1

Research and Experimental Design

EDWARD O. GARTON,
JON S. HORNE,
JOCELYN L. AYCRIGG,
AND JOHN T. RATTI

INTRODUCTION

WILDLIFE MANAGEMENT PROGRAMS must be based on quality scientific investigations that produce objective, relevant information; and quality science is dependent upon carefully designed experiments, estimates, comparisons, and models. This chapter provides an overview of the fundamental concepts of wildlife research and study design and is a revision of Ratti and Garton (1994) and Garton et al. (2005).

Emergence of Rigor in Wildlife Science

Wildlife science is a term the wildlife profession has only recently nurtured. Our profession of wildlife conservation and management was built on natural history observations and conclusions from associations of wildlife population changes with environmental factors, such as weather, habitat loss, or harvest. Thus, we have a long tradition of wildlife management based on laws of association rather than on experimental tests of specific hypotheses (Romesburg 1981).

Although Romesburg (1981, 1989, 1991, 1993) and others (Steidl et al. 2000, Anderson 2001, Anderson et al. 2003, Belovsky et al. 2004) have been critical of wildlife science and its resulting management practices, the wildlife biologist is confronted with tremendous natural variation that might confound the results and conclusions of an investigation. Scientists conducting experiments in cell biology and biochemistry have the ability to control variables associated with an experiment, isolating the key components, and repeating these experiments under the same conditions to confirm their results. They also have the ability to systematically alter the nature or level of specific variables to examine cause and effect.

The wildlife scientist often conducts investigations in natural environments over large geographic areas, making it difficult to control potentially causal factors. Responses, such as density of the species in question, are simultaneously subject to the influences of factors, such as weather, habitat, predators, and competition, that may change spatially and temporally. Thus, rigorous scientific investigation in wildlife science is challenging and requires careful design (Steidl et al. 2000).

Experimental versus Descriptive Research

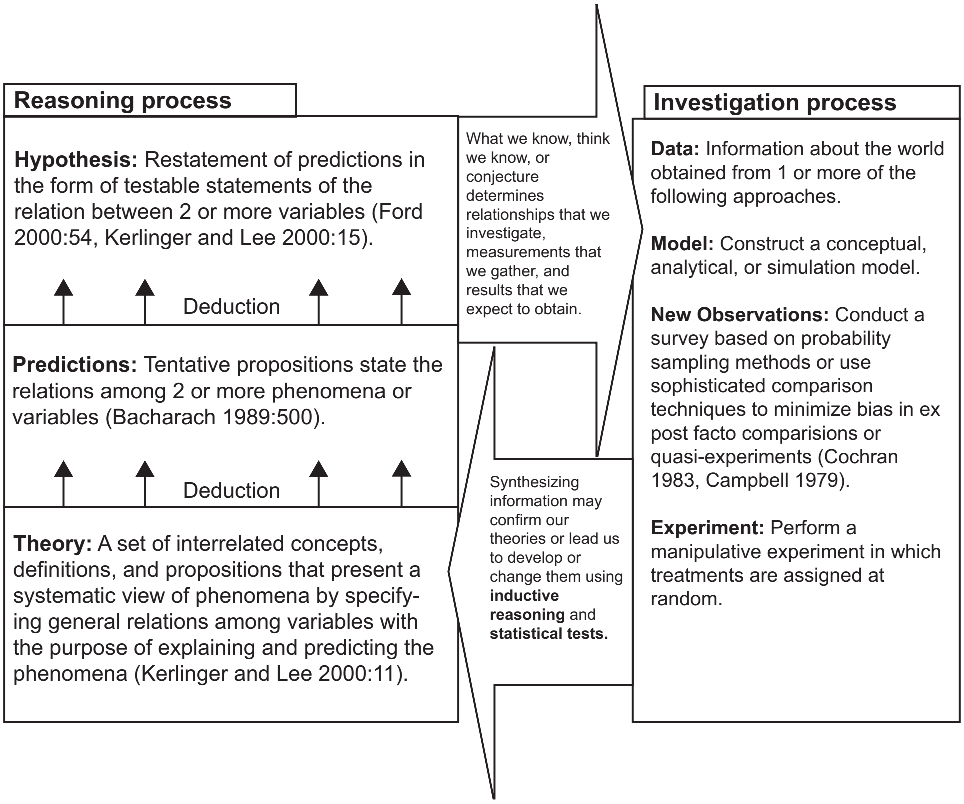
Most wildlife research prior to 1985 was descriptive. **Experimental research** is the most powerful tool for identifying cause and effect, and it should be used more in wildlife studies. However, descriptive natural history studies, field studies, and carefully designed comparisons based on probability sampling continue to be useful. **Descriptive research** is an essential initial phase of wildlife science and can produce answers to important questions, but it must be expanded to embrace interacting causes and variable results.

Descriptive research often involves broad objectives rather than tests of specific hypotheses. For example, we might have a goal to describe and analyze gray partridge (*Perdix perdix*) breeding ecology. Thus, we might measure characteristics of nesting habitat, clutch size, hatching success, brood use of habitat, food habits of chicks and adult hens, and mortality due to weather events and predators. From this information, we can learn details of gray partridge biology that will help us understand and manage the species. If we observe that 90% of gray partridge nests are in vegetation type "A," 10% in vegetation "B," with none in "C" and "D," we are tempted to manage for vegetation type "A" to increase nesting density. However, many alternatives must be investigated. Possibly vegetation type "A" is the best available habitat, but gray partridge experience high nest mortality in this type. Maybe vegetation type "X" is the best habitat for nesting, but it is not available on the study area. What vegetation types do gray partridge in other regions use? How does nest success and predation differ among regions with differing distributions of vegetation types, species of predators present, gray partridge densities, and climatic conditions? These questions show that defining quality nesting habitat is complex.

Combining descriptive studies with other studies published in the scientific literature should provide sufficient information to develop a **research hypothesis** (i.e., theory or conceptual model; Fig. 1.1) that attempts to explain the relationship between vegetation type and nesting success of gray partridge. Such models are general, but can help define specific predictions to be tested to examine validity of the model. These predictions can be stated as hypotheses. We can test hypotheses by gathering more descriptive observations or by conducting an experiment (Fig. 1.1) in which manipulated treatments are compared with controls (no treatment) to measure the change in sign (+ or –) resulting from experimental treatments. Random assignment of plots to treatment and control groups dramatically increases our certainty that measured differences are due to treatment effects rather than some ancillary factor.

Consider again the gray partridge study, and assume we have developed a **theory** (Fig. 1.1) that gray partridge adapted to be most successful at nesting in areas resembling their native habitat in Eurasia with its natural complement of predators, food sources, and vegetation cover. From this theory, we predict that partridge nesting success in grasslands in North America would be highest in undisturbed native prairie resembling native Eurasian gray partridge habitat and least successful in highly modified agricultural monocultures of corn, wheat, etc. We then formulate the hypothesis that gray partridge nesting density and nest success are higher in areas dominated (e.g., >75% of the available landscape) by native prairie than in areas dominated by cultivated fields of corn or wheat. The strongest test of this hypothesis we could perform would involve a **manipulative experiment** (Fig. 1.1), for which we must establish a series of control and experimental study plots. Our **study plots** would be randomly chosen from large blocks of land where agricultural practices have not changed in recent years, which contain the types of agricultural practices common to the region where we want to apply our findings. Some of these study plots (commonly 50%) will be randomly selected to act as control plots and will not change throughout the duration of the study. On the experimental plots (the remaining randomly selected plots in the same region as our control plots), cultivated fields will be planted to native prairie grass to test the validity of our hypothesis and predictions regarding the effect of habitat on gray partridge nesting. This process is difficult, because it requires large blocks of habitat, cooperation from landowners, several years to establish native prairie grass on the experimental plots, and additional years of study to measure the response of gray partridge to vegetative changes. The comparison between control and experimental plots will provide a basis to reject the null hypothesis of no effect, so we can conclude that increasing cover of native prairie grass, which could be within Conservation Reserve Program (CRP) fields in agricultural areas, will increase nesting density and success of gray partridge. If we fail to reject the null hypothesis, we cannot draw a firm conclusion, because the failure to reject might be due to insufficient sample size. If other studies have already shown higher nest success in areas of grass or CRP, then we must move beyond the potentially silly null hypothesis of no effect of grass cover (Johnson 1999, Läärä 2009). Instead we should focus on estimating the magnitude of effects from management efforts directed at gray partridge nesting success, so that we can build predictive models widely applicable to gray partridge management.

Some questions concerning wildlife science are not amenable to experimentation (e.g., effects of weather on populations, or differences in survival rates between gender or age classes). Other potential treatment effects are too expensive or difficult to accomplish. Some treatments may require substantial effort to convince the interested public of the value of applying them in any single treatment area. Finally, the need to evaluate effects of many habitat or population factors simultaneously may preclude experimentation. In these cases, construction of multiple biologically plausible models that seek to explain or predict observable phenomena can be a powerful tool for advancing knowledge (Hilborn and Mangel 1997) when combined with new information theoretic tools designed to identify the most likely explanatory model (Burnham and Anderson 2002). Incorporating modeling into the management process is an effective



Hypothesis: Restatement of predictions in the form of testable statements of the relation between 2 or more variables (Ford 2000:54, Kerlinger and Lee 2000:15).

Predictions: Tentative propositions state the relations among 2 or more phenomena or variables (Bacharach 1989:500).

Theory: A set of interrelated concepts, definitions, and propositions that present a systematic view of phenomena by specifying general relations among variables with the purpose of explaining and predicting the phenomena (Kerlinger and Lee 2000:11).

What we know, think we know, or conjecture determines relationships that we investigate, measurements that we gather, and results that we expect to obtain.

Synthesizing information may confirm our theories or lead us to develop or change them using **inductive reasoning** and **statistical tests.**

Data: Information about the world obtained from 1 or more of the following approaches.

Model: Construct a conceptual, analytical, or simulation model.

New Observations: Conduct a survey based on probability sampling methods or use sophisticated comparison techniques to minimize bias in ex post facto comparisions or quasi-experiments (Cochran 1983, Campbell 1979).

Experiment: Perform a manipulative experiment in which treatments are assigned at random.

Fig. 1.1. Iterative nature of the scientific method. Data are synthesized inductively to develop theories. These theories form the basis for deductively derived predictions and hypotheses that can be tested empirically by gathering new data with experiments, new observations, or models. *Modified from Ford (2000:6).*

strategy for predicting consequences of management actions while simultaneously learning about key processes affecting wildlife populations and their habitats (Walters 1986). A key requirement for this process to be successful is the need to monitor consequences of management actions through an **adaptive management process** (Walters 1986). This adaptive learning process might be facilitated by application of **Bayesian statistics,** which use additional observations to improve estimates of key relationships assumed prior to the management action (Hilborn and Mangel 1997, Gelman et al. 2003, Bolstad 2007).

Scientific Method

In one of the early papers published on the scientific method in *Science* in 1890 Chamberlin (republished in1965) emphasized the need to examine multiple working hypotheses to explain an observation. Popper (1959, 1968) formalized an approach to testing individual hypotheses, referred to as the **hypothetico-deductive method,** that became the accepted standard in science. The method is a circular process in that previous information is synthesized into a theory; predictions are deduced from the theory; the predictions are stated explicitly in the form of hypotheses; hypotheses are tested through an investigation involving experimentation, observation, models, or a combination of these; the theory is supported, modified, or expanded on the basis of the results of these tests; and the process starts again (Fig. 1.1). Platt (1964) re-emphasized the importance of **multiple competing hypotheses** and proposed a systematic pattern of inquiry, referred to as **strong inference,** in which the investigator devises alternate hypotheses, develops an experimental design to reject as many hypotheses as possible, conducts the experiment to achieve unambiguous results, and repeats the procedure on the remaining hypotheses. Other major works that provide detailed discussions of the scientific method include Dewey (1938), Bunge (1967), Newton-Smith (1981), Ford (2000), and Gauch (2003).

The most successful applications of the hypothetico-deductive method have been in physics, chemistry, and molecular biology, where experiments can isolate the results from all but a small number of potentially causal factors. The classic methods of natural history observation in wildlife science and other natural sciences have expanded to include experimentation, hypothesis testing, and quantitative modeling. James and McCulloch (1985:1) described this transition for avian biologists: "traditional ornithologists accumulated facts, but did not make generalizations or formulate causal hypotheses . . . modern ornithologists formulate hypotheses, make predictions, check the predictions with new data sets, perform experiments, and do statistical tests." Measuring simultaneous effects of multiple interacting causes (Quinn and Dunham 1983) may be facilitated by application of **information theoretic tools** to models incorporating multiple causes (Burnham and Anderson 2002). In addition

Box 1.1. Systematic outline of sequential events in scientific research with an example of elk in the northern Rocky Mountains

1. Identify the research problem.	What are the influences of environmental factors, such as wildfire and winter severity, on the carrying capacity of elk winter range?
2. Conduct literature review of relevant topics.	Excellent earlier work by Houston (1982), Merrill and Boyce (1991), DelGiudice (1995), Coughenour and Singer (1996).
3. Identify broad and basic research objectives.	(a) Determine temporal and spatial differences in food habits that may affect elk nutritional condition during winters of varying severity; (b) examine the relationship between energy intake and mobilization of energy reserves at the population level throughout winter.
4. Collect preliminary observations and data as necessary.	(a) Winter severity data for 1987–1988, 1988– 1989, and 1989–1990 including snow depth; (b) monthly precipitation during 1988 reflecting 100-year drought; (c) wet summers contributed to increases in elk population; (d) substantial winter kill first post-fire winter.
5. Conduct exploratory data analysis.	(a) Analyze food habits data for 2 different spatial locations pre-fire; (b) estimate energy intake by elk pre-fire.
6. Formulate a theory (conceptual model or research hypothesis).	Carrying capacity of elk winter range is influenced by wildfire and winter severity.
7. Formulate predictions from conceptual model as testable hypotheses (Fig. 1.1).	(a) Carrying capacity of elk winter range increases in post-fire areas; (b) carrying capacity of elk winter range decreases with increasing winter severity.
8. Design research and methodology for each hypothesis.	(a) Collect samples of urine during the same month of each winter to assess nutritional condition of elk from each study area. Only include urine samples from cows and calves. Collect samples in both burned and unburned areas. (b) Construct simulation model to translate individual responses to nutritional condition to population level responses.
9. Conduct a pilot study to test methodologies and estimate costs and variances.	Pilot study collects urine samples and estimates costs and variances.

to James and McCulloch (1985), other excellent reviews of scientific approaches applicable to natural systems include Romesburg (1981), Diamond (1986), Eberhardt and Thomas (1991), Murphy and Noon (1991), Sinclair (1991), Hilborn and Mangel (1997), Boitani and Fuller (2000), Williams et al. (2002*a*), and Morrison et al. (2008).

The first steps in the **scientific method** begin with a clear statement of the research problem (Box 1.1), followed by a careful review of literature on the topic and preliminary observations or data collection. Preliminary data can be combined with published data to conduct an exploratory data analysis (Tukey 1977). Established theory, including principles, concepts and widely accepted models (Pickett et al. 2007), should be combined with creative ideas and potential relationships resulting from the biologist's observations and exploratory data analysis to develop a **conceptual model** (i.e., theoretical framework or general research hypothesis, Andrienko and Andrienko 2006). This conceptual model is essentially a broad theory (Fig. 1.1) that offers explanations and possible solutions, and places the problem in a broader context (Box 1.1). The next step is to develop **predictions** from the conceptual model (i.e., statements that would be true if the conceptual model were true). These predictions are then stated as **multiple testable hypotheses.** Research should be designed to test these hypotheses; ideally experimentation should be used when possible. A **pilot test** at this

10. Estimate required sample sizes and anticipate analysis procedures with assistance from a statistical consultant.	Estimated sample sizes feasible and analysis procedures successful with pilot survey data.
11. Prepare written research proposal that reviews the problem, objectives, hypotheses, methodology, and procedures for data analysis.	Prepare written proposal: Combine steps 1, 3, 6, and 8 to provide background, justification, and methodology for research.
12. Obtain peer review of the research proposal from experts on the research topic and revise if necessary.	Seek out experts in state wildlife agencies as well as authors of papers found during literature review.
13. Perform experiments, collect observational data, or construct a model.	(a) Collected elk urine samples from each winter and each study area; (b) constructed model to simulate energy intake and movements for the elk population.
14. Conduct data analysis.	(a) Non-normally distributed urine sample data analyzed using nonparametric statistics; (b) measured and simulated nutritional conditions compared using urine samples with unpaired *t*-tests.
15. Evaluate, interpret, and draw conclusions from the data.	Combined use of urine samples and model simulations provided strategic approach for assessing subtle changes in nutritional condition, physical condition, and mortality rates of elk. During winter 1988–1990, snow depth had a pronounced impact on nutritional condition; the most dramatic temporal and spatial effects occurred during the most severe winter in 1989.
16. Speculate on results and formulate new hypotheses.	Carrying capacity of elk winter range influenced more by winter severity than by wildfire.
17. Submit manuscript describing the research for peer-reviewed journal publication, agency publication, and/or presentation at scientific meetings.	Combine steps 9, 11, 12, 13, and 14 to create a well-written and concise manuscript of research findings, which were published in this case as DelGiudice et al. (2001*b*).
18. Repeat the process with new hypotheses (starting at step 6 or 7).	Repeat process with new hypotheses.

Based on DelGiudice et al. (2001*b*).

stage is invaluable in testing methodologies and gathering estimates of cost and variances. Included in the design, with the assistance of a statistician, is calculation of the **sample sizes** required to detect the hypothesized effects as well as decisions about how the data will be analyzed. Peers and a statistician should **review** the proposed design before data collection begins. The data are collected using **quality control.** Data **analysis** with appropriate statistical procedures is conducted to test the theory by rejecting fallacious hypotheses, selecting the best models of relationships or differences, obtaining unbiased estimates, or selecting the best alternative. Final conclusions usually result in further speculation, modification of the original conceptual model and hypotheses, and formulation of new hypotheses. The **publication** process is the last, but essential, step, and peer-review comments should be considered carefully before research on new hypotheses is designed.

PHILOSOPHICAL FOUNDATION

Why should wildlife biologists and managers care about the seemingly endless esoteric debates by philosophers of science? One reason is that modern philosophers have reached a perspective on how to gain truth and knowledge that is consistent with the approach of practicing wildlife biologists, managers, and scientists. Modern philosophers assert that classic views of the scientific process are outmoded or inappropriate and propose replacing them with a new integrated approach directly applicable to wildlife science and ecology (Pickett et al. 2007). Their approach is founded on 3 beliefs inherent in **scientific realism** (Boyd 1992). First, the universe is real, and it is possible to gain true knowledge about it (Scheiner 1994). Second, knowledge includes ideas that we posit in theories, but can only sense indirectly (e.g., electrons, plant communities, and carrying capacities). Third, all such theories must ultimately be tested empirically (Scheiner 1994). The goal of wildlife research and experimental design must be to advance our **knowledge** by gathering new information to test and improve our evolving wildlife theory, which consists of a set of interrelated **concepts, definitions,** and **propositions** (i.e., models and confirmed generalizations often referred to as principles).

The scientific method consists of an efficient approach to expanding this evolving knowledge base. This expansion can be accomplished by gathering new observations to obtain unbiased estimates of important characteristics (e.g., age-specific survival rates), testing proposed theories (e.g., harvest and starvation of subadults are compensatory), inferring new patterns or processes (e.g., harvest is additive to cougar mortality in adult elk [*Cervus canadensis;* Polziehn and Strobeck 2002]), and restricting or expanding the domain of inference for models of patterns or processes (e.g., deeper snows decrease winter survival of elk and deer, but thresholds for the effects differ among species). This **integrated approach** estimates strength of contributions (Quinn and Dunham 1983) by multiple simultaneously acting causes (e.g., survival of elk calves depends on date of birth; milk production of cows; quality and quantity of hiding cover; and density of bears, cougars, and wolves) rather than attempting to falsify all but one causal mechanism (Platt 1964).

INITIAL STEPS

Problem Identification

The initial step in most wildlife research is **problem identification** (Box 1.1). Most research is either applied or basic. **Applied research** usually is related to management concerns, political controversy, or public demand. For example, we may study specific populations because the hunting public has demanded greater hunting opportunity or a nongame species decline raises concerns about its long-term survival. Other applied studies may be politically supported due to projected loss of habitat by development or concerns over environmental problems, such as contamination from agricultural or industrial chemicals. **Basic research** seeks to gain knowledge for the sake of knowledge and a more complete understanding of factors that affect behavior, reproduction, density, competition, mortality, habitat use, and population fluctuations. Research on management questions can often be designed so that basic research on underlying principles can be conducted for minimal extra cost as data are gathered to solve the immediate management problem.

