likelihood $\ln[L(\hat{\theta})|\text{data}]$ for AIC in eqn. 2.1.
- $P(D)$ The averaged probability of the data across all hypotheses.

And Bayes' Theorem is

$$P(H|D) = P(D|H) * \frac{P(H)}{P(D)} \tag{2.3}$$

The denominator is basically a scaling constant that can be factored out to give a simple statement of Bayes' Theorem:

$$P(H|D) \propto P(D|H) * P(H) \tag{2.4}$$

This says that the posterior probability for a hypothesis or parameter is proportional to its prior probability multiplied by the degree to which the hypothesis explains the data. Or, what we think now depends on what we thought before, modified by the insight we just got from the new data.

Bayes' Theorem can be extended to assess the relative probabilities of multiple working hypotheses. For example, for two hypotheses the posterior probabilities in favor of each hypothesis would be as follows. For hypothesis 1:

$$P(H_1|D) = \frac{P(H_1)P(D|H_1)}{P(H_2)P(D|H_2) + P(H_1)P(D|H_1)} \tag{2.5a}$$

For hypothesis 2:

$$P(H_2|D) = \frac{P(H_2)P(D|H_2)}{P(H_2)P(D|H_2) + P(H_1)P(D|H_1)} \tag{2.5b}$$

Although there are lots of sophisticated ecological examples of Bayesian analysis (see Ellison 2004), for the purpose of distilling the basic approach here is a simple example (Phillips 1973). Suppose an unscrupulous gambler carries around a biased coin (it favors heads slightly, with a 60% chance of coming up heads and 40% tails[14]), but after getting some change he is suddenly not sure whether a coin in his pocket is fair or biased. How does he decide? First he embraces his uncertainty by setting the probability of either hypothesis (biased or fair coin) as equal:

H_1: fair coin; $P(H_1) = 0.5$

H_2: biased coin; $P(H_2) = 0.5$

[14]In case the terms **heads** and **tails** are unfamiliar to you, those are just words we use in the USA to refer to the two sides of a coin.

Now he needs some data. He tosses the coin twice, realizing that the biased coin is more likely to be heads (H). He gets two heads. The likelihoods of two heads with the fair or biased coins are:

$$P(D|H_1) = \text{Probability of H and H given a fair coin} = 0.5 * 0.5 = 0.25$$

$$P(D|H_2) = \text{Probability of H and H given a biased coin} = 0.6 * 0.6 = 0.36$$

Plugging the prior probabilities and the likelihoods from the data into Bayes' Theorem (eqn. 2.5), he gets the following posterior probabilities:

$$P(H_1) = 0.41$$

$$P(H_2) = 0.59$$

Our gambler is now 59% sure that he holds his beloved biased coin. But he wants to be more sure. The posterior probabilities become the prior probabilities, and he continues to flip: he gets TH, HH, HH, and TH. Now, after 10 flips, eight of which came out heads, the gambler is 73% sure that he holds the biased coin. Depending on how sure he wants to be, he can keep going, or just pocket the coin and look for his next gambling victim.

Although the complexities of real-world Bayesian analyses has hampered their application, ever-growing computing power and ability to perform simulations are increasing their application[15]. Proponents of Bayesian approaches in ecology note that probability distributions and their uncertainty are simple to understand, directly show biological relevance, and easily allow comparison of different models. Whereas there are certainly detractors and controversies (e.g. Dennis 1996, Taper & Lele 2004), it seems clear that, like model selection using AIC, Bayesian approaches are established as a viable alternative to traditional *P* value-based hypothesis testing for some applications in wildlife population biology.

Ethics and the wildlife population biologist

It may seem a bit odd to have an ethics section in a book chapter on how we know what we know in wildlife population biology. But think about it: ethical transgressions in the pursuit of knowledge make everything else irrelevant. No amount of statistical rigor or experimental elegance can undo damage made by actions that violate ethical standards.

I will not dwell on the ethical obligations of biologists to address applied issues, or on the implications of the **land ethic** (eg Pister 1999, Leopold 2004). Instead, I will simply reiterate that at its core, ethics has to do with the standard of behavior within

[15]Currently, for Bayesian statistics ecologists use software such as WinBUGS (www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

our profession. Following those ethical norms guides the integrity of our inference from idea through collected data, all the way to conclusion and application of a study to real-world wildlife population issues.