Literature Review

Once a research problem has been identified, research should begin with a thorough **literature review,** including collecting published and unpublished management agency data. Searching Google Scholar (http://www.scholar.google.com) and other free online databases provides instant access to titles with links to abstracts and often the full text of published peer-reviewed literature. Membership in The Wildlife Society and other professional organizations (Ecological Society of America, Society for Conservation Biology, American Fisheries Society, etc.) as well as many public libraries provide access to full-text databases of every paper published in societies' refereed journals and monographs. Broadscale Internet searches using Google and other search engines may provide unpublished information of value from public agencies and institutions, but information posted by individuals or unknown organizations should be treated with substantial skepticism. Using a variety of sources for your literature review will ensure that you have compiled the most relevant and recent information pertaining to your objectives.

Biological, Political, and Research Populations

Wildlife professionals work with 3 types of populations that impact study design: biological, political, and research populations. Mayr (1970:424) defined a **biological population** as a group "of potentially interbreeding individuals at a given locality" and **species** as "a reproductively isolated aggregate of interbreeding populations." Thus, a **population** is an aggregation of individuals of the same species that occupies a specific locality at a particular time, and often the boundaries can be described with accuracy. For example, the dusky Canada goose (*Branta canadensis*) population breeds in a relatively small area on the Copper River delta of Alaska and winters in the Willamette Valley near Corvallis, Oregon (Chapman et al. 1969). Between the breeding and wintering grounds of the dusky Canada goose is the more restricted range of the relatively nonmigratory Vancouver Canada goose (Ratti and Timm 1979). Although these 2 populations are contiguous with no physical barriers between their boundaries, they remain reproductively isolated and independent.

For most populations, such as red-winged blackbirds (*Aegolius phoeniceus*), grouping individuals into a hierarchical organization of demes, populations, and metapopulations within the species may require careful consideration of 5 facets (Fig. 1.2, Box 1.2): (1) geographical distribution of individuals, (2) geographical distribution of habitats (resources), (3) correlations in demographic rates (Bjørnstad et al. 1999, Post and Forchhammer 2002, Palsbøll et al. 2006), (4) genetic relationships (Manel et al. 2005), and (5) patterns of movement. Identifying the appropriate level in this hierarchy to

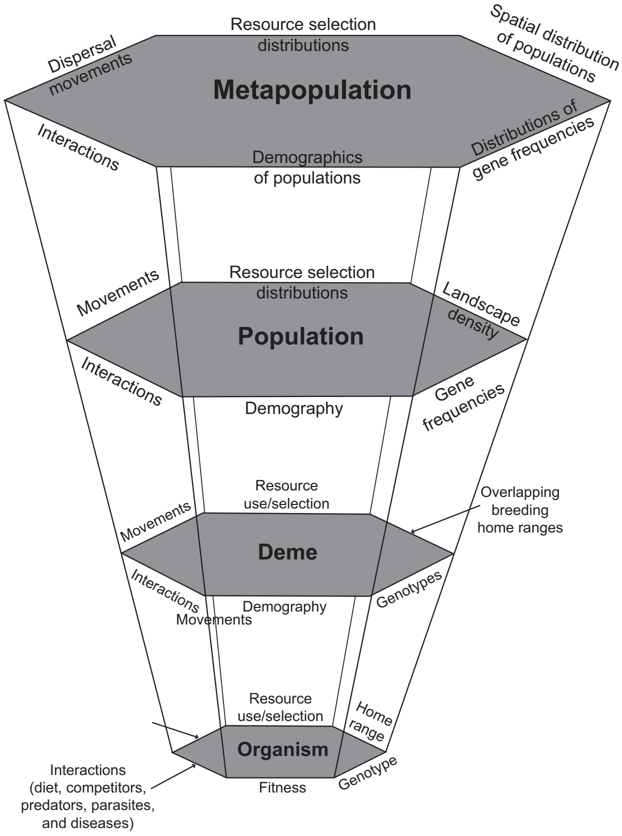


Fig. 1.2. Hierarchical arrangement of individuals from organism to metapopulation, illustrating multiple facets to consider in delineating levels to study: demography, genetics, geographical distribution of individuals, distribution and selection of resources, patterns of movement and interactions (e.g., diet, competitors, predators, parasites, and diseases). Processes operating at 1 level in this ecological hierarchy are influenced by the processes and characteristics at both lower levels (i.e., mechanisms) and higher levels in the ecological hierarchy (i.e., context). *Modified from Pickett et al. (2007:29).*

sample or assign treatments is critical to obtaining precise estimates and performing valid, powerful tests of ideas (i.e., theory), but keep in mind that processes operating at one level are influenced by those occurring at both lower levels in the hierarchy (i.e., mechanisms) and higher levels (i.e., context). Choosing the level in the biological hierarchy to study (Box 1.2) defines the research population or domain (Pickett et al. 2007) to which inferences and conclusions apply.

Beletsky and Orians (1996:152) and refuge biologists studied red-winged blackbirds at Columbia National Wildlife Refuge, Othello, Washington, and demonstrated, with 20 years of banding data, that territorial males and associated females occupying a set of discrete patches of marsh vegetation associated with ponds or streams on the refuge constituted a **deme** (Fig. 1.3). High correlations in demographic rates among demes and genetic similarity due to dispersal among demes make the entire red-winged blackbird population on the refuge an appropriate biological population for management (Beletsky 1996, Garton 2002:665). Surrounding irrigated farmlands isolate red-winged blackbird populations at refuges from one another to some degree, but populations at refuges throughout the Columbia Basin could be treated as a **metapopulation** within the subspecies (*A. phoeniceus nevadensis;* Fig. 1.3). Another example of biological populations with separate boundaries is the bison (*Bison bison*) populations in Yellowstone National Park in the northwestern United States (Olexa and Gogan 2005). Biological populations for other species may not be so geographically distinct as those for Canada geese, red-winged blackbirds, and Yellowstone bison, in which case the researcher will have to carefully consider from which biological aggregation their samples are selected and to which their findings will apply.

The **political population** has artificial constraints of political boundaries, such as county, state, or international entities. For example, a white-tailed deer (*Odocoileus virginianus*) population in an intensively farmed agricultural region in the Midwest might be closely associated with a river drainage system due to permanent riparian cover and food critical for winter survival. The biological population may extend the entire length of the river drainage, but if the river flows through 2 states, the biological population is often split into 2 political populations that are subjected to different management strategies and harvest regulations. Traditionally, this problem has been common in wildlife management. When biological populations have a political split, it is best to initiate cooperative studies, in which research personnel and funding resources can be pooled to benefit both interested agencies.

Ideally, the **research or statistical population** should conform closely to the biological population, so that inferences can be applied to the chosen biological population. Due to logistical constraints, we often take a sample from this research population (i.e., sample frame; Scheaffer et al. 2005). Thus, sampling methodology is critical, for it provides the only link between samples and the research population. In rare instances, a population may be studied that represents all individuals of a species (e.g., an endangered species with few individuals, such as whooping cranes [*Grus americana*]). Or the research population might represent an entire biological population, such as one of the bison herds in Yellowstone National Park (Olexa and Gogan 2005). However, the research population usually is only a portion of the biological population and a small segment of the species. Carefully specifying a research or statistical population is essential in the planning phase of an investigation and may require thorough investigation of the existing literature on the species to determine breeding biology and dispersal patterns, geographic sampling to identify distribution of individuals and resources, and reviews of the literature on biological aggregations (Mayr 1970, Selander 1971, Stebbins

Box 1.2. Hierarchy of spatial population units

Deme	The smallest grouping of individuals approximating random breeding within the constraints of the breeding system, where it is reasonable to estimate birth, death, immigration, and emigration rates. Animals in this grouping are ideally distributed continuously across one patch of homogeneous to heterogeneous habitat, and their movements are restricted to home ranges for breeders during the breeding season. The size of this patch ideally would be related to the dispersal distance of juveniles or perhaps equal an area 20–50 times the size of a female breeding home range (e.g., Fig. 1.3 and Garton [2002] for red-winged blackbirds). Note: for some species demes are not feasible to delineate because of complex mating patterns and movements (e.g., in mallards, *Anas platyrhynchos,* males and females form pair bonds on wintering areas and males follow females to nesting areas the following spring, which may be quite distant from their natal area; Bellrose 1976:236).
Population	A collection of demes or individuals at one point in time, typically the breeding season, with strong connections demographically (very high correlations in vital rates), geographically (close proximity), genetically (Manel et al. 2005), and through frequent dispersal. The population occupies a collection of habitat patches (relative to dispersal distance) without large areas of nonhabitat intervening. The area is typically 100 times the size of an average female home range and is not larger than the dispersal distance of 95% of natal dispersers, but it may be much larger if habitat patches are linear in shape and widely dispersed (e.g., all red-winged blackbirds occupying Columbia National Wildlife Refuge during the breeding season might be reasonably treated as a population; Garton 2002; Fig. 1.3). A population is dynamic through time: demes or groups of individuals show correlated fluctuations associated with the effects of broad-scale environmental factors (e.g., weather and fires) or other populations (e.g., competitors, predators, and disease outbreaks).
Metapopulation	A collection of populations sufficiently close together that dispersing individuals from source populations occasionally colonize empty habitat resulting from local population extinction (Levins 1969). Populations in a single metapopulation may show low or high correlations in demographic rates, but the low rates of dispersal are sufficient to maintain substantial genetic similarity (e.g., red-winged blackbird populations distributed among the 7 national wildlife refuges along 200 km of the Columbia River in the south-central part of Washington constitute a metapopulation; Garton 2002; Fig. 1.3). Numerous types of metapopulations have been described, from source-sink to nonequilibrium to classic (or Levins) metapopulations (Harrison and Taylor 1997).
Subspecies	A collection of populations as well as metapopulations, if present, in a geographic region where very rare dispersals maintain genetic, morphological, and behavioral similarity. However, populations and metapopulations occupy habitat patches that may be separated by large areas of nonhabitat, resulting in substantial demographic independence among populations or metapopulations (Mayr 1982, Garton 2002; Fig. 1.3).
Species	The collection of interbreeding populations as well as metapopulations and subspecies, if present, encompassing the entire distribution and geographic range of the populations. The populations may show substantial differences in phenotypes (vegetation association, physiology, and behavior) and genotypes (Garton 2002; Fig. 1.3).

Modified from Garton (2002).

1971, Ratti 1980, Wells and Richmond 1995, Garton 2002, Hanski and Gaggiotti 2004, Cronin 2006).

Conclusions from research are directly applicable only to the research population from which the samples were drawn. However, biologists usually have goals to obtain knowledge and solve problems regarding biological populations and species. The key questions are: (1) Is the sample an unbiased representation of the research population? (2) Is the research population an unbiased representation of the biological population? (3) Is the biological population representative

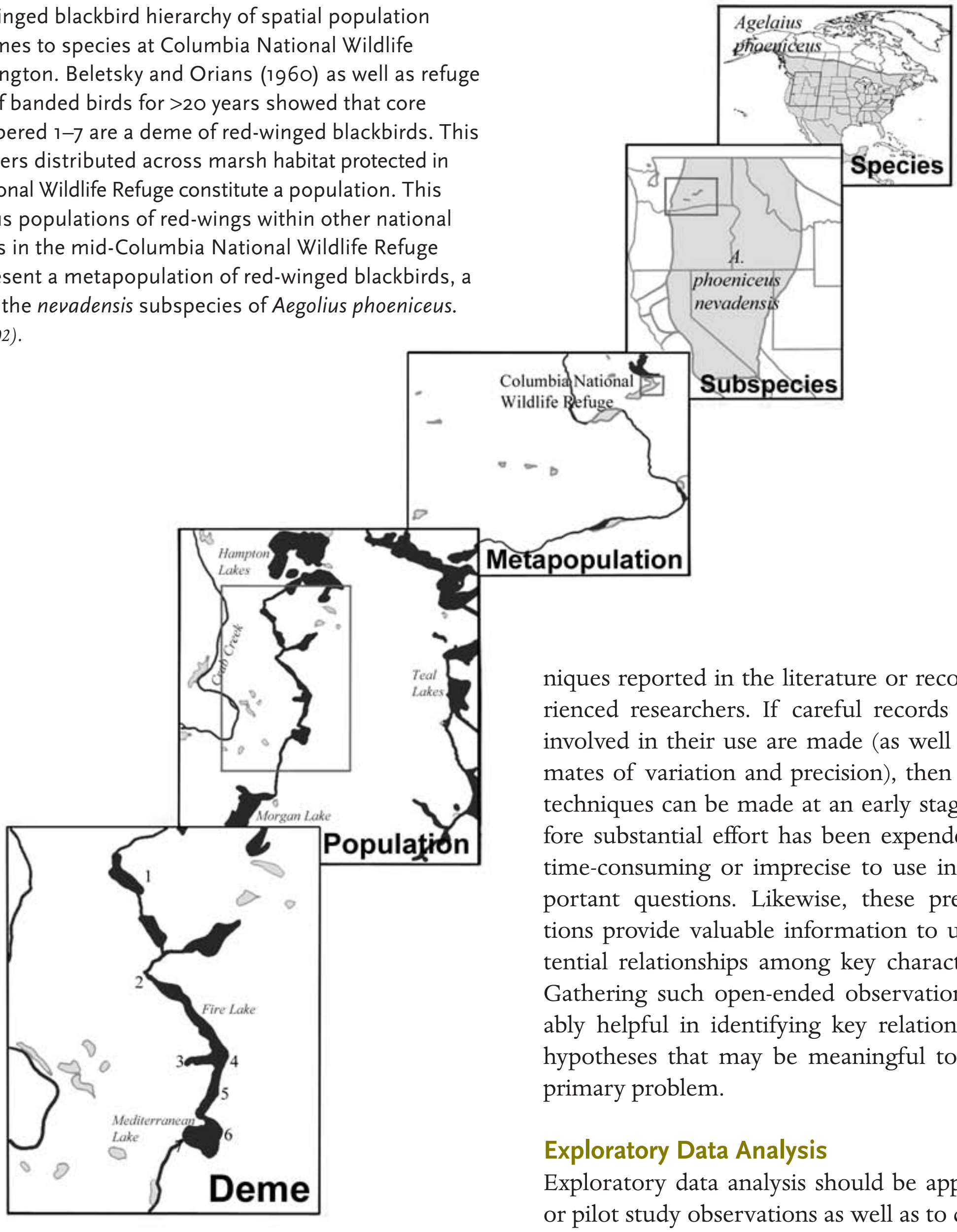


Fig. 1.3. Red-winged blackbird hierarchy of spatial population units from demes to species at Columbia National Wildlife Refuge, Washington. Beletsky and Orians (1960) as well as refuge staff studies of banded birds for >20 years showed that core marshes numbered 1–7 are a deme of red-winged blackbirds. This deme plus others distributed across marsh habitat protected in Columbia National Wildlife Refuge constitute a population. This population plus populations of red-wings within other national wildlife refuges in the mid-Columbia National Wildlife Refuge Complex represent a metapopulation of red-winged blackbirds, a subdivision of the *nevadensis* subspecies of *Aegolius phoeniceus*. *After Garton (2002).*

of the species? Because traits among segments of biological populations (and among populations of a species) often differ, broad conclusions or inferences relative to a research hypothesis should be avoided until several projects from different populations and geographic locations provide similar results. Combining and synthesizing replicate studies across large spatial extents should be a long-term goal, but may require the use of new techniques, such as meta-analysis (Osenberg et al. 1999).

Preliminary Data Collection

Making an effort to gather preliminary observations at this stage can pay great dividends in the end by allowing the researcher to explore a variety of potential research techniques reported in the literature or recommended by experienced researchers. If careful records of time and effort involved in their use are made (as well as preliminary estimates of variation and precision), then optimal choices on techniques can be made at an early stage in the design, before substantial effort has been expended on methods too time-consuming or imprecise to use in answering the important questions. Likewise, these preliminary investigations provide valuable information to use in exploring potential relationships among key characteristics of interest. Gathering such open-ended observations also are remarkably helpful in identifying key relationships and alternate hypotheses that may be meaningful to understanding the primary problem.

Exploratory Data Analysis

Exploratory data analysis should be applied to preliminary or pilot study observations as well as to data from the literature or public agencies and institutions (Tukey 1977, James and McCulloch 1985, Andrienko and Andrienko 2006). During this process data are quantitatively analyzed in terms of means, medians, modes, standard deviations, and frequency distributions for important groups, and scatter plots of potential relationships are generated. Exploration of the data should be as complete and biologically meaningful as possible, which may include comparison of data categories (e.g., mean values, proportions, and ratios), multivariate analysis, correlation analysis, and regression. The "basic aim of exploratory data analysis is to look at patterns to see what the data indicate" (James and McCulloch 1985:21). If the research topic has received extensive previous investigation, the exploratory phase might even take the form of a meta-analysis of previous data gathered on the question (Osenberg et al. 1999). This phase often involves extensive discussions with other investigators with field or experimental experience on the topic.

THEORY, MODELS, PREDICTIONS, AND HYPOTHESES

Exploratory data analysis, literature reviews, and perceived associations should lead to the development of a theoretical framework (i.e., conceptual model; Fig. 1.4) of the problem. **Wildlife theories** (Fig. 1.1) are a set or system of interrelated concepts, definitions, assumptions, facts, confirmed generalizations, and propositions (Kerlinger and Lee 2000, Pickett et al. 2007) that present a structured view of wildlife ecology and management by specifying general relations among variables (e.g., waterfowl populations, annual rainfall, abundance of ponds and riparian habitat, and hunter harvest), with the purpose of explaining and predicting the phenomena (e.g., changes in waterfowl abundance; Office of Migratory Bird Management 1999, Ford 2000, Conroy et al. 2005).

We now explore the meaning and value of theory by considering our conceptual model of waterfowl population dynamics (Fig. 1.4), which expresses in a simple way complicated patterns of autumn waterfowl populations being positively influenced by spring breeding population size, number of ponds, and quantity and quality of wetland habitat, and negatively affected by nest predators, whose influence likely interacts with quality and quantity of wetland habitat around ponds. Likewise harvest influences spring population sizes the following year (i.e., $t + 1$ in Fig. 1.4), but the interaction may be complex, with either or both compensatory and additive effects coming into play. Utilizing this theory to understand dynamics of any particular waterfowl population requires stating a domain of interest and inference. For example, Conroy et al. (2005) studied an American black duck (*Anas rubripes*) metapopulation breeding in 3 regions and harvested in 6 regions in Canada and the United States (Box 1.2). Any individual investigation asks important questions and evaluates alternative hypotheses (e.g., models of harvest) in a restricted portion of the entire theory. For example, Conroy et al. (2005) used Bayesian methods to evaluate harvest models for American black ducks for this metapopulation. Often, important variables (e.g., abundance of nest predators) are very difficult to estimate, so their influence must be inferred through changes in nest success and fledging rates resulting from experimental manipulations (e.g., predator removal or manipulation of nesting cover).

Ford (2000:43) identifies 2 parts of a theory, consisting of (1) a working part providing information and a logical basis for making generalizations and (2) a motivational or speculative part that defines a general direction for investigation. Stating our theoretical framework (conceptual model) explicitly requires careful thinking and analysis of accepted generalizations (principles) stated in classic textbooks, reviews, and the published peer-reviewed literature on the topic. **Predictions** or deductive consequences of theory form the basis for hypotheses, which are variously described as assertions subject to verification (Dolby 1982, James and McCulloch 1985; Fig. 1.1) or testable statements derived from or representing various components of theory (Pickett et al. 2007:63; Box 1.3). Normally, the primary research hypothesis is what we initially consider to be the most likely explanation, but if the question has been placed into the proper theoretical framework, several alternate hypotheses are presented as possible explanations for observed facts (Fig. 1.1). Modern hypotheses commonly take the form of quantitative models that explicitly describe the relationships or magnitude of differences (Box 1.3).

We take an important step from descriptive natural history when we formulate conceptual models as research hypotheses. Interpretation of exploratory data analysis, creation of a theoretical framework, deduction of predicted

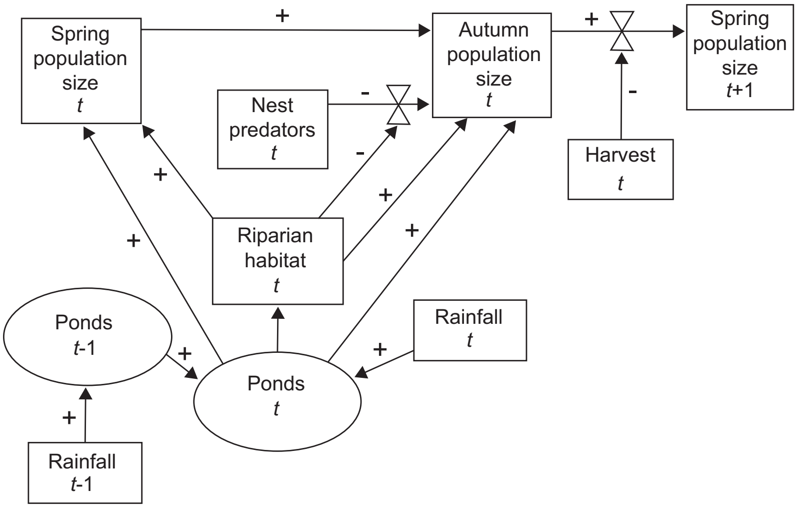


Fig. 1.4. Conceptual model of waterfowl population dynamics.

Box 1.3. Components of theory

Component	Example
Domain. The scope in space, time, and phenomena addressed by a theory.	An individual waterfowl population or metapopulation in North America (e.g., American black duck, *Anas rubripes*) during 1971–1994.
Assumptions. Conditions needed to build the theory.	Conroy et al. (2005) assumed survival and productivity rates estimated for 3 regions in Canada and harvest rates from 6 regions in Canada and the United States to determine the dynamics of this metapopulation.
Concepts. Labeled regularities in phenomena.	**Harvest** refers to waterfowl shot during a legal hunting season and retrieved by the hunter.
Definitions. Conventions and prescriptions necessary for the theory to work with clarity.	Conroy et al. (2005) defined harvest rate as the probability of harvest based on direct recoveries (hunter reports of banded birds shot or found dead in the hunting season immediately following release; Williams et al. 2002*a*).
Facts. Confirmable records of phenomena.	All data on harvest regulations (season length and bag limit) and hunter numbers for 1971–1994 were obtained from the Canadian Wildlife Service and U.S. Fish and Wildlife Service (Conroy et al. 2005).
Confirmed generalizations. Condensations and abstractions from a body of facts that have been tested or systematically observed.	Harvest rates of male and female waterfowl generally differ, and Conroy et al. (2005) estimated harvest rates for males only to eliminate the need for estimating sex-specific harvest rates.
Laws or principles. Conditional statements of relationship or causation, statements of identity, or statements of process that hold in a domain.	Better wetland habitat conditions positively influence productivity in waterfowl populations (Fig. 1.4).
Models. Conceptual constructs that represent or simplify the structure and interactions in the material world. (**Scientific models** can project consequences of ideas; **statistical models** draw inferences and discriminate among competing ideas based on limited observations).	Conroy et al. (2005) developed statistical models for harvest rates in American black ducks. They found that harvest rates depended on both season length and bag limit, but differed between years and areas during 1971–1994.
Translation. Procedures and concepts needed to move from the abstractions of a theory to the specifics of applications or test or vice versa.	Annual changes in wetland conditions are estimated from aerial strip transect counts of pond densities throughout waterfowl breeding areas in North America (U.S. Fish and Wildlife Service and Canadian Wildlife Service 1987).
Hypotheses. Testable statements derived from or representing various components of theory.	Harvest rates for American black ducks increase with season length and bag limits in an area (tested and confirmed by Conroy et al. 2005).
Framework. Nested causal or logical structure of a theory.	During the fall, groups of American black ducks join with other groups on the same wetlands and other nearby wetlands to form populations that join 3 other populations in Canada during their migration south; they form a metapopulation occupying 6 regions of Canada and the United States (Conroy et al. 2005; Fig. 1.2).

After Pickett et al. (2007:63).

consequences, and formulation of testable hypotheses as alternative models are difficult aspects of science that require creativity and careful reasoning, but they are essential to the future of wildlife science.

OVERVIEW OF STUDY DESIGN

Introduction

Many different study designs are available for answering questions about the biology and management of wildlife species (Eberhardt and Thomas 1991, Morrison et al. 2008; Fig. 1.5). These options differ dramatically in terms of 2 criteria: How certain are the conclusions reached? How widely applicable are the conclusions? No single option is perfect. The biologist must weigh the available options carefully to find the best choice that fits the constraints of time and resources. Here we provide an overview of the most prominent study designs with further explanation in subsequent sections.

Experiments consisting of manipulative trials are underused in wildlife science (Fig. 1.5). **Laboratory experiments,** in which most extraneous factors are controlled, provide the cleanest results with the most certainty, but results generally have only narrow inference to free-ranging wildlife populations, unless they concern basic processes (e.g., disease susceptibility or nutritional biology). **Natural experiments,** in which large-scale perturbations (e.g., wildfires, disease outbreaks, and hurricanes) affect populations and landscapes naturally, provide only weak conclusions because of lack of replication and inability to control extrinsic factors through random assignment of treatments (Diamond 1986, Underwood 1997, Layzer 2008, Diamond and Robinson 2010; Fig. 1.5). **Field experiments,** in which manipulative treatments are applied in the field, combine some of the advantages of laboratory and natural experiments (Hurlbert 1984, Scheiner and Gurevitch 2001; Fig. 1.5). They have singular advantages, because truly replicated field experiments combine both breadth of inference and relatively certain conclusions (Johnson 2002). By assigning treatments to field replicates randomly, we can be certain that conclusions are valid rather than resulting from extrinsic factors beyond our control.

Some questions of importance in wildlife biology and management are not appropriate for experimentation. For example, we may be interested in the effects of weather on a particular animal population, but we cannot manipulate weather at will, in spite of the apparent human impact on its long-term trajectory. In addition, we may be interested in the relative importance of such factors as predation, habitat, and food limitations on population changes (Quinn and Dunham 1983, Mills 2007). In these cases we should formulate primary and alternate hypotheses in the form of models, estimate their maximum likelihood parameters, and test them statistically with likelihood ratios or compare them with information criteria (Hilborn and Mangel 1997, Burnham and Anderson 2002). Case studies consisting of unreplicated natural history descriptions are most useful at early stages in development of the research process (Fig. 1.5). **Pseudoreplicated field studies,** in which replicates are not statistically independent or samples rather than treatments are replicated, are only slightly better than **descriptive natural history studies.** At the other extreme are **replicated field studies,** wherein no manipulation or randomization of treatments occurs, but true replication occurs in a probability sampling framework, and information is gathered to evaluate alternate hypotheses. Conclusions from replicated field studies are broadly applicable, but are less certain than those from replicated field experiments.

Designing good field studies is more difficult than designing good experiments because of the potential for extraneous factors to invalidate tests or comparisons. One key step for both experiments and field studies is designing a sampling procedure to draw observations (experimental units or sample units) from the populations of interest. Only if

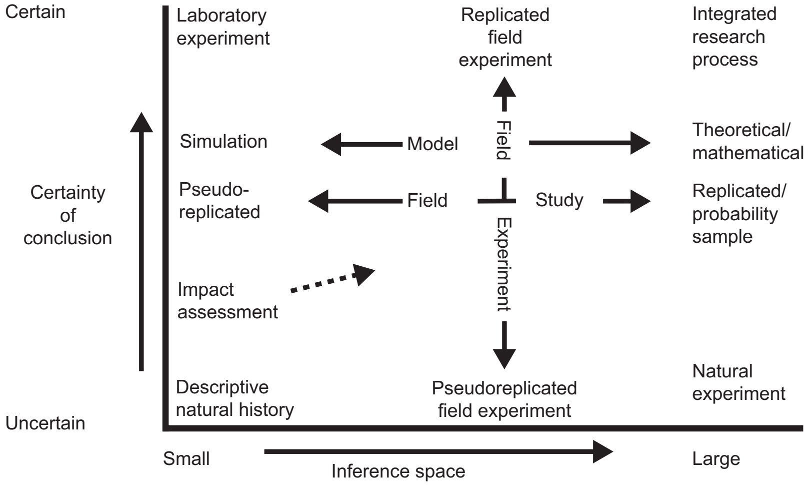


Fig. 1.5. Potential for wildlife study designs to produce conclusions with high certainty (few alternative hypotheses likely) and widespread applicability (a diversity of research populations in which the inferences apply).

this step is done properly can conclusions of the tests be applied to these populations. Survey sampling provides methods that are helpful in designing such sampling procedures (Cochran 1977). These methods are particularly important for field studies, but also are useful in field experiments for drawing experimental units and subsamples (samples within an experimental unit).

Impact assessments are another type of study design, but typically there is no replication, because the impact only occurs at a single site (e.g., an oil spill in a national wildlife refuge). However, they are useful for collecting baseline data as long as the type, time, and place of the impact are known; germane variables can be measured; and spatial and temporal controls exist (Green 1979, Williams et al. 2002*a*). Frequently, impact assessments are planned (e.g., prescribed fire), which allows for before-and-after measurements. The inference can be improved by monitoring both impact and nonimpact sites at several replicated sites of both types rather than only monitoring impact sites (Anderson 2002*b,* Williams et al. 2002*a*).

Models, which are a simplified representation of a system or process, are a versatile way to address a wide range of research questions that emphasizes being certain the conclusions do follow from the estimates, relationships, and assumptions (Fig. 1.5). The inference space of such models spans a continuum from large and general for theoretical or mathematical models (simple differential or difference equations) to smaller, more realistic for simulation models (complex multicausal, multiscale simultaneous differential and/or difference equations; Fig. 1.5). The certainty of conclusions based on models is in part influenced by the measurement of model variables and estimation of model parameters; it can thus be portrayed in predictions of models incorporating both process and estimation uncertainties (Hilborn and Mangel 1997:59). Models provide an important framework from which to begin to understand the processes influencing questions in wildlife science. They can help gauge the influence of one variable on others. For example, models can be used to assess how much juvenile dispersal influences population growth. By holding all other variables that influence population growth constant and then varying juvenile dispersal rates in the model, we can estimate how much the population growth is altered by small or large variation in juvenile dispersal rates. Conducting this type of **sensitivity analysis** makes such models an important tool for wildlife scientists.

In a subsequent section we describe an **integrated research process** that combines many aspects of study design, such as natural history observations, natural experiments, and laboratory experiments (Fig. 1.5). Using this process makes the research inference space large and increases the certainty of research conclusions by combining multiple study components. The complexities of the ecosystems in which wildlife science takes place are best addressed with the integrated research process, because it enables the wildlife scientist to capture more of the natural variability inherent in ecosystems (Clark and Stankey 2006, Morrison et al. 2008).

Once a research option has been chosen for each hypothesis or predictive variable, careful planning of the actual testing process can proceed. We must identify exactly what data will be collected for each hypothesis or predictive variable as well as when, how, how much, and for how long. Furthermore, how will these data be treated statistically? Will the data meet assumptions of the statistical test? Is the sample size adequate? Will the statistical hypothesis provide information directly related to the theory or model? Do biases exist in data collection, research design, or data analysis that might lead to a spurious conclusion? These questions must be considered carefully for each hypothesis before fieldwork begins. Consulting a statistician is important, and the statistician should understand the basic biological problem, the overall objectives, and the research hypotheses.

Peer review (evaluation by independent qualified reviewers) of the proposed research, including both study design and subsequent data collection and analysis, should be obtained from several people with expertise and experience in the research topic. Peer review will usually improve a research design and may disclose serious problems that can be solved during the planning stage. Unfortunately, most peer reviews occur too late for remedial work: after data collection, when the final report or publication manuscript is written.

Laboratory Experiments

Drawing inferences from **laboratory experiments** is easy because of the high level of control, yet this advantage must be weighed against the disadvantages (Table 1.1) in terms of (1) scale—laboratory experiments are restricted to small spatial scales and short time periods, (2) scope—only a restricted set of potential manipulations is possible in the laboratory, (3) realism—the laboratory environment places many unnatural stresses and constraints on animals, and (4) generality—some laboratory results cannot be extrapolated to natural communities. In a continuation of our example, laboratory experiments could be designed to examine whether geese really can select the most nutritious forage when given several alternatives in a cafeteria feeding trial. Diamond (1986) provided examples of the 3 types of experiments (laboratory, natural, and field) and made excellent suggestions for improving each type. Other examples and discussions of experiments are provided by Cook and Campbell (1979), Milliken and Johnson (1984), Kamil (1988), Hairston (1989), Underwood (1997), Tilman et al. (2006), and Chalfoun and Martin (2009).

Laboratory experiments in biology have been most useful for studying basic molecular or biochemical processes common to all organisms of a class. Laboratory experi-

Table 1.1. Strengths and weaknesses of different types of experiments

	Experiment type		
	Laboratory	Field	Natural
Control of independent variables[a]	Highest	Medium	Low
Ease of inference	High	Medium	Low
Potential scale (time and space)	Lowest	Medium	Highest
Scope (range of manipulations)	Lowest	Medium	High
Realism	Low	High	Highest
Generality	Low	Medium	High

Modified from Diamond (1986).

[a] Active regulation and/or site matching.

ments also have provided valuable information on emerging issues, such as wildlife diseases (e.g., Cooke and Berman 2000, Woodhams et al. 2008), efficacy of fertility control (Chambers et al. 1999, Hardy et al. 2006), and interactions between exotic and native species (e.g., Komak and Crossland 2000, Kopp and Jokela 2007).

Identifying one research design as best for all situations is not possible. All options should be considered as possibilities for evaluating hypotheses. Sometimes the best evaluation of a hypothesis involves using a combination of field studies and several types of experiments. For example, field observations by Ratti et al. (1984) indicated that spruce grouse (*Dendragapus canadensis*) fed exclusively on certain trees while ignoring numerous other similar trees of the same species. This observation led to a laboratory experiment with captive birds that tested the hypothesis that trees selected for feeding had higher nutritional content than did trees selected at random (Hohf et al. 1987).

Natural Experiments

Natural experiments are similar to field studies, except that in them we study the effects of uncontrolled treatments, such as wildfires, hurricanes, mass mortality from diseases, agricultural practices, and range expansions by animals or plants (Layzer 2008, Diamond and Robinson 2010). A key problem in evaluating natural experiments is that we cannot assign treatments randomly and therefore cannot be certain that any differences between treated and untreated units are not due to other factors that differed between them before some were "treated." In natural experiments the treatment precedes the hypothesis and most comparisons must be made after the fact. With our Canada goose example, a natural experiment might be to survey farmers in the region to locate pastures that have been fertilized and those that have not been fertilized in recent years. If our observations of feeding geese show more use of pastures that had been fertilized, we have more evidence indicating the birds select more nutritious forage. However, many alternative explanations remain. For example, perhaps those pastures that were fertilized were grazed later in the summer, and geese preferred fields with the shortest grass, where their ability to detect approaching predators is greatest. Many hypotheses of interest to wildlife biologists can be tested only with natural experiments, yet it is difficult to draw inferences from such experiments. The applied nature of wildlife management makes the realism and generality of natural experiments an important advantage, but their applicably to other populations is questionable unless multiple similar natural events are analyzed.

Field Experiments

Field experiments span a range from pseudoreplicated field experiments (Hurlbert 1984), in which no true replication is used (or possible) and conclusions are not certain, to replicated field experiments, for which conclusions are relatively certain (Johnson 2002). **Replicated field experiments** provide conclusions that are broadly applicable to free-ranging wildlife populations. Field experiments offer advantages over natural experiments in terms of certainty of inference and control of confounding factors, but they suffer the disadvantages of restricted scale and lower generality (Table 1.1). Compared to laboratory experiments, field experiments have greater scope and realism. Their main advantage is that we can randomly assign treatments and thereby eliminate fallacious conclusions due to effects of confounding factors. In field experiments, manipulations are conducted, but other factors are not subject to control (e.g., weather). In many situations in wildlife science, field experiments offer the best compromise between the limitations of laboratory and natural experiments (Wiens 1992, Krebs et al. 2001). In our Canada goose example, a subsequent field experiment would be to select random pairs of plots in known foraging areas. One member of each pair would be randomly assigned to be fertilized to learn whether geese select fertilized plots more often than they do the nonfertilized control plots. If they do select fertilized plots more often, a stronger inference about selection of nutritious foods could be made, because random assignment of a large number of plots to fertilization and control groups should have canceled effects of extraneous confounding factors. Interspersion of treatment and control plots (Hurlbert 1984, Johnson 2002) in fields naturally used by geese strengthens our belief that our conclusion would apply in systems where geese typically forage. **Adaptive management** could successfully incorporate field experiments by breaking management zones into replicates that are assigned various treatment levels for comparison to a standard management action (Connelly et al. 2003*a*). The strong advantages of field experiments are that random assignment of treatments to units interspersed among units to which the conclusions will apply protects against reaching invalid conclusions due to extrinsic factors.

Field Studies

Field studies may appear similar to experiments when they are conducted to test hypotheses, but they differ in that treatments are not assigned at random. For example, in a field study of dietary selection by Canada geese we might randomly select plots where flocks of geese have fed and those where they have not fed to examine whether geese choose areas with vegetation that is more nutritious. If they do, a weak inference would be that geese are choosing nutritious food, but numerous alternative explanations remain untested (e.g., maybe geese preferred hilltop sites, where visibility was good, and coincidentally these also were sites farmers fertilized most heavily to compensate for wind-driven soil erosion from previous years of tillage). Making inferences from field studies is difficult, because we make ex post facto comparisons among groups (Kerlinger 1986). Drawing firm conclusions is difficult, because these groups also differ in many other ways. The important characteristic of a field study is that we have comparison groups (e.g., use versus nonuse plots), but we have no treatments. Well-designed field studies can make important contributions to wildlife science and management (e.g., Paltridge and Southgate 2001), but their limitations must not be overlooked.

Impact Assessment

The most basic form of **impact assessment** compares measurements of wildlife and other characteristics at a site potentially affected by pollution or development to similar measurements at an unaffected reference site (Anderson 2002*b;* Fig. 1.5). This most simple form of impact assessment provides almost no basis for inference, because the reference site may differ for a multitude of reasons besides absence of the pollution source or development. Green (1979) noted the potential improvement in this design that results from making measurements before and after development at both reference and development sites. The basic before–after/control–impact (**BACI**) design has become standard in impact assessment studies (Anderson 2002*b,* Morrison et al. 2008) and also has been used in predator removal studies (e.g., Risbey et al. 2000). However, differences from before to after at reference (control) and impacted (treatment) sites are confounded by natural temporal variation and may not be produced by the impact itself (Hurlbert 1984, Underwood 1994, Williams et al. 2002*a*). In contrast to a well-designed field experiment, neither reference nor impacted sites are chosen randomly over space, and treatments are not assigned randomly. These limitations severely reduce the certainty of conclusions and the application of inferences to other areas. The goal is not to make inferences to all possible sites (Stewart-Oaten et al. 1986) for a power plant, for example, but to the particular power plant site being developed. For larger impact studies in which the goal is to make inferences with more certainty that are applicable to more sites (Fig. 1.5), the basic BACI design must be improved by the addition of replication and randomization (Skalski and Robson 1992, Underwood 1994). Stewart-Oaten et al. (1986) emphasized the value of expanding the BACI design to include temporal replication and noted the advantage of taking samples at irregular time intervals rather than on a fixed schedule. Hurlbert (1984) emphasized that comparing abundances of wildlife from repeated surveys at 1 impact and 1 reference site constitutes pseudoreplication that is only eliminated by having several replicated impact and reference sites. Replicated reference sites with environmental characteristics similar to the impact site are quite possible and highly desirable; however, replicated impact sites are only feasible in large-scale impact studies, typically involving meta-analysis of many single impact site studies.

Modeling

Modeling can be used as a deductive tool to synthesize theoretical understanding together with creative ideas about potential solutions to a problem or question. Creating a quantitative model makes the assumptions, accepted facts, generalizations, and laws or principles explicit for use in making valid and/or testable predictions. Kitching (1983:31) suggested this process of modeling involves 18 steps that correspond exactly to steps in the **scientific method** (Box 1.1; see details below under Modeling). Starfield and Bleloch (1991) describe this process in a straightforward manner with many wildlife examples created in spreadsheets. Clark (2007) presents a very rigorous account of ecological modeling utilizing free statistical and modeling software, such as R (R Development Core Team 2006), and Otto and Day (2007) provide a more mathematical, but very readable treatment of ecological modeling for biologists.

Modeling currently plays an essential role in 2 widely practiced processes of wildlife science: adaptive management and population viability analysis. **Adaptive management** requires building predictive models that summarize what is known or assumed about a management issue to examine alternative management actions. Managers choose one of the alternatives, and monitoring is conducted to: (1) ensure the action was accomplished; (2) evaluate whether the predicted consequences did in fact result; and (3) use feedback of results to improve understanding of the system, its behavior, key parameters, and relationships incorporated into the model. **Population viability analysis** uses models and data for populations to estimate the probability that populations of rare species will persist for specified times into the future (Mills 2007:254). These forecasts are essential to make scientifically defensible decisions concerning the listing or delisting of a species under the Endangered Species Act of 1973 (U.S. Fish and Wildlife Service 1973). Clearly building models such as these is an application of the scientific method that produces knowledge in the form of forecasts, but other applications of modeling strive to increase our general understanding of interrelationships (e.g., long-term impacts and dynamics of wolf, cougar, and coyotes on deer and elk; Garton et al. 1990, Varley and Boyce 2006), which are difficult to manipulate experimentally. Likewise building conceptual and quantitative models acts as a helpful early step in any investigation, because it sharpens our focus on identifying critical relationships and assumptions. It is an essential step in an integrated research process.

Integrated Research Process

The **integrated research process** (Fig. 1.5) builds on a solid base of natural history observations. Field observations and conceptual models should lead to experiments, and the results of natural experiments should lead to field and laboratory experiments. For example, Takekawa and Garton (1984) observed birds feeding heavily on western spruce budworms (*Choristoneura occidentalis*) during a budworm outbreak, which suggested that birds were a major source of budworm mortality. Field experiments were conducted to test this hypothesis by placing netting over trees to exclude birds. Survival of budworms on trees with netting was 3–4 times higher than on the control trees exposed to bird predation (Takekawa and Garton 1984). The level of certainty increases as many predictions from the research hypothesis are supported and alternate hypotheses are rejected in successively more rigorous tests that use replicated research options. After such findings are repeated over broad geographic areas or throughout the range of the species, the research hypothesis may become a principle of wildlife science (Johnson 2002). The integrated research process should be the goal of wildlife science (Clark and Stankey 2006, Morrison et al. 2008).