Most of the scientific societies in our discipline, including the Ecological Society of America, The Wildlife Society, and the Society for Conservation Biology, have codes of ethics. Every student of applied population biology should know about them. Jack Ward Thomas (1986) has paraphrased, in simple yet eloquent terms, the code of ethics for The Wildlife Society:

> **1** *Tell folks that your prime responsibility is to the public interest, the wildlife resource and the environment.*
> **2** *Don't perform professional services for anybody whose sole or primary intent is to damage the wildlife resource.*
> **3** *Work hard.*
> **4** *Don't agree to perform tasks for which you aren't qualified.*
> **5** *Don't reveal confidential information about your employer's business.*
> **6** *Don't brag about your abilities.*
> **7** *Don't take or offer bribes.*
> **8** *Uphold the dignity and integrity of your profession.*
> **9** *Respect the competence, judgment and authority of other professionals.*
>
> *Implied but not specifically mentioned is the requirement simply to tell the truth . . . Tell the truth, all the truth, all the time. It's the right thing, the healthy thing, the professional thing to do.*
>
> **(Thomas 1986)**

The rules are simple, but can be hard to follow in a complicated world. Thomas' last comments about truth, in particular, can be difficult when an applied biologist feels outgunned, outspent, or out-politicked. It boils down to the question: "Irrespective of the righteousness of the cause, is distortion of the truth ever permissible?" (Erman & Pister 1989). The short answer, which goes to the heart of the integrity of our profession, is no.

A few years ago I learned first-hand the brutal consequences that can occur when these simple ethical guidelines are violated in a wildlife study. In the USA, Canada lynx became a species of special concern to land managers in the late 1990s, and was listed in March 2000 as a federally threatened species in the contiguous USA. At the time of listing it was not known precisely where lynx occurred. To provide a basis for subsequent monitoring, Kevin McKelvey of the US Forest Service (USFS) Rocky Mountain Research Station and I led a project called the National Lynx Survey to provide a consistent, standardized, reliable process for determining the current range of lynx in the USA. We designed the National Lynx Survey around non-invasive genetic sampling using hair rub pads (to be described more in Chapter 3).

Before initiating the study we carefully developed, validated, and exposed to peer-review a DNA-based species-identification protocol, using both blind and widespread geographic range tests (Mills et al. 2000a, Mills 2002). Although the collection of samples across 16 states in the northern USA was administered through the USFS

infrastructure, people from several agencies participated; approximately 800–1000 field helpers deployed and collected the hair pads, then sent them to my laboratory for analysis. Detailed protocols for collecting the hair samples – including discussion of the controls that we had used to develop the species-identification test – were included in all collection kits sent to field workers.

However, a handful of field workers in Oregon and southern Washington (where no actual lynx were detected) took it upon themselves to label some lynx samples as if they had been collected from the field (complete with slope, elevation, location, and vegetation types filled out on data forms) when really they were collected from captive or wall-mounted animals.

These mislabeled samples (which we correctly identified to species) would have been folded into our analysis of samples collected as part of the National Lynx Survey if not for a telephone call months later from one of those who mislabeled samples. When McKelvey and I found out about the mislabeling, our response was to deal with it internally, re-iterating to the hundreds of people collecting data on this project that internal controls were all in place and that the most important thing they could do was to ensure the integrity of samples coming from the field to the laboratory. But before we could do that, a political and media frenzy erupted. In December 2001, the *Washington Times* broke the story of mislabeled samples as a symptom of rampant fraud among applied biologists. Some in the US Congress followed up on the frenzy, saying that all actions on species protection should come into question. These are extreme interpretations to be sure, but in response others at the opposite end of the political spectrum – committed above all else to defending what they perceived as the higher goal of endangered species protection against attacks from Congress – espoused an equally extreme view that mislabeling samples had been an appropriate and even noble thing to do[16]. Three government investigations and one Congressional committee were launched to investigate this matter, and it was covered by dozens of journalists.

In my testimony to the US House of Representatives on March 6, 2002, I noted that we can never know the motivations for those who mislabeled samples. However, I stressed that, although there was no scientifically valid explanation for the mislabeling of samples, the actions of these few should not condemn the credibility of applied biology:

> *My experience throughout my career in working with hundreds of biologists and field personnel – including employees of USFWS, USFS, NPS, state wildlife departments, private groups, and several universities – is that they have exceptionally*

[16]I actually had one activist tell me that I should publicly announce that the field workers had good reason to mislabel samples as a test to expose incompetence because my laboratory was unreliable! He told me that if I did not help him make the argument that the field workers had done the right thing that I would be playing into the hands of those in Congress who wanted to bring down the Endangered Species Act. It horrified me that this person was seriously suggesting that the scientific process by which applied biologists contribute to important policy decisions should be twisted into a political tool.