Outstanding examples of integrated research programs include long-term research on red grouse (*Lagopus lagopus scoticus*) in Scotland (Jenkins et al. 1963, Watson and Moss 1972, Moss et al. 1984, Watson et al. 1994, Kerlin et al. 2007, New et al. 2009), red deer (*Cervus elaphus*) on the Isle of Rhum, Scotland (Lowe 1969, Guinness et al. 1978, Clutton-Brock et al. 1985, Coulson et al. 1997, McLoughlin et al. 2008, Stopher et al. 2008, Owen-Smith 2010), and snowshoe hare (*Lepus americanus*) in North America (Keith 1963, 1974; Windberg and Keith 1976; Keith and Windberg 1978; Keith et al. 1984; Krebs et al. 2001). Research on red grouse and snowshoe hare has focused on hypothesized causes of population cycles, whereas research on red deer has focused on population regulation and density-dependent effects on survival, fecundity, reproductive success, spacing behavior, and emigration. Research on snowshoe hare has evaluated the role of predators (i.e., lynx [*Lynx lynx*] primarily, but other mammals and birds, too) as well as alternate proposed causes of the classic 10-year cycle in snowshoe hare and lynx numbers. For all 3 example species, descriptive studies and field observations formed the groundwork for subsequent research that included a series of innovative field studies and experiments (natural, field, and laboratory).

For example, preliminary studies of red grouse in Scotland (Jenkins et al. 1963) provided information on fundamental population parameters: births, deaths, immigration, and emigration. This information was used to form research hypotheses about causes of population fluctuations. Postulated causes initially included food quality, breeding success, spacing behavior, and genetics (Watson and Moss 1972, Kerlin et al. 2007). Using data from long-term field studies coupled with field and laboratory experiments, Watson and Moss (1972) concluded that quality of spring and summer foods (heather [*Calluna vulgaris*] shoots and flowers) affected egg quality, breeding success (viability of young), and spacing behavior of males and females, but territory size ultimately affected recruitment and population density (but see Bergerud [1988] for a critique of the self-regulation hypothesis and inferences based on red grouse research). Watson et al. (1984*b*) tested these conclusions with innovative field experiments, in which they (1) fertilized fields to assess grouse response to increased nutritional quality of the heather and (2) implanted males with time-release hormones to monitor changes in territory size associated with aggressiveness induced by higher or lower levels of androgens and estrogens (Watson 1967). Additional and more rigorous research rejected hypotheses that nutrition, genetics, and parasitism were causal factors (although Dodson and Hudson [1992] make a counterargument for the role of the parasite *Trichostrongylus tenuis*), and instead focused on emigration as the key factor in population declines (Moss et al. 1984, 1990; Watson et al. 1984*a;* New et al. 2009). These findings led to more research, because the mechanisms underlying density-dependent relationships, including summer and winter emigration, were unclear. Recent research has focused on the hypothesis of kin selection and differential aggression between kin and non-kin to explain cyclic changes in red grouse (Moss and Watson 1991, Watson et al. 1994) and synchronization of cycles across large regions according to weather (Watson et al. 2000, Kerlin et al. 2007). Thus, the integrated research process continues.

EXPERIMENTAL DESIGN

A variety of designs is available for researchers planning an experiment or quasi-experiment. This brief overview of some designs that have seen wide and innovative application to wildlife science should augment information provided in standard courses and references on experimental design (Underwood 1997, Scheiner and Gurevitch 2001, Quinn and Keough 2002, Morrison et al. 2008).

Single-Factor versus Multifactor Designs

Single-factor analyses are the simplest, because they involve only comparisons between 2 or more levels of 1 factor. Evaluating the simultaneous effect of 2 or more independent variables (**multifactor designs**) at once requires the use of complicated statistical methods, which should be discussed with a statistician. Under many conditions we can test 2 factors at once without expending more effort than would be required to test either of the factors alone. A complicating issue is the potential for interaction among factors (Steel and Torrie 1980). An **interaction** occurs if the effects of one factor on the response variable are not the same at different levels of another factor. For example, if we are interested in the effect of snowmelt date on nest success by arctic-nesting polymorphic snow geese (*Chen caerulescens*), we might discover an interaction between color phase and the onset of spring snow melt. Thus, darker, blue-phase birds would have higher nesting success during early snowmelt years, because they are more cryptically colored once snow has melted and experience less nest predation. During late snowmelt years white-phase birds are more cryptically colored and experience less nest predation. Many observations might be required to clarify possible relationships in these situations.

Dependent Experimental Units

Special designs have been developed to handle many types of dependency in experimental units, where **dependence** means that units tend to be more similar to one another than if we were to pick units at random from the entire population. For example, animals in one group tend to be more similar to one another (e.g., doe–fawn groups of deer have few bucks), and vegetation plots that are spatially proximate tend to be more similar to one another than are plots picked at random from the entire study area. A common design involves pairing. In a **paired design** we match experimental units in pairs that are as similar as possible. The treatment is then applied to one member of each pair at random. If there is a confounding factor, which we succeed in matching in the pairs, this approach will lead to a more powerful test than if pairing is not performed. For example, if we were studying the effects of spring burning on northern bobwhite (*Colinus virginianus*) habitat, we could establish pairs of plots throughout our study area, being careful to place each pair in a homogeneous stand of vegetation. We would then randomly assign one member of each pair to be burned in the spring. The analysis would then examine the differences between the members of a pair and test for a consistent improvement or decline in the burned member of the pair. Pairing would remove the effects of vegetation difference from one part of the study area to another and would result in a more sensitive experiment. If members of pairs are not more similar than members of the general population, the test will be less powerful because of the pairing.

When more than 2 levels of a factor are compared, pairing is referred to as **blocking.** A block is a set of similar experimental units. Treatments are randomly assigned to units in each block, and the effectiveness of blocking can be tested during the analysis. For example, if we expanded our study of burning to include spring and autumn burning as treatments, a block design would be appropriate. Three adjacent plots would be placed in homogeneous vegetation stands, and spring and autumn burning would be applied randomly to 2 of the 3 plots in each block (e.g., set of 3). This powerful design is normally referred to as a **randomized block.**

Another common form of dependency occurs when **repeated measurements** are taken on the same experimental unit through time. This practice is common in wildlife research, wherein the effects of treatments may change over time and must be monitored over a series of years. For example, in our study of spring and autumn burning the effects may be different in the first, second, and third growing seasons after treatment. The plots should be monitored over several years to measure these effects. The measurements are repeated on the same plots, so they are not independent. This repetition must be treated correctly in the analysis by using repeated measures or multivariate analysis of variance (ANOVA; Milliken and Johnson 1984, Johnson and Wichern 1988, Williams et al. 2002*a*). Dependency also is common in count data, especially when animals occur in groups (Eberhardt 1970). This lack of independence is often referred to as **overdispersion.** To properly cope with significant overdispersion the dependency should be modeled. Unless the biologist has extensive training in this topic, close cooperation with a consulting statistician is essential when designing and analyzing experiments involving such complicated designs.

Crossover Experiments

Crossover experiments provide a powerful tool to evaluate treatments that do not produce a long-lasting effect. Selecting pairs of experimental units and randomly assigning one member of each pair to be treated during the first treatment period initiates a crossover experiment. The second member serves as the control during this treatment period. In the second treatment period, the control unit becomes the treatment and the former treatment becomes the control. In this way the effects of any underlying characteristics of experimental units are prevented from influencing the results. This technique is valid only if treatment effects do not persist into the second treatment period.

Consider the following example. Suppose we wanted to test the hypothesis that mowing hay before 4 July decreases ring-necked pheasant (*Phasianus colchicus*) nest success. We could test this idea by dividing our study area into 5 homogeneous hayfield regions and then dividing each region into 2 portions. In one randomly selected portion of each region we could pay farmers not to mow their hay fields until after 4 July (treatments). In the other portion of each region, hay mowing would proceed as in most years, with the first cutting during mid-June; these portions would serve as controls. To monitor nest success, we locate nests by systematic field searches, being sure to search treatment and control areas with identical methodology (e.g., search intensity and seasonal timing). Nest success would be measured with standard techniques. After 1 year, we might measure significantly higher nesting success in the treatment portions (i.e., those areas with delayed hay mowing). However, the number of treatments is small, and we are not able to conclude with confidence whether higher nest success resulted from the treatment or from some undetected, inherent differences in treated portions of each region, such as nest predators. We would implement the crossover experiment by switching in the second year, so the original control portions of the study regions now have mowing delayed until after 4 July (new treatments), and the original treatment portions revert to the standard practice of first cutting in mid-June (new controls). If the portions with late cutting treatments again have higher nest success, we have better evidence that delayed mowing is responsible for higher nest success than we had at the end of the first year (i.e., we have better evidence for a cause-and-effect relationship). If even stronger support for the hypothesis is desired, the crossover experiment might be repeated in the same region and in other farming regions.

Fixed, Random, Mixed, and Nested Effects

One of the most critical decisions we must make in design concerns choosing the population for which we want to make inferences. If only a few levels of a treatment factor are relevant or would occur, we set a limited number of values at which the treatment would be applied, and the factor is termed a **fixed effect** (**Model I**)**.** If we want the conclusion to apply to any level of a treatment factor, we must select the treatment levels as a random sample from the population of potential values, so that a conclusion drawn about the effect of this factor applies across all levels at which it occurs. This design is termed a **random effect** (**Model II**)**.** A **mixed model** (**Model III**) includes both fixed and random effects. In simple 2-factor or multifactor designs all levels of each factor are applied to all levels of other factors, and the design is considered to be a **crossed design.** When this is not possible, the design must use approaches in which one factor is nested in another. A **nested design** can be described as hierarchical, which occurs most commonly where certain levels of one factor only occur in some levels of another factor. For example, a study evaluating the effect of vegetation treatment on bird communities might have 3 plant communities (ecological systems) with treatments of clearcut, burn, partial-cut, and controls. These factors would need to be nested if one of the plant communities was a shrub community where timber harvest does not occur. Decisions about the design of experiments must be reflected correctly in the analysis, as different measures of variance are appropriate for fixed, random, mixed, or nested effects.

Replication

Sample size refers to the number of independent random sample units drawn from the research population. In experiments, sample size is the number of replicates to which a treatment is assigned. For logistical reasons, we may measure numerous **subsamples** closely spaced in a single sample unit. However, we must be careful to distinguish these subsamples from independent random samples. Subsamples are not independent random sample units, because they typically are more similar to one another than are widely spaced samples. Similarly, subsamples in experiments are not true replicates if they cannot be independently assigned to a treatment category. The precision of a statistic is measured by its standard error. **Standard error** is calculated from the variation among the true sample units or replicates and the number of samples. If subsamples are mistakenly treated as true sample units or replicates, sample variance will underestimate the actual amount of variation in the populations; sample size will overestimate true sample size; and we will be overconfident in the precision of the estimate, because its true standard error will be underestimated.

To illustrate this point, suppose we wanted to evaluate the effect of prescribed fire on northern bobwhite habitat in a large valley (1,000 km^2). We might conduct research on a habitat improvement project that involves burning 1 km^2 of grassland and brush (e.g., Wilson and Crawford 1979). We could place 20 permanent plots in the area to be burned and 20 in an adjacent unburned area. Measurements on burned and unburned plots before and after the fire could be compared to examine the effects of fire on bobwhite habitat. However, the 20 plots on the burned area are not really replicates, but are merely subsamples or **pseudoreplicates** (Hurlbert 1984). In fact, we have only one observation, because we have only one fire in a 1-km^2 plot in the 1,000-km^2 valley. What would happen if we were to redesign the study to conduct 20 burns on 20 randomly chosen areas scattered throughout the valley? We would expect to see more variation among these plots than among 20 plots in a single burned area. The fallacy of the first design is obvious. A statistical test would evaluate only whether the burned 1-km^2 area differed from the unburned 1-km^2 area and could lead to false conclusions about effects of burning on bobwhite habitat in this area. A more appropriate design would require randomly selecting 40 sites from throughout the entire valley and randomly assigning 20 to be burned (treatments) and 20 to be control (unburned) sites. Each burned and control site would be sampled with 5 plots to measure bobwhite habitat before and after the treatment, and data would be analyzed by ANOVA; the 40 sites are samples and the 5 plots per site are subsamples. Thus, the 20 sites of each type would be true replicates. Treating the 100 burned and 100 unburned plots as experimental replicates would be an example of **pseudoreplication.** Psuedoreplication is a common problem, and investigators must understand the concept of replication and its importance in ecological research (Hurlbert 1984, Johnson 2002).

Controls

In experimental research, a **control** may be defined as parallel observations used to verify the effects of experimental treatments. Control units are the same as experimental units except they are not treated; they are used to eliminate the effects of confounding factors that could potentially influence conclusions or results. Creative use of controls would improve many wildlife studies. Experimental studies in wildlife that involve repeated measurements through time must include controls because of the importance of weather and other factors that vary with time (Morrison et al. 2008). Without adequate controls, distinguishing treatment effects from other sources of variation is difficult. For example, in the northern bobwhite study, control sites were required to distinguish the effects of burning from those of rainfall and other weather characteristics that affect plant productivity. There might be an increase in grass production in the year following burning because the rainfall was higher that year. Without control sites we cannot tell whether increased grass production resulted from increased rainfall, from burning, or from a combination of both factors. Thus, we cannot evaluate the relative importance of each factor.

Determining Sample Size

One of the more challenging steps prior to starting actual data collection is to set goals for sample size using a prospective **power analysis.** The **power** of any hypothesis test is defined as the probability of rejecting the null hypothesis when, in fact, it is false. Power depends on the magnitude of the effect (e.g., magnitude of difference between treatment and control or a bound on the estimate), variation in the characteristic, significance level (α), and sample size. Zar (1999) provides formulas to calculate power and sample size for hypothesis tests, but a statistician should be consulted for complicated experimental designs and analyses. Many statistical packages (e.g., Statistical Analysis System; SAS Institute 2008) or specialized analysis software (e.g., MARK; White and Burnham 1999) provide capability to generate sample data for analysis to determine in advance how large the sample size should be to detect effects expected.

Effect size (magnitude of effect) is an important factor influencing sample size requirements and the power of a test. However, power and sample size calculations should be based on a biologically meaningful effect size. Identifying a biologically significant effect usually involves expressing the conceptual model as a quantitative model plus value judgments about the importance of a biological response. Estimating power of the test and calculating sample size requirements forces the investigator to evaluate the potential significance of the research prior to beginning fieldwork. Sample size analysis may lead to substantial revision of the goals and objectives of research.

Checklist for Experimental Design

The design of any experiment must be developed carefully or the conclusions reached will be subject to doubt. Four particularly critical elements in the design of a manipulative experiment are (1) specification of the research population, (2) replication with independent units, (3) proper use of controls, and (4) random assignment of treatments to experimental units. An experimental design checklist, such as the one listed in this section, is useful for providing a series of questions to assist in addressing these critical elements. Many of the questions will be helpful with the design of data gathering for studies involving nonexperimental hypothesis testing. Some experimental designs may address several hypotheses simultaneously (e.g., **factorial designs**); in other designs, each hypothesis may require independent experimental testing.

1. **What is the hypothesis to be tested?** The hypothesis developed from the conceptual model must be stated clearly before any experiment can be designed. For example, we could test the hypothesis that nest predation on forest songbirds is higher at sharp edges, such as occur at typical forest clearcuts, than at feathered edges (partial timber removal), such as occur at the boundary of selectively logged areas (Ratti and Reese 1988, Chalfoun et al. 2002, Stephens et al. 2003).

2. **What is the response or dependent variable(s) and how should it be measured?** The **response variable** should be clear from the hypothesis (e.g., nest predation), but selecting the best technique to measure it might be more difficult to determine. We must consider all possible methods and identify one that will simultaneously maximize precision and minimize cost and bias. It is often helpful to contact others who have used the techniques, examine the assumptions of the techniques, and conduct a pilot study to test the potential them. In our example, we might search for naturally occurring nests along forest edges and use a generalized Mayfield estimator of mortality rate (Heisey and Fuller 1985, Jehle et al. 2004, King et al. 2009). This response variable is continuous, and we could apply any of a variety of designs termed **general linear models** (GLM; e.g., ANOVA, linear regression, or analysis of covariance) under a hypothesis testing framework though application of information theoretic methods to these models. Alternately, we could measure the response for each nest as successful (at least one young fledged) or unsuccessful and use appropriate analysis methods, such as chi-squared statistics applied to contingency tables or log-linear models (Fienberg 1970, 1980; Hazler 2004).

3. **What is the independent or treatment variable(s) and what levels of the variable(s) will be tested?** The **indepen-**

dent variable(s) should be clear from the hypothesis (sharp and feathered forest edges in our example), but selecting levels to test will depend on the population for which we want to make inferences. If we want to test the effects of the independent variable at any level, we must select the levels to test at random (random effects [Model II]; Zar 1999). If we are interested in only a few of the levels that our independent variable could take, we use only those levels in our experiment and make inferences only to the levels tested (fixed effects or Model I; Zar 1999). For example, if we wanted to evaluate the effects of forest edges of any type on predation rates, we would select types of forest edges at random from all types that occur and apply a random effects model to analyze the data. In our example we are interested only in the 2 types categorized as sharp and feathered, so a fixed effects model is appropriate. Additionally, our independent variable must be identified and classified clearly or measured precisely. Finally, how can we use controls to expand our understanding? In our example, comparing nest predation in undisturbed forests to predation at the 2 types of edges might be enlightening, and we would analyze the data with fixed effects models. Our final conclusions would not apply to predation rates in all types of forest edges, but only to the 2 types that we compared to undisturbed forest.

An alternative approach to the design would be to treat the independent (treatment) variable as being **continuous** and use regression rather than a classified grouping of treatment categories. Under this design we might specify the treatment would consist of some level of overstory removal on one side of the forest edge, and we would apply regression forms of GLM under either hypothesis testing or information theoretic model evaluations. The response could be measured as the difference in predation rates between the 2 sides of the boundary, which would be predicted from percentage of overstory removed. Here it becomes critical to select **treatment levels** (e.g., percentage of overstory removed) across the full range of forest treatments to which we want to apply our conclusions.

4. **For which population do we want to make inferences?** If the results of the experiment are to be applied to the real world, our experimental units must be drawn from some definable portion of that world, the **research population.** The dependent and independent variables chosen should define the relationship(s) examined and place constraints on the definition of this population. We must also consider the impact of potential **extraneous factors** when selecting the population of interest. If the population is defined so broadly that many extraneous factors affect the results, the variation might be so large that we cannot test the hypothesis (low internal validity). If the population is defined so narrowly that we have essentially a laboratory experiment, application of the results might be severely limited (low generality or external validity).

Reaching the proper balance between internal and external **validity** takes thought and insight. For example, we might want to compare nest predation rates in sharp and feathered forest edges throughout the northern Rocky Mountains, but the logistics and cost would make the study difficult. Thus, we might restrict the study population to one national forest in this region. Next we need to consider the types of forests. We might want to test the hypothesis for the major forest types, but we know the species of birds nesting in these forests and their nest predators differ among forest types. Thus, we may need to restrict our population to one important type of forest to remove extraneous factors that could impact the results if we sampled a large variety of forest types. We need to ask what types of sharp and feathered edges occur to decide which we will sample. Sharp edges are commonly produced by clearcuts, power line rights-of-way, and road rights-of-way. These 3 types differ dramatically in such factors as size, shape, human access, and disturbance after treatment. Additionally, our ability to design a true experiment involving random assignment of treatments is severely limited for all but the clearcuts. Therefore, we might restrict the populations to sharp edges created by clearcuts and feathered edges created by selective harvests.

5. **What is the experimental unit?** What is the smallest unit that is independent of other units, which will allow random assignment of a treatment? This element must be identified correctly or the resulting experiment might not have true replication, but instead represent a case of pseudoreplication (Hurlbert 1984). For example, we might erroneously decide the experimental unit for our nest predation study will be an individual nest. The resulting design might entail selecting 3 areas and randomly assigning them to be clearcut, control, and selectively logged. By intensive searching, we find 20 nests along the edge of each area and monitor them for predation. The resulting data would suggest 20 replicates of each treatment, but, in fact, only a single area was given each treatment. Only 1 area was randomly assigned each treatment, and the 20 nests are subsamples. Thus, pseudoreplication restricts the potential inferences. In effect, we have sampled from populations consisting only of 2 logged areas and 1 unlogged area, and our inferences can be made only for those 3 areas, not to clearcuts, selective cuts, or undisturbed forests in general.

In some situations, **pseudoreplicated designs** are unavoidable, but interpretation of their results is severely restricted, because without replication, confounding factors rather than the treatment could have caused the results. For example, in our nest predation experiment if one of the areas was in the home range of a pair of common ravens

(*Corvus corax*) and the other areas were not, this single confounding factor could affect the results regardless of treatment. A more reliable experiment would require that we identify several areas with potential to be logged, perhaps 15, sufficiently far apart to be independent of one another, and that we randomly assign 5 each to be clearcut, selectively harvested, and controls. We would locate and monitor several nests in each area. The nests in a single area would be correctly treated as subsamples, and their overall success treated as the observation for that area. This approach attempts to remove the effects of confounding factors and to allow development of a conclusion with general application to the populations sampled (i.e., edges created by clearcuts and selective cuts in this habitat type in this region). Including control stands without an edge provides invaluable information for assessing the biological significance of the difference between the 2 types of edges.

6. **Which experimental design is best?** A few of the most widely used designs are described, but we advise consulting texts on experimental design and a statistician before making the final selection (Scheiner and Gurevitch 2001, Quinn and Keough 2002, Morrison et al. 2008). The choice depends primarily on the type of independent and dependent variables (categorical, discrete, or continuous), number of levels of each, ability to block experimental units together, and type of relationship hypothesized (additive or with interactions). For our study of nest predation along 2 types of forest edges, a single-factor design would be appropriate, but Hurlbert's (1984) argument for interspersion of treatments and controls could be incorporated by using a more sophisticated design. For example, 3 adjacent stands in 5 different areas might be randomly assigned to treatment and controls, with areas cast as blocks, resulting in a randomized complete-blocks design (Zar 1999).

7. **How large should the sample size be?** Estimating sample size needed for proper analysis is essential. If the necessary sample size were too costly or difficult to obtain, it would be better to redesign the project or work on a different question that can be answered. Sample size depends on the magnitude of the effect to be detected, variation in the populations, type of relationship that is hypothesized, and desired power for the test. Typically some preliminary data from a pilot test or from the literature are required to estimate variances. These estimates are used in the appropriate formulas available in statistical texts (e.g., Zar 1999) and incorporate a prospective power analysis to ensure that we have a high (80–90%) chance of detecting biologically meaningful differences between the treatment and control categories. Powerful analysis programs like SAS (SAS Institute 2008) provide tools to perform prospective power analysis for complicated designs.

8. **Have you consulted a statistician and received peer review on the design?** Obtaining review by a statistician before the data are gathered is essential. The statistician will not be able to help salvage an inadequate design after a study is completed. Peer review by other biologists having experience with similar studies also could prevent wasted effort if measurements or treatments are proposed that will not work on a large scale in the field. Now is the time to get these comments!

MODELING

"All models are wrong, but some are useful" (Box 1979:2). Rigorously evaluating ideas concerning wildlife habitats and populations by using experimental manipulations may be difficult, because we cannot randomly assign treatments and the high cost of treatments precludes adequate replication in many cases. However, modeling methods provide an alternative route to finding solutions to pressing problems (Starfield and Bleloch 1991, Shenk and Franklin 2001), selecting the best of alternative choices (Holling 1978, Walters 1986, Clemen and Reilly 2001, Conroy and Peterson 2009), determining the relative magnitude of effects from multiple causes acting simultaneously (Wisdom and Mills 1997, Saltelli et al. 2001), and evaluating population viability (Mills 2007:254). A biologist's goal should be to build the simplest model that describes the relationships between causative factors and the effects they produce. It is most likely that a wildlife scientist will select a modeling strategy at the simple, empirical ends of the continua in terms of model complexity (Table 1.2) or in Levins's (1966) terms, sacrifice generality for realism and precision. Long-term monitoring data and extensive measurements of demographic rates and habitat relationships provide the basis for more complex models.

In most cases the goal is to model the responses of wildlife populations or habitats with the smallest number of predictors necessary to make good predictions. Note, this use of the term model corresponds to what Williams et al. (2002*a*:23) refer to as a **scientific model** rather than a statistical model. **Statistical models** are the foundation for all statistical estimation, hypothesis testing, and statistical comparison among competing models through an inductive process based on limited observations (see later sections under Parameter Estimation and Confronting Theories with Data). Scientific models, described in this section, are used deductively to project system dynamics based on a set of ideas expressed as characteristics and relationships estimated inductively from statistical models. We use these 2 types of models cooperatively to help answer important questions about wildlife. Scientific models are commonly referred to as **simulation models,** because they simulate the dynamics of a system described in terms of the assumptions, charac-

Table 1.2. Modeling strategies along gradients of simple to complex for scientific and statistical models

	Gradient	
	Simple	Complex
Scientific models		
Quantification	Conceptual (verbal) →	Quantitative
Theoretical	General →	Complex simulation
Relationships	Linear →	Nonlinear
Variability	Deterministic →	Stochastic
Time scale	Time-specific →	Dynamic
Mathematical formulation	Difference equations →	Differential equations
Number of factors	Single →	Multifactor
Number of sites	Single site →	Multisite
Number of species	Single species →	Multispecies
Statistical models		
Sampling	Simple random →	Stratified, clustered or multistage
Hypothesis testing	Fixed or random effects →	Mixed fixed and random effects
Independence of observations	Complete independence →	Dependence among observations in space, time, or both
Errors	Single term →	Separate process and observation errors

teristics, relationships, and variability observed. When variability is a key component, they are referred to as **Monte Carlo** scientific models. Kitching (1991) suggested a variation of the following 8 steps to build an ecological model. These steps are directly applicable to building scientific wildlife models.

Steps to Build a Model

Problem Definition

The problem of interest must have been identified earlier as one of the first steps in the scientific method, and the relevant theory, previous observations, conceptual model (Fig. 1.4), predictions, and hypotheses must be stated clearly. Someone proposing to build a model to answer the question must now explain why a numerical or mathematical model is an appropriate way of tackling the problem (Kitching 1991:31). A good example of an appropriate question is: which of the available management options are more likely to recover an endangered species and prevent its extinction? It is important to embrace the modeling approach to this problem as a pragmatic one. "There is no point at all in building an ecological model that is more complex, more complete or more time-consuming than is justified by the terms of reference of the problem to which the model is a response" (Kitching 1991:31). The better the problem(s) is identified, the more useful the model will be.

System Identification

After identifying the problem(s) it is critical to define the **system boundary** and the **level of resolution** to model in the hierarchy of ecological levels (ranging from individual animals with associated spatial extent to population or metapopulation; Fig. 1.2). The biologist must then select a set of components to model (see the examples in Fig. 1.4). One strategy is to pursue a parsimonious approach, making the model as simple as possible, by selecting only critical components essential to describe the system. This approach is used for developing general **theoretical models** (Table 1.2) taking the form of analytical mathematical models. The other extreme is to include all components likely to be involved in the processes of interest. Such models take the form of complex simulation models. The typical route followed in wildlife models is to take the simple empirical approach, and Starfield and Bleloch (1991) recommend tending toward the parsimonious end while including enough complexity to produce realistic predictions. Once the initial set of components is defined to meet the objectives, the nature of their interactions must be defined based on creative thinking and literature as follows: positive, negative, feedback loops, and complex combinations. Creating a simple system diagram is useful for clarifying these relations (e.g., Fig. 1.4) and guiding literature searches.

Model Type Selection

The great variety of model types available (Table 1.2) may seem daunting at first, but the problem definition process described above should guide selection of the appropriate type of model along the continuum from simple to complex, with preference always for the simplest model necessary to meet the needs. Building complex models requires estimating more characteristics with more complex relationships. Fortunately most wildlife problems can be handled with simple, linear models incorporating deterministic effects of a few independent factors at a single or small number of sites. Even forecasts for population viability requiring stochastic models with time lags are easily modeled with simulations based on estimates obtainable with standard linear regression methods (e.g., Garton et al. 2010).

Mathematical Formulation

Almost all wildlife models are formulated as difference equations because of strong seasonal and annual patterns, which make estimating parameters for **continuous time models** formulated as **differential equations** difficult. Differential equation formulations have been more successful for developing general theoretical models that form the basis for many ecological theories underlying principles of wildlife population ecology (Ginzburg 1986, Turchin 2001, Berryman 2003, Colyvan and Ginzburg 2003), but translating these general models into **stochastic difference equation models** has proven very successful for modeling time series of populations with complex dynamics (e.g., population viability analysis for San Joaquin kit fox [*Vulpes macrotis mutica*] incorporating density dependence and a 2-year lag in rainfall effects on plant productivity; Dennis and Otten 2000).

Computational Method and Program Selection

Simple wildlife models formulated in commonly used general purpose spreadsheet programs can provide remarkable insight into wildlife population dynamics (Starfield and Bleloch 1991). Some specialized software designed for specific purposes, such as population viability analysis, have wide application to projecting persistence of endangered and rare species—for example, RAMAS (Akçakaya 2000b) and VORTEX (Lacy 1993). Programs designed for statistical analysis—for example, SAS (SAS Institute 2008) and R (R Development Core Team 2006)—are equally adaptable to simulating both deterministic and stochastic models as they are to estimating the parameters for these models (Bolker 2008, Garton et al. 2010).

Parameter Estimation

Sampling methods, least squares for GLM, and maximum likelihood methods are all useful for estimating parameters for alternative models. Information-theoretic approaches to evaluating competing models (see further details later under the section Confronting Theories with Data) provide excellent tools to evaluate relative precision of alternative models in predicting responses (Burnham and Anderson 2002). Burnham and Anderson (2002) contend that information theoretic methods, such as using **AIC** to assess the information content of a model, should be applied where we cannot experimentally manipulate causes or predictors. Model averaged parameter estimates are readily calculated within this framework using Akaike weights (Burnham and Anderson 2002:133ff.).

Model Validation

Validation of a model should take at least 2 forms. Comparing the predictions of the model to data that were analyzed to build the model provides a preliminary validation or **verification** (Oreskes et al. 1994) that is always performed as part of constructing the model. Clearly this step is essential to verify the model is performing as the investigator expects. A real test of the **validity** of the model requires comparing output from the model to independent data not used in its construction (Gardner and Urban 2003). The comparison is usually made with standard statistical tools, such as correlation and regression, which may be evaluated from a frequentist perspective by using either hypothesis tests or likelihood measures. Because models using all data possible maximize precision in parameter estimation, approaches, such as jack-knifing, in which each individual observation is predicted from models fit to all the rest of the data are applied (Efron and Tibshirani 1993).

Model Experimentation

Once the previous seven steps are completed, the model is ready to be used to address the original questions that initiated the modeling process. **Experiments** are performed by manipulating key input parameters to assess the response of model output characteristics to anticipated alternative management actions and/or potential environmental trends, changes, or variation. A useful model is an invaluable aid to both research and management, but the **veracity** of any predictions rests firmly on assumptions built into the model structure, the relationships modeled, and the validity of any parameters estimated from field observations. Scheller et al. (2010) provide further details on the approach outlined here, which applies modern software engineering techniques as part of a process to increase the reliability of ecological models. A useful model should be used interactively with ongoing research and management activities, so that modeling exercises help identify critical relationships and parameters that are then investigated in the field by gathering new observations or performing experiments. In the management context this process is **adaptive management:** model predictions guide management actions and continued monitoring provides feedback to validate and improve model assumptions expressed as model parameters and relationships.

SAMPLING

Most information gathered by wildlife biologists is used to meet descriptive rather than experimental objectives, but obtaining precise estimates is equally important for both experiments and descriptive research. Examples include estimates of population size, recruitment, herd composition, annual production of forage species, hunter harvest, and public attitudes. In these efforts biologists attempt to obtain estimates of characteristics that are important for management decisions. We want to obtain the best estimates possible within the constraints of our time and money resources. A large body of statistical literature exists to help; these types of studies are referred to as **surveys,** and the topic is known as **survey sampling** (Cochran 1963, 1983; Scheaffer et al. 2005) or **finite population sampling.**

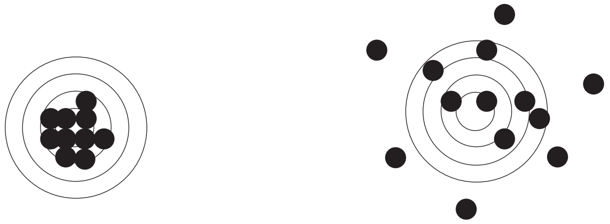
The research population is typically synonymous with the statistical population, but a powerful approach is to redefine the **statistical population** geographically in terms of units of space or habitat. Defining the statistical population as drainages, forest stands, individual ponds, or square-kilometer blocks often facilitates estimating total numbers of animals and the composition of a population. Sampling smaller units of habitat is more likely to be logistically feasible. Likewise this redefinition of the research (statistical) population makes it feasible to apply the powerful tools for sampling from finite populations.

Sampling also is a critical part of experimental research and the test of formal statistical hypotheses. All field studies and most field experiments require creative sampling designs to reduce variation among observations in the treatment or comparison categories. For example, stratification and clustering can sharpen comparisons, but data collected using these methods require analysis by more complicated designs (e.g., block or split-plot designs; Zar 1999). Choice of specific sampling methods is dependent on the objectives or hypotheses being addressed, the nature of the population being sampled, and many other factors (e.g., species, weather conditions, topography, equipment, personnel, time constraints, and desired sample sizes). A variety of sampling designs is available for biologists to use in wildlife surveys and experimental research (Thompson et al. 1998, Scheaffer et al. 2005, Morrison et al. 2008).

Precision, Bias, and Accuracy

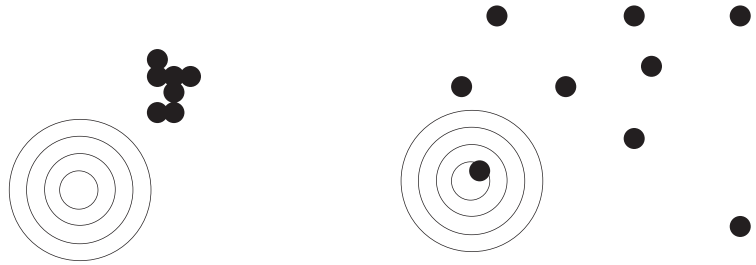
One measure of quality of estimates is their precision. **Precision** refers to the proximity of repeated measurements of the same quantity (Cochran 1963, Krebs 1999, Zar 1999). Precision of an estimate depends on variation in the population and the size of the sample. Indicators of the precision of an estimator are **standard error** and **confidence intervals.** Larger variation in the population leads to lower precision, whereas a larger sample size produces higher precision in the estimator. Another measure of the quality of an estimator is **bias.** Bias describes how far the average value of the estimator is from the true population value. An unbiased estimator centers on the true value for the population. If an estimate is both unbiased and precise, we say that it is **accurate** (defined as an estimator with small mean-squared error; Cochran 1963). Accuracy is the ultimate measure of the quality of an estimate (Fig. 1.6) and refers to the small size of deviations of the estimator from the true population value (Cochran 1963).

Let us illustrate these concepts with a typical population survey. Suppose we were interested in estimating the density of elk on a large winter range. One approach might be to divide the area into a large number of count units of equal size and draw a sample of units to survey from a helicopter. This approach would define the research population in terms of a geographic area rather than in terms of animals. The elements of the target population are count units, and we select a sample of these units using an objective sampling design (a probability sample). Using the helicopter, we search each sampled unit, attempting to count all elk



A Unbiased and precise = accurate

B Unbiased but not precise = not accurate



C Biased and precise = not accurate

D Biased and not precise = not accurate

Fig. 1.6. Concepts of bias, precision, and accuracy illustrated with targets and a shot pattern. *Modified from Overton and Davis (1969), White et al. (1982).*

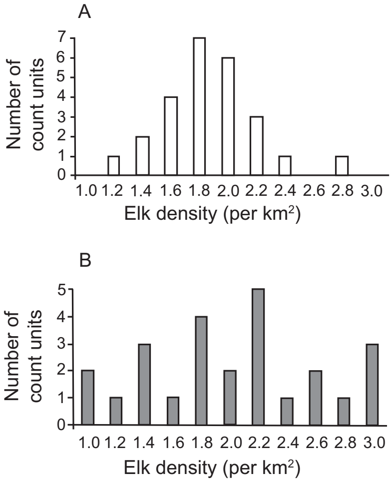


Fig. 1.7. Hypothetical example of elk counts and density estimates in (A) Area 1 and (B) Area 2.

present in it. We divide the number of elk counted in a unit by the size of that unit to obtain a density estimate for each unit (Fig. 1.7A). The histogram suggests little variation in density on this winter range, as most spatial units (80%) have densities between 1.5 and 2.3 elk/km^2. We need a single value that is representative of the entire winter range, and we choose the mean from the sample as the best estimate. The variation from one unit to the next is small; thus, the mean from our sample is a fairly precise estimate. But, suppose we had obtained different results (Fig. 1.7B). Now the variation from one unit to the next is great, and the sample mean is less precise and not as reliable as the previous estimate. Thus, for a given sample size, the former estimate is more precise because of less variation in the population.

Would the mean from the sample in Area A (Fig. 1.7A) be an accurate estimate of the mean density of elk on this winter range? To answer this question, we must evaluate the bias in the estimate. If the winter range was partially forested or had tall brush capable of hiding elk from view, aerial counts in each unit would underestimate the true number of elk present (Samuel et al. 1987). In this example the mean density from the sample would be a biased estimate of elk density on the winter range and, therefore, not highly accurate. If the winter range were a mixture of open brush fields and grasslands, where all animals would be visible, mean density from the sample could be an accurate estimate of elk density on the entire winter range. We strive for accuracy in our estimates by selecting the approach with the least bias and most precision, applying a valid sampling or experimental design, and obtaining a sufficiently large sample size to provide precise estimates.

Evaluating bias in an estimate is difficult and, in the past, has been based on the researcher's biological knowledge and intuition. If the bias is constant, the estimate can be used to make relative comparisons and detect changes (Caughley 1977). Usually it is not constant (Anderson 2001), but its magnitude often can be measured so that a procedure to correct estimates can be developed (Rosenstock et al. 2002, Thompson 2002*b*). For example, Samuel et al. (1987) measured visibility bias in aerial surveys of elk from helicopters, and Steinhorst and Samuel (1989) developed a procedure to correct aerial surveys for this bias.

Sampling Designs

Simple Random

A **simple random sample** requires that every sample unit in the population has an equal chance of being drawn in the sample and the procedure for selecting units is truly random. This can be accomplished by assigning each member of the population a number and then picking numbers, to identify members to sample, from a table of random numbers or a random number generator on a computer or calculator. For example, suppose that for a special hunt in which a limited number of permits was issued, we wanted to estimate the number of successful hunters. We might decide to contact a sample of permit buyers by telephone after the season to measure their hunting success. A survey design checklist (Box 1.4) helps design such a survey properly. The **population** that we want to make statements about is all persons who obtained a permit. The list of the members of the population is usually called the **sampling frame** (Scheaffer et al. 2005). It is used to draw a random sample from the population. The sampling frame must be developed carefully, or the resulting estimates may be biased. For example, if a portion of the permit buyers did not have telephones and we decided to drop them from the list, the results could be biased if such hunters had different hunting success than did permit buyers with telephones. To draw a random sample for the survey, we could assign a number to each person who purchased a permit and select the numbers to be contacted by using a table of random numbers or a random number generator.

In other types of surveys, obtaining a truly random sample of the population might be difficult. In such instances another method, such as systematic sampling, should be used. When the research population consists of animals that would be difficult to sample randomly, one approach is to change the design. We do this by making small geographic units, such as plots or stands, the **sample units** (or **experimental units,** if we are developing a sampling design for an experimental treatment) and making the measurement on each plot a number or density of animals. Thus, we can take a random sample of spatial units and use it to infer abundance across the entire study area sampled. A valid random sampling procedure must be independent of investigator decisions. For example, an excellent procedure for ran-

Box 1.4. Survey design checklist

Question	Example
1. What is the survey objective?	Estimate the percentage of successful hunters
2. What is the best technique	Telephone survey of permit holders or method?
3. To which population do we make inferences?	Everyone who has a permit for this hunting period
4. What will be the sample unit?	Individual permit holders
5. What is the size of the population to be sampled (N)?	$N = 350$ (for special permit hunt)
6. Which sample design is best?	Simple random sample (Scheaffer et al. 2005).
7. How large should the sample be?	$n = \frac{Np(1-p)}{(N-1)B^2/4 + p(1-p)}$ $n = \frac{Np(1-p)}{(N-1)(B^2/4) + p(1-p)}$, where: N = population size (350) p = proportion of permit holders who harvested deer (from pilot survey = 0.24) B = bound on the estimate = 0.05 (we want an estimate with $p \pm 0.05$ confidence) Therefore $n = \frac{350(0.24)(1-0.24)}{(350-1)(0.05)^2/4) + 0.24(1-0.24)}$, $n = 159$ (i.e., we should contact approximately 160 permit holders
8. Have you contacted a statistician to review design?	Yes!

domly locating plots in a study area would be to use a Landsat image of the study area stored in a Geographic Information System (**GIS**) program, which allows us to select random locations within the boundary of our study area using Universal Transverse Mercator (UTM) coordinates (Fig. 1.8A). The UTM coordinates of these selected plot locations can be entered into a hand-held Global Positioning System (**GPS**) unit that will guide us to the exact location. Random-like methods, referred to as haphazard or representative, have been used in place of truly random designs, but should be avoided, because they are subject to investigator bias. An example of these methods is the technique of facing in a random direction and throwing a pin over the shoulder to determine the center for a vegetation plot. Although this procedure seems random, the odds of a field crew randomly facing away from a dense stand of thorny shrubs, such as multiflora rose (*Rosa multiflora*), and throwing the pin into the middle of such a patch is practically zero. Truly random samples occasionally produce poor estimates by chance due to poor spatial coverage of the area or population of interest (e.g., in an area with a small number of important habitat patches, all patches may be missed by a truly random approach; Hurlbert 1984, Johnson 2002).