> *high ethical standards in their pursuit of knowledge. Although inappropriate actions may occur on an individual and rare basis, my opinion is that these instances do not invalidate the larger body of biology, in the same way that inappropriate actions by a few physicians does not mean that we should shut down the practice of medicine.*"

In the end, the National Lynx Survey continued intact. However, there is no doubt that for the short term the credibility of our profession was damaged by "Lynxgate." Because wildlife and conservation biology are relatively young professions, credibility has been hard won on the backs of thousands of professional lifetimes (Thomas & Pletscher 2002). Increasingly, applied population biology work shows up on the front pages of newspapers, is heard in courtrooms, and is considered in the drafting of laws. Trust is everything to our continued relevance as applied biologists. Here are seven lessons from the mislabeled samples in the National Lynx Survey that should be considered action items for all wildlife biologists (quoted from Thomas & Pletscher 2002:1285):

1. Refresh our acquaintance with the ethical standards of our profession.
2. Assure adherence to those standards by bringing attention to actions that are inappropriate.
3. Condemn violations.
4. Consider every action by the standard of whether we would be proud to see it printed in the newspaper – because that is likely.
5. Understand that wildlife biologists now play a significant role in national affairs, and individual and collective actions will be considered in that light.
6. Know that in a mature profession with significant public trusts and responsibilities, there is simply no room or excuse for operating outside the rules of the game.
7. Recognize the responsibility – of teachers, agencies, and the profession – to state, formulate, teach, and continuously reinforce ethical standards and the need for transparent processes.

So, the very fact that all this happened convinced me that it is worth a page or two to remind all of us of our simple and perhaps obvious ethical obligations. Whether we are a scientist designing a study or a technician carrying out a study, our honesty is the bedrock on which lies all else in designing studies and interpreting population biology data.

Summary

By considering **truth** to be when an idea of the mind ($idea_{mind}$) matches a factual referent, it becomes apparent that applied population biology can only move forward by coupling creative ideas, reliable factual referents measured in wild populations, and a formal process to connect the $idea_{mind}$ and the facts. That is how reliable knowledge is gained.

Some of the prerequisites to obtaining trustworthy facts from the field include replication, randomization, and the use of controls. Accurate measurements have low bias and high precision. In the spirit of our **embrace uncertainty** mantra, variance may be more important than the mean. Process variance comes from spatial and temporal variance in nature, while sampling variance is an inevitable nuisance that arises anytime we sample from a population.

The strongest formal approach to connect facts to ideas of the mind is through the hypothetico-deductive approach. Natural history and observed correlations (induction) form an important basis for hypothesis generation, but specific *a priori* predictions should then be tested against data. This is the stage for care and forethought in developing the scientific question to be asked, and for specifying both the sampling strategy and how data will be interpreted.

The best way to distinguish among hypotheses is currently a matter of debate. The traditional approach of testing for statistical significance using *P* values has been criticized as being too reliant on an arbitrary threshold of $P = 0.05$. Also, there has often been too little consideration of statistical power – the probability of detecting an effect that is there – a real problem in studies where missing an effect could damage wildlife populations.

Some have argued that alternative frameworks for hypothesis testing should be implemented more widely. Model-selection frameworks using either AIC or Bayesian criteria are rapidly gaining footholds in wildlife population biology. In both cases, data inform the likelihood of particular models (or $\text{idea}_{\text{mind}}$) being best-suited to explain a particular set of data.

Finally, any rigor in study design or implementation is bereft without a strong foundation in ethics. Scientific societies in applied wildlife population ecology have ethics guidelines, and we should all be familiar with them. The integrity of our profession depends on it.

Further reading

Garton, E.O., Ratti, J.T., and Giudice, J.H. (2005) Research and experimental design. In: *Techniques for Wildlife Investigations and Management* (ed. C.E. Braun), pp. 43–71. The Wildlife Society, Bethesda, MD. A well-applied description of key aspects of research philosophy and experimental design.

Hilborn, R. and Mangel, M. (1997) *The Ecological Detective.* Princeton University Press, Princeton, NJ. A lucid overview, with examples, of advanced concepts in maximum likelihood, model selection, and Bayesian analysis.

Krebs, C.J. (1999) *Ecological Methodology*, 2nd edn. Benjamin Cummings, Menlo Park, CA. A much-admired classic that describes the fundamentals of the decisions and analyses made by practicing field ecologists.

Morrison, M.L., Block, W.M., Strickland, M.D., and Kendall, W.L. (2001) *Wildlife Study Design*. Springer-Verlag, New York. A fine compilation of study design with application to wildlife population and habitat ecology.