Systematic

A **systematic sample** is taken by selecting elements (sampling units) at regular intervals as they are encountered. This method is easier to perform and less subject to investigator errors than simple random sampling. For example, if we wanted to sample birdwatchers leaving a wildlife management area, it would be difficult to draw a truly random sample. However, it would be easy to draw a systematic sample of 10% of the population by sampling every tenth person leaving the area. Systematic sampling also is used extensively in vegetation measurements because of its ease of use in the field. It is almost exclusively used in geographic sampling, because it makes possible evaluation of the spatial pattern of variability (e.g., spatial autocorrelation), which is used for most modern spatial modeling. A valid application requires random placement of the first plot, followed by systematic placement of subsequent plots, usually along a transect or in a grid pattern (Fig. 1.8B). This approach often

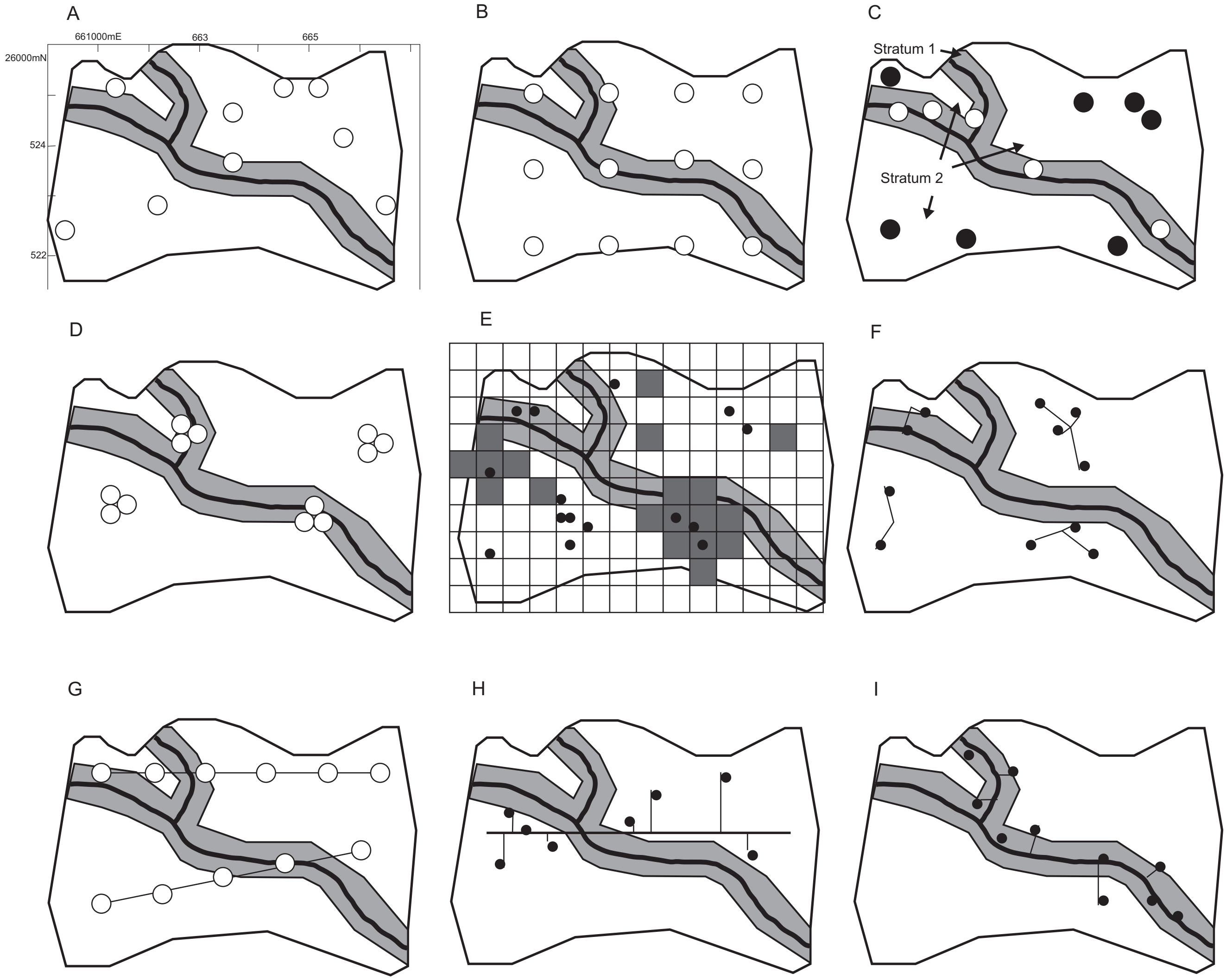


Fig. 1.8. Examples of sampling designs: (A) simple random sample, (B) systematic sample, (C) stratified random sample, (D) cluster sample, (E) adaptive cluster sample. Examples of sampling methods: (F) point sampling, (G) plots along transects, (H) line transect, (I) road sampling.

provides greater information per unit cost than simple random sampling, because the sample is distributed uniformly over the entire population or study area. For random populations (i.e., no serial correlation, cyclic pattern, or long-period trend), systematic samples give estimates with the same variance as simple random samples.

The major danger with systematic samples is they may give biased estimates with **periodic populations** (i.e., with regular or repeating cycles). For example, if we were interested in estimating the number of people using a wildlife management area, we might establish a check station and take a systematic sample of days during the season. This procedure could yield extremely biased results if we chose to take a sample of one-seventh of the days. If the day sampled fell during the workweek, we could obtain different results than if it were during the weekend. Additionally, the estimate of variance would likely be too small, leading us to conclude the estimate was much more precise than it is in reality. In this situation the population sampled obviously is periodic; in other situations the periodicity might be quite subtle. Thus, systematic sampling must be used with caution. The formal procedure is conducted by randomly selecting one of the first *k* elements to sample and every *k*th element thereafter. For example, if we wanted to sample 10% of our population, *k* would equal 10, and we would draw a random number between 1 and 10. Suppose we selected 3; we would then sample the 3rd element and every 10th element thereafter (i.e., 13th, 23rd, 33rd, . . . element). At a check station we might use this strategy to sample 10% of the deer hunters or birdwatchers who came through the station. When locating plots along a transect, we would randomly locate the starting point of the transect and then place plot centers at fixed intervals along the transect, such as every 100 m. Advantages and disadvantages of random and systematic sampling have been reviewed by Thompson et al. (1998), Krebs (1999), and Morrison et al. (2008).

Stratified Random

In many situations, obvious subpopulations exist in the total population. For example, tourists, birdwatchers, and hunters are readily divided into residents and nonresidents. A study area can be divided into habitats. A population of animals can be divided into age or gender groups. If members of these subpopulations are similar in terms of the characteristics we are estimating and the subpopulations themselves differ from one another in the characteristic of interest, a powerful design to use is **stratified random sampling.** Subpopulations are referred to as strata, and we draw a simple random sample of members from each stratum. Stratified random sampling also is useful if we are particularly interested in the estimates for the subpopulations themselves. The strata are chosen so they contain units of identifiably different sample characteristics, usually with lower variance within each stratum.

For example, if the objective of a study of moose (*Alces alces*) is to estimate moose density, we might define strata on the basis of habitats (e.g., bogs and riparian willow [*Salix* spp.] patches, unburned forests, and burned forest). We then draw a simple random sample from each stratum (Fig. 1.8C). If moose density is different among strata, variation in each stratum will be less than the overall variation. Thus, we will obtain a better estimate of moose density for the same or less cost. If strata are not different, stratified estimators may not be as precise as simple random estimators. In some instances the cost of sampling is less for stratified random sampling than for simple random sampling. A final advantage of stratified random sampling is that separate estimates for each stratum (e.g., moose density in willows or in forests) are obtained at no extra cost. The **formal procedure** for stratified random sampling consists of 3 steps: (1) clearly specify the strata (they must be mutually exclusive and exhaustive), (2) classify all sampling units into their stratum, and (3) draw a simple random sample from each stratum. Formulas are available to calculate the sample size and optimal allocation of effort to strata (Krebs 1999, Scheaffer et al. 2005). A pilot survey can be analyzed using ANOVA to learn whether stratification is indicated. If cover types define strata, most GIS software will automatically select random coordinates within cover types, making stratified random samples easy to select.

Cluster Sampling

A **cluster sample** is a simple random sample in which each sample unit is a cluster or collection of observations (Fig. 1.8D). This approach has wide application in wildlife biology, because many birds and mammals occur in groups during all or part of the year. When we draw samples from such populations, we draw clusters of observations (i.e., groups of animals). Likewise, many wildlife user groups (e.g., waterfowl hunters and park visitors) occur in clusters (e.g., boats in wetlands and vehicles along highways). Cluster sampling also is useful where cost or time to travel from one sample unit to the next is prohibitive. This situation is common in surveys of animals and habitat. The formal procedure for cluster sampling consists of 3 steps: (1) specify the appropriate clusters and make a list of all clusters, (2) draw a simple random sample of clusters, and (3) measure all elements of interest in each cluster selected.

Making a formal list of clusters is rarely possible or essential. Instead, we emphasize obtaining a random sample of clusters. If the sample units are animals, which naturally occur in groups, the size of the clusters will vary from group to group, depending on the social behavior of the species. Cluster sampling of habitat is performed by choosing a random sample of locations and then locating multiple plots in a cluster at each location. In this case, the researcher sets the cluster size. The optimal number of plots (**cluster size**) depends on the pattern of variability in habitat. If plots in a cluster tend to be **similar** (i.e., little variability in a cluster), cluster size should be small. If plots in a cluster tend to be **heterogeneous** (high variability within a cluster), it should be large. For other types of cluster samples, such as groups of animals or people in vehicles, cluster size is not under control, but is a characteristic of the population. For example, aerial surveys of elk and deer on winter ranges result in samples of animals in clusters. Estimates of herd composition (e.g., fawn:doe or bull:cow ratios) are readily obtained by treating these data as cluster samples (Bowden et al. 1984).

Adaptive Sampling

Adaptive sampling differs from the methods discussed earlier because the sample size is not set at the start of the sampling effort, but rather depends on the results obtained during sampling. Thompson and Ramsey (1983) pioneered adaptive cluster sampling for gathering information on rare animals and plants, which are often clustered in occurrence. In **adaptive cluster sampling** an initial sample of units is drawn by a random or other standard design, and neighboring units also are sampled for any unit that satisfies a criterion, such as having more than x individuals present (Thompson and Seber 1996, Williams et al. 2002*a*, Brown 2003, Thompson 2003). The initial sampling unit and neighbors (where sampled) form **neighborhoods** analogous to clusters and are treated as in cluster sampling. The size of clusters does not need to be constant, nor is it known in advance. For spatially clustered animals or plants, the neighborhood consists of adjacent spatial sample units (Fig. 1.8E). Smith et al. (1995*a*) showed that adaptive cluster sampling would be relatively more efficient than simple random sampling for estimating densities of some species of wintering waterfowl if the right sample unit size and criterion for further sampling in the neighborhood were chosen. The species for which it would be superior show more highly clustered distributions. For other species, conventional sampling designs with fixed sample sizes are superior. Numerous examples of applications of adaptive sampling under conventional sampling designs and estimation methods, as well as applications based on maximum likelihood methods and Bayesian approaches can be found in Thomas et al. (1992), Thompson and Seber (1996), Smith et al. (2003*b*, 2004), and Noon et al. (2006). Thompson et al. (1998), Williams et al. (2002*a*), and Morrison et al. (2008) also review the basic concept and provide simple examples.

Sequential Sampling

Sequential sampling differs from the classical statistical approach in that sample size is not fixed in advance (Wald 2004). Instead samples are drawn one at a time, and after each sample is taken the researcher decides whether a conclusion can be reached. Sampling is continued until either the null hypothesis is rejected or the estimate has adequate precision. This type of sampling is applicable to wildlife studies where sampling is performed serially (i.e., the result of including each sample is known before the next sample is drawn; Krebs 1999). The major advantage of this approach is that it usually minimizes sample size, thus saving time and money. After an initial sample of moderately small size is obtained, successive samples are added until the desired precision is met, the null hypothesis can be rejected, or a maximum sample size under a stopping rule has been reached. This approach typically requires 33% the sample size required in a standard design (Krebs 1999:304). For example, if we wanted to survey deer on a winter range to ensure that harvest had not reduced buck abundance below a management guideline of 5% bucks, we would develop a graph (Fig. 1.9) and plot the results of successive samples as shown (Krebs 1999:312). We must choose a level of significance for our test (e.g., $\alpha = 0.10$) and a power for the test ($1 - \beta = 0.90$) and specify an upper rejection region (>10% bucks), above which we assume the population has not been adversely impacted by buck-only harvests. Once an initial sample of 50 deer has been obtained, sequential groups of deer encountered are added and totals plotted on the graph until the line crosses one of the upper or lower lines or the stopping rule is reached. For example, the lower rejection line is reached at a sample size of 140 (Fig. 1.9). At this point the null hypothesis that bucks constitute >5% of the herd would be rejected, and the conclusion would be there are 5% bucks remaining. An important constraint is the sample must be distributed throughout the entire population, so that a simple random sample of deer groups is obtained. Achieving this sample would be most feasible using aerial surveys from helicopter or fixed-wing aircraft.

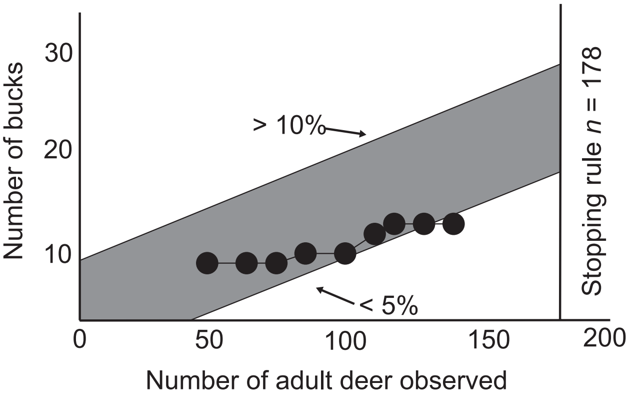


Fig. 1.9. Sequential sampling for percentage of bucks in a deer herd.

Other Sampling Designs

Many other sampling designs are available. For example, **2-stage cluster sampling** involves surveying only a portion of the members of each cluster drawn in the sample. This approach is efficient when clusters are large. Cluster sampling is one version of the more general method referred to as **ratio estimation** (Cochran 1963, Williams et al. 2002*a*). Related methods are **regression estimation** and **double sampling** (Scheaffer et al. 2005) that have great potential for wide application to wildlife research. The interested reader should consult a standard reference on sampling techniques (Scheaffer et al. 2005) and work with a statistician experienced in survey sampling. Stevens and Olsen (2004) proposed a new, efficient approach that combines the advantages of spatially systematic designs with the proven unbiased nature of random sampling. They described this approach as a **generalized random tessellation stratified** (**GRTS**) design. GRTS uses a recursive approach that converts a 2-dimensional map into a 1-dimensional one while maintaining spatial closeness in original locations. This conversion allows a valid systematic sample to be drawn that meets the requirements of random sampling while distributing the sample across the entire spatial area. Theobald et al. (2007) have provided free tools (STARMAP Spatial Sampling tools; http://www.stat.colostate.edu/~nsu/starmap/), which make it feasible to apply GRTS to generating spatially balanced probability-based survey designs.

Sampling Methodology

Plots

Plots are widely used to sample habitat characteristics and count animal numbers and sign. **Plots** represent small geographic areas (circular, square, or rectangular) that are the elements of the geographically defined population. The research population size is the number of these geographic areas (plots) that would cover the entire study area. Sufficient time, money, and personnel to study an entire area are usually not available, and a subset of plots is used with the assumption that it is representative of the area. Any of the survey designs (simple random, systematic, stratified random, cluster, etc.; Fig. 1.8) or more complicated designs, such as 2-stage designs, may be applied (Cochran 1963, Williams et al. 2002*a*). Selecting the best design requires insight into the characteristics and patterns of distribution of species across the landscape. One advantage of using plots is that size of the population is known and totals can be estimated (Seber 1982). Selection of plot size and shape, also an important consideration, has been reviewed by Krebs (1999).

Point Sampling

In **point sampling** a set of points is established throughout the population, and measurements are taken from each sample point (Fig. 1.8F*)*. A common measurement is distance from the point to a member of the population (e.g., plant or calling bird). Examples include point quarter and nearest neighbor methods used widely to estimate the density of trees and shrubs (Mueller-Dombois and Ellenberg 1974), and the variable circular plot or point transect method of estimating songbird density (Reynolds et al. 1980). If observers doing point counts for birds record the distance to each bird detected, as in the variable circular plot approach, transforming distances to areas makes it easy to apply the extensive methods and algorithms developed for line transects referred to as **distance sampling methods** (Buckland et al. 1993, 2001, 2004; Laake et al. 1994). Selection of sample points usually follows a systematic design, but other sample designs can be used, as long as points are spaced sufficiently far apart that few members of the population are sampled more than once. Necessary sample size can be estimated from formulas even if population size is assumed to be large or unknown (Zar 1999).

Transects

A **transect** is a straight line or series of straight line segments placed in the area to be sampled. Transects are used to organize or simplify establishment of a series of sample points or plots and as a sample unit themselves. Transects are widely used to obtain systematic samples of spatially distributed populations (e.g., plants). In these situations, plots along transects are actual sample units (Fig. 1.8G) and should be treated as described for systematic sampling. Plots also can be placed along transects at random intervals. When transects are used as sample units, they are commonly referred to as line transects (Burnham et al. 1980, Williams et al. 2002*a*). Measurements of perpendicular distance, or sighting distance and angle, to the sampled elements (e.g., flushing animals, groups of animals, carcasses, and snags) are recorded (Fig. 1.8H). These distances are used to estimate the effective width of the area sampled by the transect (Seber 1982; Buckland et al. 1993, 2001, 2004). Each transect is treated as an independent observation, and transects should be nonoverlapping according to established sampling designs (e.g., simple random, systematic, and stratified random). Transects are often easier to establish in rough terrain than are plots, but they must be established carefully with a compass or transit and measuring tape or with a GPS unit. Use of transects is becoming more widespread in aerial survey work because of development of precise navigational systems (Patric et al. 1988, Anthony and Stehn 1994, Marques et al. 2006). The critical assumptions for transect methods for sampling such mobile objects as animals (i.e.,100% detection for objects directly on the line and no movement toward or away from the observer before detection) must be examined carefully before this sampling method is selected (Burnham et al. 1980, Williams et al. 2002*a*). In certain cases, more sophisticated methods may be used to adjust counts for less-than-perfect detection on the line (Buckland et al. 1993, 2001, 2004; Manly et al. 1996; Quang and Becker 1996; Williams et al. 2002*a*) or near the points (Kissling and Garton 2006). A strip transect appears similar, but it is really a long, thin plot, because the method assumes all animals or objects in the strip are counted (Krebs 1999).

Road Sampling

Sampling from roads is a widely used method for obtaining observations of species sparsely distributed over large areas or for distributing observations of abundant species over a large geographic area. This sampling method is usually the basis for spotlight surveys of nocturnal species, such as white-tailed deer (Boyd et al. 1986, Collier et al. 2007), black-tailed jackrabbit (*Lepus californicus;* Chapman and Willner 1986), grassland owls (Condon et al. 2005), brood and call counts of upland game birds (Kozicky et al. 1952, Kasprzykowski and Golawski 2009), scent-station surveys (Nottingham et al. 1989, Preuss and Gehring 2007, Mortelliti and Boitani 2008), and the Breeding Bird Survey (Robbins et al. 1986, Sauer et al. 2008). This approach involves drawing a sample from a population defined as that population occupying an area within a distance x of a road (Fig. 1.8I). The distance x is generally unknown and varies with any factor that would affect detection of an animal, such as conspicuousness, density, type of vegetation cover, or background noise for surveys based on aural cues.

Roads rarely provide unbiased estimates for a region, because they are generally placed along ridges or valleys and avoid steep or wet areas. Furthermore, roads modify habitat for many species and may attract some wildlife. For example, during snow periods some bird species will come to roads for grit and spilled grain. Thus, sampling along roads rarely provides a representative sample of habitat (e.g., Hanowski and Niemi 1995) or wildlife populations (Pedrana et al. 2009). Although this bias is well known, it is often ignored in exchange for a method that is cost efficient and easy. As with all indices, every effort should be made to standardize counting conditions along fixed, permanently located routes (Caughley 1977, Sauer et al. 2008); however, this alone does not guarantee reliable counts (Anderson 2001, Thompson 2002*b*). Sampling along roads can be an efficient approach if it is designed as a random sample from a stratum adjacent to roads that is one element of a stratified random sample of the entire area, including other strata distant from roads (Bate et al. 1999, Langen et al. 2009).

Dependent (Paired) and Independent Observations

If we wish to make population comparisons, **pairing observations** is a powerful tool for detecting differences. If there is a correlation between members of a pair, treating them as dependent or paired observations can improve the power of tests for differences. For example, to compare diets of adult female mountain sheep (*Ovis canadensis*) and lambs, we might treat a ewe with a lamb as a pair and measure the diet of each animal by counting the number of bites of each plant they eat while foraging together. Treating these observations as pairs would sharpen comparisons between age classes, because it would compare animals foraging together and experiencing the same availability of plants. Pairing is a powerful technique in other contexts for which there is dependency between the observations. Pairing should be used only if an association really exists, otherwise the power of comparison will be decreased.

Pairing also can be used to help answer a different question. For example, studies of habitat selection are often made by locating areas used by a species (i.e., nest sites or radio locations) and measuring habitat characteristics at these use sites with sample plots. Available vegetation types are measured from random sample plots throughout the study area (Fig. 1.10A). A comparison of use and random plots can identify characteristics of areas selected by the species. An alternative approach involves pairing use and random plots by selecting a random plot within a certain distance of the use plot (Fig. 1.10B). For analysis, use and random plots are paired (i.e., random plot locations are dependent on use sites). This comparison could produce different results from the unpaired comparison, because it tests for habitat differences in areas used by the species (microhabitat selection). In contrast the unpaired comparison (e.g., independent plots) tests for habitat differences between areas used by the species and typical vegetation types available in the general study area (macrohabitat selection). Choosing a paired or unpaired design will depend on the objectives of the study, but both may be useful when applying a hierarchical approach to studying habitat selection (Wiens 1973, Johnson 1980, Cruz-Angón et al. 2008, Schaefer et al. 2008).

CONFRONTING THEORIES WITH DATA

Confronting theories with data involves evaluation and interpretation, which is a creative phase, similar to hypothesis formulation. The quality of conclusions drawn is dependent on the biologist's past educational and professional experience as well as a willingness to consider standard and less traditional interpretations. One great danger in wildlife science (and other fields) is that researchers often have a conscious or unconscious expectation of results. This **bias** might begin with the development of the overall research objective and carry through to the interpretation phase. This danger is so great that in some fields, such as medicine, experiments are performed with a double-blind approach: neither researcher nor subjects know membership of treatment and nontreatment groups. A scientist must not design research or interpret data in a way that is more likely to support preconceived explanations of biological systems. Biologists who are consciously aware of their own biases and strive to keep an open mind to new ideas are most likely to make revolutionary discoveries.

The objective is to organize, clearly and concisely, the results of data collection, exploratory data analysis, and specific statistical analyses. These results must be transformed from a collection of specific information into a **synthesis** explaining the biological system. Do statistical evaluations support one or more of the theories and hypotheses and clearly reject others? Do the results provide a reasonable ex-

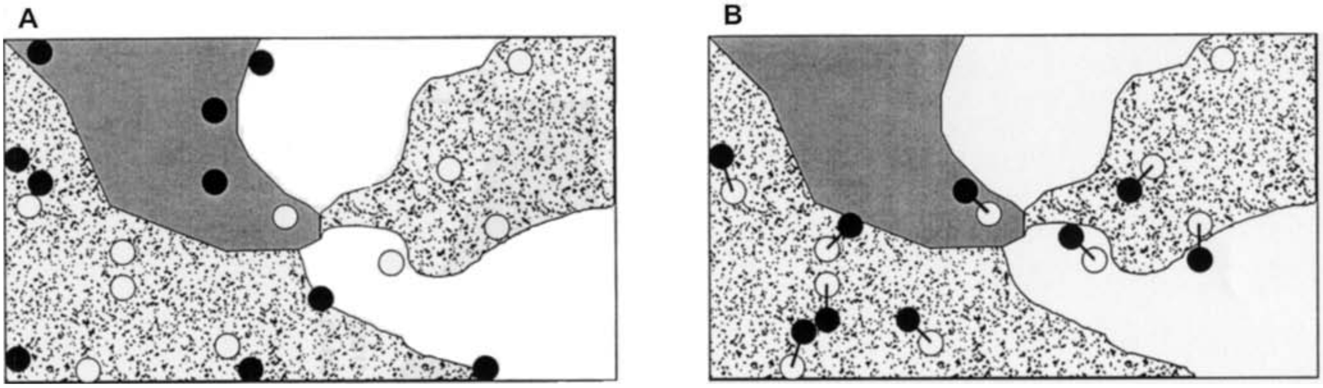


Fig. 1.10. Illustrative examples of (A) use (○) and random plots (●), and (B) use plots paired with random plots.

planation of the biological system? Are there alternative explanations of the data and statistical tests? Are there specific problems with the data that should be identified, such as inadequate sample sizes or unusual variation in specific variables measured? What could have introduced bias into the estimates? Are additional data required? These questions must be considered carefully, and if concerns are identified, they must be noted in reports and publications.

During this phase, the biologist usually reaches some conclusions based on the data and results of statistical evaluations. If the data support the hypothesis, we cannot conclude the theory (model) is true, but only that it has not been rejected (James and McCulloch 1985). The central issue is that we do not prove a research hypothesis or theory to be correct—indeed some would argue that *all* hypotheses are to some degree incorrect. Instead, the **credibility** of the hypothesis increases as more of its predictions are supported and alternative hypotheses are rejected. We can assist other biologists by carefully considering how broadly our conclusions can be generalized to other areas or populations and not allowing our conclusions to go beyond the data. Interpretation of research data must clearly separate conclusions and inferences based on data from speculation. For example, if we demonstrate that droppings from spruce grouse are most abundant under lodgepole pine (*Pinus contorta*) and Engelmann spruce (*Picea engelmannii*), we can conclude that grouse use both tree species for some behaviors, but the type of behavior (e.g., roosting or feeding) is mere speculation without additional data (e.g., observations of feeding activity and crop or fecal analyses). Likewise, **replication** of studies across space and time "provides us greater confidence that certain relationships are general and not specific to the circumstances that prevailed during a single study" (Johnson 2002:930).

Data Collection

Most novice research biologists are anxious to initiate data collection because of the attractiveness of working outdoors and the pleasure derived from observing wildlife-related phenomena. However, the design phase should not be rushed to initiate fieldwork more quickly. Successful research biologists often spend about 40% of their time in design and planning phases, 20% in actual fieldwork, and 40% in data analysis and writing publications. Data collection can be physically difficult and highly repetitious.

All data should be recorded on preprinted data sheets or entered directly into a handheld data logger, computer, or personal digital assistant. This practice ensures that each field person collects exactly the same data, as **consistent** collection of data simplifies analysis. Data sheets should be duplicated after each field day (e.g., computer entry, photocopies, or transcribed) and stored in a separate location from the original data set. Data entered electronically in the field should be downloaded daily and backed up for storage at another location. Transcription of data (including computer data entry) must be followed by careful proofreading, which is greatly facilitated by checking for valid entries by using database queries and spreadsheet scripts. All field personnel should receive careful instructions regarding data collection, and the principal researcher must check periodically to see that each person has similar skills and uses the same methods for observation, measurement, and recording (Kepler and Scott 1981). The principal researcher is responsible for quality control, and the validity of research results depends on the quality of research design and data collection.

Pilot Study

A **pilot study** is a preliminary short-term trial through all phases of a research project. Pilot studies are an important, but often neglected step in the research process. Information can be obtained that will help the researcher avoid potentially disastrous problems during or after the formal research phase. Pilot studies often will disclose hidden costs or identify costs that were over- or underestimated. **Optimal sample allocation** (Scheaffer et al. 2005) incorporates cost estimates to maximize the benefits obtained from limited research budgets. Use of a pilot study should reveal basic logistical problems, such as travel time among study plots being underestimated or expectations for overall sample sizes being infeasible without additional personnel and funding. Statistical procedures for estimating needed sample sizes require variance estimates of variables that will be measured, and these variance estimates are often available only from data gathered in a pilot study. These preliminary data might disclose the variance of the population is so large that obtaining adequate sample sizes will be difficult. It is far better to discover these problems before time, energy, personnel, and critical research dollars are committed to a research project doomed to fail. If the research is part of an ongoing project, or if much research on the topic has been published, costs, methodology, and variance estimates may already be firmly established.

Power Analysis

In descriptive studies, power analysis provides sample size requirements for obtaining an estimate of desired precision and can be calculated after an estimate of population variance is obtained from previous studies or a pilot study. Formulas for sample size are available for standard survey designs (Thompson et al. 1998, Scheaffer et al. 2005) and for typical hypothesis tests (Zar 1999). In studies involving experiments or other types of comparisons, sample size is increased to improve the **power of a hypothesis test** (defined as the probability of detecting a real difference) and to prevent erroneous conclusions. Power analysis for hypothesis tests depends on several factors, including sample size, level of significance (α), variance in the populations, effect size (the true change that occurred), and efficiency of the test or design (Steidl et al. 1997). In contrast to this essential prospective power analysis during the design phase, performing a retrospective power analysis after the data are collected, during the analysis phase, is controversial or contraindicated (Thomas 1996, Steidl et al. 1997). Retrospective power analysis is uninformative unless effect sizes are set independently of the observed effect (Steidl et al. 1997).

To illustrate power of a test, consider the following example. Suppose we were using fawn:doe ratio as an indicator of production for a mule deer (*Odocoileus hemionus*) herd (i.e., the biological population is our research population). We want to know whether the fawn:doe ratio has declined. There are 4 possible outcomes from sampling the herd and testing for a decline in the fawn:doe ratio (i.e., the null hypothesis is there is no change; Table 1.3). We evaluate whether the fawn:doe ratio has declined by comparing the test statistic calculated from our data to a value for this statistic at the chosen level of significance (α). The **level of significance** represents the chance of concluding the ratio changed when in fact it did not. An $\alpha = 0.05$ indicates that we would make this error only 5 times if the population really did not decline and we tested it by drawing a sample 100 times. This error is referred to as a **Type I error.** But, we could make another error. We could conclude the ratio had not changed when in fact it had declined. For the situation where we count 500 deer, we would fail to detect the decline in the fawn:doe ratio 50% of the time (Table 1.3). This type of error is referred to as **Type II error,** and its likelihood is measured by α. When we perform a test, we typically set α low to minimize Type I errors. But, Type II errors might be as important (Alldredge and Ratti 1986, 1992) or even more important than Type I errors. Obviously, we want to detect a change when it occurs; the probability of detecting a change is called the **power of the test.** The power of the test is calculated as the probability of not making a Type II error ($1 - \alpha$).

We cannot control natural variation in the population or the actual change that occurred, but we can control the other 3 factors (i.e., sample size, efficiency, and significance level). **Parametric tests** (based on a normal distribution, e.g., *t*-tests, *F*-tests, and *Z*-tests) have the highest efficiency for normally distributed populations and for large samples. **Nonparametric tests** (based on distributions other than the normal distribution, e.g., Mann-Whitney, Wilcoxon signed-ranks tests) are superior when sample sizes are small (30) and populations are not normally distributed (Johnson 1995, Cherry 1998). The power of a test declines as the level of significance is made more stringent (decreasing α). In the example (Table 1.3), this problem is critical, because the Type II error (failing to detect declining production) is the more serious error than detecting a declining production when it is actually increasing. It would be preferable to increase α so that power of the test could be increased. In other situations the Type I error will be more serious, and α must be kept low. Increasing sample size increases power of the test. Calculating sample size necessary for a desired level of power is essential to designing a high quality study (Toft and Shea 1983, Forbes 1990, Peterman 1990). However, such calculations should be based on meaningful effect sizes (i.e., one that constitutes a biologically significant result; Reed and Blaustein 1997, Cherry 1998, Johnson 1999).

The importance of **sample size** cannot be overemphasized. Sample size and experimental design are the major factors under the control of the biologist that strongly influence power of the test (i.e., the likelihood of detecting a significant difference when one really occurs). Inadequate sample size usually results from: (1) inadequate consideration of population variance; (2) inability to collect data (e.g., observe a rare species); or (3) insufficient funding, time, or personnel. Often a sample size problem is overlooked initially because of failure to consider sample size reduction throughout the study (i.e., we focus mostly on the initial sample size and not on the final sample size that represents the most important data for consideration of a hypothesis). For example, in a study of mallard (*Anas platyrhynchos*) brood movements almost 10 times as many nests were required to be found as the sample size of broods indicated because of an

Table 1.3. Possible outcomes of a statistical test for declining production in a deer herd. Counts of 500 antlerless deer (adult does and fawns) were obtained each year, and tests of the null hypothesis of no change in the fawn:doe ratio were performed at the 5% level of significance ($\alpha = 0.05$).

	Fawns per 100 does							
	Actual herd value			Count value				
Case	188	1989	Change	1988	1989	Conclusion from test	Result of test	Likelihood of this result (%)
1	60	60	None	61	59	No change	No error	95 ($1 - \alpha$)
2	60	60	None	65	50	Declined	Type I error	5 (α)
3	65	50	Declined	65	50	Declined	No error	50 ($1 - \alpha$)
4	65	50	Declined	62	57	No change	Type II error	50 (α)

89% sample size reduction from nests located to actual brood data (Rotella and Ratti 1992*a, b*).

Another common problem is fairly large overall data sets that are not sufficiently similar across years (or seasons) to combine, resulting in annual sample sizes that are too small for analysis. At the beginning of a research project we often set our desired sample size based on combining data collected over several continuous years. However, if the characteristic of interest were different across study years, combining the data would not be valid. For example, in a study of habitat selection by red fox (*Vulpes vulpes*), habitat use might differ between mild and severe winters with heavy snow cover. In this example, combining the data would not be valid, yet the sample size in each year may be too small to detect selection (Alldredge and Ratti 1986, 1992).

Approaches to Data Analysis

At this point, researchers have developed well-planned and biologically meaningful hypotheses; decisions have been made regarding study, experimental, and sampling designs; and empirical data have been collected to shed light on the validity of the hypotheses. Now researchers must decide on a statistical approach. Unfortunately, this decision has become less clear over the past decade (Butcher et al. 2007). General approaches for data analysis include Bayesian versus frequentist paradigms with distinct differences in how probability should be interpreted (Cox 2006). Within the **frequentist** paradigm, one could choose null hypothesis significance testing (**NHST**), point and interval estimation of effect sizes, likelihood-based and information theoretic methods, or some combination of these (Läärä 2009). Unfortunately, the statistical approach that is most familiar and widely used (i.e., NHST) in wildlife science has continued to be criticized (e.g., Yates 1951, Cherry 1998, Johnson 1999, Wade 2000, Fidler et al. 2006, Läärä 2009), causing confusion and frustration for researchers (Butcher et al. 2007). We introduce these various approaches and point out some of the key differences while purposefully not recommending one over another. We think it is more important to expose researchers to the relevant discussions, so they can make an informed selection of the best approach.

Ellison (2004) summarized the main differences between Bayesian and frequentist approaches to statistical inference (also see Ellison 1996, Dennis 1996, Taper and Lele 2004, Hobbs and Hilborn 2006). The first is a difference in what is considered a random outcome. **Frequentist inference** considers the model and the true parameter values to be fixed quantities, whereas the observed data are random outcomes from this process. Thus, frequentists refer to the probability of the data (Y) given a particular hypothesis (H), as defined by the model and parameters: $\text{Prob}(Y|H)$. In contrast, **Bayesian inference** treats both the data and model as random, allowing quantification of the probability of a hypothesis being true given the observed data: $\text{Prob}(H|Y)$.

This distinction brings up the second major difference between these approaches—the definition of ***probability.*** Frequentist inference defines probability as the relative frequency of a particular outcome if the process was repeated an infinite number of times. For example, the probability of obtaining a heads with a flip of a coin is the number of times a head turns up divided by the number of flips, where the number of flips is repeated to infinity. Bayesian approaches define probability quite differently: it is the degree of belief in the likelihood of an event occurring.

Finally, the 2 approaches differ in the way **prior knowledge** is incorporated. For Bayesian inference, it is required that prior knowledge is translated into a probability distribution, which is then combined with the sample data to make an inference. Frequentist inference generally uses only the observed data, although prior knowledge can be incorporated by combining likelihoods from previous studies with the likelihood of the observed data (see Hobbs and Hilborn 2006:10). Although the decision of whether to use Bayesian versus frequentist approaches is often made on practical grounds (Lele et al. 2007), we end with a quote from Ellison (2004:517) that we believe is particularly relevant:

> Deciding whether to use Bayesian or frequentist inference demands an understanding of their differing epistemological assumptions. Strong statistical inference demands that ecologists not only confront models with data, but also confront their own assumptions about how the world is structured.

Hypothesis Testing

Significance testing as a statistical approach for confronting hypotheses with empirical data has been the subject of fervent debate in many disciplines (Fidler et al. 2004), including wildlife and ecological science (e.g., Anderson et al. 2000, Eberhardt 2003, Guthery et al. 2005, Lukacs et al. 2007, Steidl 2007, Stephens et al. 2007, Läärä 2009). Nonetheless, it remains a viable option for practicing wildlife researchers (Robinson and Wainer 2002, Butcher et al. 2007). **Hypothesis testing** is rooted in the philosophical idea of falsification, in which an attempt is made to disprove a hypothesis, leaving the alternative to be tentatively accepted (Underwood 1997). Johnson (1999) described the 4 basic steps of statistical hypothesis testing that mirror the approach suggested by Underwood (1997). The researcher develops a hypothesis that reflects his or her ideas about a particular ecological process or the effects of some treatment. The logical opposite of this hypothesis is usually taken as the **null hypothesis,** and data are collected to assess the validity of the null hypothesis. A statistical test of it involves calculating a *P*-value, which is then used to decide the fate of the null hypothesis. Strictly speaking, a ***P*-value** is the probability that if the null hypothesis were true and the test were hypothetically redone, one would observe data at least as extreme as those which were observed. Thus, a study that results in a *P*-value of 0.05 means that if the null hypothesis were true and the study were repeated 20 times, you would expect only 1 of these 20 studies to produce results at least as different from the null hypothesis as your study. Obviously, the definition is quite cumbersome and likely has led to much confusion, misuse, and misinterpretation of a statistical hypothesis test (Johnson 1999).

To more fully understand the role of hypothesis testing in wildlife science, it is helpful to have some historical perspective. Robinson and Wainer (2002) provide a concise description of hypothesis testing as it was originally intended by the famous statistician R. A. Fisher, who used it to assess potential innovations in agriculture. A few key points from this description are:

1. It is often legitimate to assume a particular innovation would produce no effect, and thus testing a null hypothesis of no effect is not considered trivial.
2. No single test should be the end of the discussion, because there is a chance (depending on the significance level for a particular test) that an effect can be suggested even when there is none, an effect should only be accepted if **repeated studies** continue to provide significant results.
3. Hypothesis testing only makes sense if continued research seeks to identify the **size** and **direction** of the effect.

Given these original intentions, it is not hard to see why so many have been critical of hypothesis testing in wildlife science. Several have argued that it is exceedingly rare to legitimately propose a zero effect or alternatively that some set of parameters are exactly equal (Cherry 1998; Johnson 1999; Anderson et al. 2000, 2001*a;* Läärä 2009). These point null hypotheses are often deemed silly nulls, because they are almost certain to be false a prori. Additionally, although replication was a cornerstone of Fisher's approach, true replication in wildlife science is not the normal procedure, which instead relies on "single-shot studies" designed to reach conclusions based on a one-time interpretation of a *P*-value (Robinson and Wainer 2002:265). Although **replication** is an important component of the scientific method regardless of the statistical approach used, because of the definition of a *P*-value, it is particularly relevant to hypothesis testing. These issues are especially problematic when hypothesis testing is applied to field studies without random assignment of treatments. Many statisticians strongly object to performing hypothesis tests on observational data or recommend alternative approaches for evaluating the data, such as confidence intervals for estimates, information measures for models, or Bayesian confidence measures (Cherry 1998, Johnson 1999, Anderson et al. 2000, Hobbs and Hilborn 2006, Läärä 2009).

Despite these criticisms, most statisticians agree that hypothesis testing can play a valuable, but limited role in data analysis (Cherry 1998, Johnson 1999, Stephens et al. 2007), especially if accompanied by estimates of **effect sizes** and a measurement of the **precision** of these estimates (Robinson and Wainer 2002). One improvement might be for researchers to adopt a trinary decision approach that is likely a more productive route than interpreting results of a hypothesis test (Jones and Tukey 2000). Using this approach, the conclusions of a hypothesis test are either $\mu_1 > \mu_2$, $\mu_2 > \mu_1$, or the direction of the difference is undetermined. Using this language avoids the temptation to accept a null hypothesis that is likely untrue while stressing the need for continued research to determine the direction and magnitude of the effect (Robinson and Wainer 2002).

Information-Theoretic Model Selection

Information-theoretic model selection offers a distinct alternative to hypothesis testing, and the approach has seen widespread growth in wildlife and ecological sciences (Hilborn and Mangel 1997, Burnham and Anderson 2002, Johnson and Omland 2004, Richards 2005). In contrast to hypothesis testing, model selection seeks to identify the hypotheses that are closest to the truth out of a **set of competing ideas** while fully acknowledging that all are wrong or incomplete characterizations of the process. The philosophical basis for this approach is more in line with that of Lakatos (1978:24): "All theories . . . are born refuted and die refuted. But, are they equally good?" He considered it nonsensical to retain only unfalsified hypotheses because of the philosophy that hypotheses may never be truly falsified and, more importantly, science will keep a hypothesis that is known to be wrong if there is not a better one available to take its place. Thus, a hypothesis is falsified only if a better one with greater empirical support is available to replace it. The information-theoretic model selection approach also closely follows Chamberlin's (1890, 1965) view of science by advocating the construction of multiple working hypotheses that are subject to repeated confrontation with empirical data. Those supported by the data tend to be retained, whereas those with little support tend to be dropped from consideration (Burnham and Anderson 2001).

Using the information-theoretic model selection approach, several competing models are suggested to reflect different hypotheses about how a process works or the effects of a particular treatment. An appropriate study is designed to collect empirical data that will be used as the arbiter in a **contest among rival hypotheses.** The metric for deciding among hypotheses is how close each model is to the truth. Due to an explicit link with information theory (hence "information-theoretic"), Kullback-Leibler distance has been promoted as an appropriate measure of the distance each competing model is from the true data-generating model (see Burnham and Anderson 2002:50–54). Several criteria may be used to estimate the relative expected Kullback-Leibler distance (Shibata 1989, Burnham and Anderson 2002), including Takeuchi's information criteria; likelihood cross-validation criteria (Stone 1977); and Akaike's information criteria (**AIC;** Akaike 1973), which are the most common in the wildlife and ecological literature. By focusing on the best explanation for an observed phenomenon, information-theoretic model selection does not rely on a binary decision process characteristic of hypothesis testing, instead allowing models to be differentiated according to the amount of support they receive from the data. Several practical guidelines for using information-theoretic model selection approaches have been published (Anderson et al. 2001*a,* Anderson and Burnham 2002, Richards 2005). In addition to the comprehensive treatment in Burnham and Anderson (2002), see Guthery et al. (2005) for a more critical review.

Effect Size and Interval Estimation

Most researchers agree that hypothesis testing and model selection are only one component of statistical inference and that estimation of effect sizes and measures of their precision are at least as important (Johnson 1999, Robinson and Wainer 2002, Stephens et al. 2007). Quinn and Dunham (1983:613) suggested: "The objective of biological research typically is to assess the relative contributions of a number of potential causal agents operating simultaneously." If this is the case, then estimation of **effect sizes** is of primary importance to wildlife science and these results should be emphasized in data analysis. Others have echoed this sentiment: "The very basic tools for statistical reasoning on the strength of associations and the sizes of differences and effects are provided by point estimates, their standard errors and associated confidence intervals" (Läärä (2009:152). Reporting effect sizes is not only important for practical interpretation of the focal study, but they also are the critical components for any subsequent meta-analysis (Gurevitch et al. 2001, Hobbs and Hilborn 2006). Läärä (2009) contains several practical recommendations for presenting and interpreting effect sizes that should be especially useful to practicing wildlife professionals.

Regression and General Linear Models

One of the most flexible approaches to identifying predictive and potentially causal relationships between wildlife and environmental or management characteristics involves use of ordinary least squares to estimate parameters of **regression models** or **GLM** (Fig. 1.11). Experimental manipulations that produce different levels of predictor variables are more readily analyzed by ANOVA, regression, or analysis of covariance versions of GLM under a Fisherian philosophy (Fig. 1.11), named after R. A. Fisher, who pioneered a "spirit of reasonable compromise, cautious, but not overly concerned with pathological situations" (Efron 1998:99) in the analysis of experiments. Designing a study to gather data on a variety of potential causal variables rather than manipulating those variables through a designed experiment is an appealing alternative, but yields inferences of much lower certainty (Fig. 1.5). Performing hypothesis tests on such data (e.g., testing point null hypotheses) is easily performed with modern regression programs. However, it may not be justified as an inferential approach and may readily lead into a "fishing expedition" doomed to failure due to high Type I errors. Many statisticians refuse to analyze such data by using hypothesis tests and instead encourage biologists to apply maximum likelihood and information-theoretic model approaches under a modeling perspective, that is, identifying the most parsimonious model with good predictive ability (Milliken and Johnson 1984, Anderson et al. 2000, Burnham and Anderson 2002).

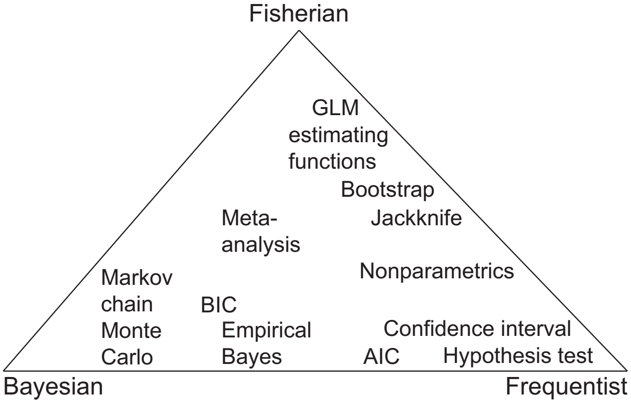


Fig. 1.11. Selecting analysis methods from 3 dominant statistical philosophies. GLM = general linear model, BIC = Bayesian information criteria, AIC = Akaike's information criteria. *Modified from Efron (1998).*

It is essential in designing manipulative or observational studies, if one plans to estimate linear models, to strive to obtain observations throughout the full range of the predictive variables. It is especially important to obtain observations at both low and high values of the predictive variable, because they set limits for the range of values that can be used later for prediction. The values at the ends of this range have the most leverage on slope estimates. If too narrow a range is measured, a significant relationship may not be detected among the variability. However, a relationship may be linear only through a portion of its range, such that beyond a certain level an increasing effect may turn into a negative effect at progressively higher levels. Such situations should be apparent from exploratory data analyses (Anderson 2001, Johnson 2002).

Bayesian Approaches

Bayesian data analyses are described as "practical methods for making inferences from data using probability models for quantities we observe and for quantities about which we wish to learn" (Gelman et al. 2003:3). One of the primary appeals of Bayesian statistics is that after sampling a population and calculating statistics, such as the mean, variance, and confidence interval for the mean, Bayesian analysis allows us to state that we are 95% certain the true mean for the population is within this 95% confidence interval. Johnson (1999) provided an easily understood description of the conceptual differences between the frequentist and Bayesian approaches (Fig. 1.11). A Bayesian analysis requires performing 3 basic steps (Gelman et al. 2003).

1. **Specify a probability distribution for all quantities** (i.e., use prior studies and creative thinking to specify a particular **prior** probability for the parameter[s]). We begin by stating the range of all possible values for the characteristics we are attempting to measure and make our best guess of a probability distribution for a parameter (based on earlier studies and clear thinking) if any values are more likely than others. This step is controversial, because it introduces subjective decisions into the process and has potential for misuse if one's goal is to cook the books to produce a particular result (Dennis 1996). However, well-designed research should gather historical data, so that knowledge is available on the probability distribution of the parameter(s) (Box 1.1) or failing that, it should specify minimum and maximum values with equal chances of intermediate values (i.e., a flat prior).

2. **Use the observed data to calculate a posterior distribution for the parameter of interest as a conditional probability distribution.** This second step in Bayesian analysis follows data collection. We improve our prior guess of the value of the characteristic by combining it with the new data gathered to state conclusively our best posterior guess of the value of the characteristic. This step is performed using **Bayes's rule,** and this Bayesian estimate might be considered as a weighted average estimate based on the sample data and the assumed prior value, where weights are proportional to the precision of the observed and prior values (Gelman et al. 2003:43). As sample size increases, the Bayesian value approaches the maximum likelihood estimate and any influence of the prior probability vanishes. **Markov chain Monte Carlo** methods are used widely for these calculations (Fig. 1.11).

3. **Evaluate the fit of the model and the implications of the resulting posterior distribution.** This step in Bayesian analysis (Gelman et al. 2003:3) consists of "evaluating the fit of the model and the implications of the resulting posterior distribution: does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions?"

Validating Parametric and Simulation Models

The validation and experimental phases of the modeling process described earlier really constitute an effort to **confront theories with data.** The models express our theoretical understanding of the system, its characteristics, and its processes. Validation and experimentation confront this theory with data, especially when we conduct these activities in an adaptive management framework, where management actions are accompanied by monitoring to simultaneously validate the predictions of the models (our theory or understanding of the system) and probe the behavior of the system (Walters 1986:250). Comparing model predictions to data potentially completes the feedback loop that can be used to improve our understanding, but the natural tendency of managers and biologists is to break the loop by ignoring any inconsistencies detected. This tendency is natural because of the considerable effort expended in developing the models and trade-offs in selecting management actions. Ignoring inconsistencies leads to passive adaptation rather than a probing through experimental management actions. "Conservative, risk-averse decision making creates a particularly difficult situation for learning" (Walters 1986:251), especially when the effects of management are compounded with environmental changes and there are lags inherent in the responses. Where the desired outcome is a harvestable surplus of a game species, the manager and biologist face substantial social, economic, and political pressure to find the "right" answer (see the section Adaptive Management: Connecting Research and Management below). Models are invaluable in efforts to ensure that management and ecological understanding are based on valid estimates and relationships rather than on wishful thinking, but their results are often attacked by the interested public, whose values and "gut instincts" are opposed to model predictions.

SPECULATION AND NEW HYPOTHESES

Rarely does a single research project provide the last word on any problem (Johnson 2002). More commonly, research will generate as many questions as answers. Speculation, based on inconclusive or incomplete evidence, is one of the most important aspects of science. **Speculation** must be identified and should not be confused with conclusions based on data. But, speculation is the fuel for future research. Many facts of nature have been discovered by accident—an unexpected result from some associated research effort. However, most research is directional (i.e., it attempts to support or falsify a theory reached by speculating from facts). **New hypotheses** can be considered a form of speculation, which is verbalized in a more formal fashion than speculation and has a specific testable format. For example, considering spruce grouse, we can formulate a basically untestable hypothesis that spruce grouse have evolved a preference for use of lodgepole pine and Engelmann spruce trees. This statement is too vague and requires historical data that cannot be collected. However, we can hypothesize that spruce grouse use lodgepole pine and Engelmann spruce trees for: (1) feeding or (2) roosting. Testing these hypotheses, we might learn that 80% of the spruce grouse diet is lodgepole pine, even though Engelmann spruce is more abundant. We may then speculate (i.e., hypothesize) that needles from lodgepole pine provide higher nutritional quality than needles from Engelmann spruce.

PUBLICATION

The final step of the scientific method is publication of research. Unfortunately, many research dollars are wasted, because knowledge gained was not published and the information is buried in file cabinets or boxes of data sheets. The **publication process** is the most difficult phase for many biologists. Clear concise scientific writing is difficult, because most biologists have little formal training in or inclination for this activity. Peer review may damage the ego, because we must subject our work to anonymous critiques used by editors to judge whether the manuscript is acceptable for publication.

Agency administrators often do not encourage or reward employees for publishing their work and discourage publication in some instances. Administrators are pressured with calls for immediate answers to management problems; thus, they devalue the long-term benefits of the publication process. Effective administrators recognize that **peer review** and **publication** will: (1) **correct errors** and possibly lead to a better analysis, (2) help authors reach the most **sound conclusions** from their data, (3) make it easier to **defend** controversial policies, (4) help their personnel **grow** as scientists by responding to critical comments and careful consideration of past errors (which may have been overlooked without peer review), and (5) make a **permanent contribution** to wildlife management by placing results in a literature format available to other agencies, researchers, and students.

Publication is essential to science. Peer reviews normally improve the quality of a manuscript, but some research may not be suitable for publication. This observation emphasizes the importance of careful planning, design, data collection, etc. Rarely would any research effort that is properly planned, designed, and executed (including a well-written manuscript) be unpublishable. However, the revision process (i.e., responding to criticisms from the editor and referees) may be painful and frustrating to authors. Overall, the system is necessary to ensure quality publications, and authors should not be discouraged by the necessity to defend their work and revise manuscripts. Research is not complete and does not make a contribution to knowledge and sound management of wildlife resources until results are published in a way that effectively communicates to the scientific community and user groups (e.g., wildlife managers). In addition to publication in peer-reviewed journals, research findings will improve wildlife management immediately if they are communicated in **other forums,** such as professional meetings, workshops, seminars, general technical reports, informational reports, and articles in the popular press.

COMMON PROBLEMS TO AVOID

Procedural Inconsistency

Procedural inconsistency is another common problem that can be prevented with proper research design. Problems of this type occur from seemingly minor variations or alterations in methodology. For example, if a project is dependent on field personnel to accurately identify songs of forest passerine birds, the data set may be biased by identification errors (Cyr 1981). In this situation, the magnitude of the bias will depend on the rate of errors by individuals, differences in the rate of errors among individuals, and relative proportion of data collected by each individual. Research methodology should be defined with great detail, and all individuals collecting data should have similar skills and knowledge of the methods used (Kepler and Scott 1981). If inconsistencies cannot be eliminated through selection and training of field workers, the design must incorporate double sampling or similar procedures to remove inherent biases (Farnsworth et al. 2002). One unfortunate aspect of biases of this type is they are often overlooked (or ignored) as potential problems and are seldom reported in research publications.

Nonuniform Treatments

A common bias stems from **nonuniform treatments.** This problem is illustrated by considering 2 previous research examples. In the discussion of crossover experiments, we described a 2-year study on pheasant nest success, in which mowing on treatment areas was delayed until after 4 July. Assume that in the first year of this study, all treatment areas were mowed between 4 and 7 July, as planned. But, during year 2 of the study, a 3-day rainstorm began on 4 July, and the treatment areas were not cut until 9–12 July. Although this 5-day difference in mowing the treatment areas may seem insignificant, the impact on the results and interpretation of our experiment is really unknown—and may be serious. Thus, the second year of the experiment should be repeated. Because dates of pheasant nesting and plant growth varies from year to year in response to temperature and rainfall patterns, a better way to set the date for the mowing treatment might be based on the cumulated degree-days widely published in farm journals.

In the second example, we want to evaluate effects of sharp and feathered edges on nest success of forest birds. If we had used both clearcuts and road ways as sharp edges, we might have hopelessly confused the treatment results because of differences in the attractiveness of sharp edges near roads, where carrion is an abundant attractant to such generalist predators as ravens. High variability between replicates in nonuniform treatments substantially reduces our power to detect biologically significant effects.

Pseudoreplication

Pseudoreplication occurs when sample or experimental units are **not independent** (i.e., they are really subsamples rather than replicates, but are treated as though they were independent samples or experimental units). This problem is widespread in field ecology (Hurlbert 1984) and should be avoided when possible. Experimental units are independent in manipulative experiments only if we can **randomly assign treatments** to each unit. In field studies, a simple test for pseudoreplication is to ask whether the values for 2 successive observations are more similar than values for 2 observations drawn completely at random from the research population (e.g., Durbin and Watson 1971). If so, the successive observations are probably not true replicates and the research should be **redesigned,** or this lack of independence must be treated correctly in the analysis. This treatment can be done by using cluster sampling, adjusting the degrees of freedom for tests (Porteus 1987, Cressie 1991), or applying Monte Carlo approaches to evaluate test statistics (as is widely done for spatially correlated data; Dale and Fortin 2002).

There must be a direct tie between the sample or experimental unit and the **research population.** If the research population consists of 1 meadow in Yellowstone National Park, then 2 or more samples drawn from that meadow would be replicates. In this example, our inferences or conclusions would apply only to that single meadow. If our research population consisted of all meadows in Yellowstone National Park, then 2 plots in the same meadow would not constitute true replicate samples. Also, repeated sampling of the same radiomarked animal often constitutes a form of pseudoreplication (e.g., if our research population consisted of moose in one ecoregion, repeated observations of habitat use by a single animal would not be true replicates; a similar problem would arise if 2 radiomarked animals were traveling together, so their habitat selection would not be truly independent). The data would have to be summarized into a single value, such as the proportion of observations in a certain habitat, for statistical analysis. This compression would reduce the sample size to the number of radiomarked moose. Treating repeated observations as replicates is strictly justified only when the individual animal is the research population. In this situation, tests for **serial correlation** (Swihart and Slade 1985) should be conducted to ensure that observations are not repeated so frequently they are still pseudoreplicates.

ADAPTIVE MANAGEMENT: CONNECTING RESEARCH AND MANAGEMENT

Wildlife management programs should be developed from the application of scientific knowledge based on research (i.e., we should apply scientific facts and principles resulting from research on specific topics, e.g., population ecology, habitat selection, or behavior). Initially, this practice is a sound one for the development of a new management program. The logic behind formulation of a management program is similar to the formulation of a research hypothesis: both provide opportunity for predictive statements. Our management prediction is that our plan of action will achieve a desired result. However, a major problem with nearly all wildlife management programs throughout the world is the lack of research on the effectiveness of programs (Macnab 1983, Gill 1985). Seldom is the question "does our management lead to the desired result?" addressed in formal, well-designed, long-term research projects. For example, research indicates that spinning-wing decoys make mallard breeding populations more vulnerable to harvest (Szymanski and Afton 2005). A potential long-term management response would be to create more restrictive hunting regulations as the use of spinning-wing decoys increases. The assumption is that if using spinning-wing decoys increases mallard harvest rates, then hunting regulations are needed to ensure mallard populations over the long term do not shrink with increased vulnerability. However, we should consider several important questions. Does increased vulnerability translate to increased harvest? What segments of the mallard populations are most vulnerable to the use of spinning-wing decoys? Will mallards become accustomed to spinning-wing decoys over time and thereby decrease their vulnerability to harvest? These questions and more should be addressed, because imposing more restrictive hunting regulations could backfire if the answers to these questions do not support it.

A second common example is prescribed burning as a management practice to increase deer and elk populations. The effectiveness of this management has not been addressed directly, and most evaluations have only noted increases in browse forage species and changes in animal distributions (Stewart et al. 2002, Van Dyke and Darragh 2007, Long et al. 2008*a, b*). Increased population levels in response to prescribed burning have not been adequately documented or thoroughly studied (Peek 1989).

A third example is the use of population indices to monitor changes in population levels (e.g., ring-necked pheasant crowing counts, lek counts, track counts, catch-per-unit-effort, and aerial surveys). The primary assumption for use of a population index is the index is directly related to density. Although nearly every wildlife management agency uses trend data from population indices for management decisions, only a few examples of index validation exist (e.g., Rotella and Ratti 1986, Crête and Messier 1987, Marchandeau et al. 2006, Forsyth et al. 2007). Some studies have disclosed that index values are not related to density (Smith et al. 1984, Rotella and Ratti 1986, Nottingham et al. 1989, Rice 2003).

Walters (1986) proposed a systematic solution to these problems, which he called **adaptive management.** It involves a more formal specification of management goals and responses to management actions through the use of **predictive models** (Table 1.2) based on multiple working hypotheses, which can be compared to actual system responses through detailed **monitoring** (Thompson et al. 1998, Sauer and Knutson 2008, Conroy and Peterson 2009). Management actions are treated as experiments, which must be monitored carefully to ascertain whether goals were met and to identify errors in understanding the dynamics of the natural systems being managed. **Actual responses** to management actions are compared to predictions from our models based on current knowledge and assumptions (e.g., adaptive harvest management; Williams and Johnson 1995, Williams et al. 1996, Johnson and Williams 1999, Johnson et al. 2002). **Adaptive resource management** is an interactive process in which learning over time improves management as long as a monitoring program provides feedback to both our understanding of the system and the effects of management (Conroy and Peterson 2009).

Adaptive resources management is a specific case of **structured decision-making,** a process that addresses complexity, uncertainty, multiple objectives, and different perspectives to achieve management objectives (Clemen 1996, Clemen and Reilly 2001). Structured decision-making has multiple steps: problem definition, objectives, alternatives, consequences, trade-offs, uncertainty, risk tolerance, and linked decisions (Conroy et al. 2008). The basic strength of this decision-making approach is that it allows wildlife scientists to make effective decisions more consistently and to provide guidance for working on hard decisions (Clemen 1996, Clemen and Reilly 2001). Wildlife scientists are faced with difficult decisions regarding both the management and conservation of wildlife. For example, how can bison be restored to their former range, which would benefit other threatened prairie species, while also considering the economic and social impacts to cattle ranchers if brucellosis spread from bison to cattle? Both structured decision-making and adaptive resource management are being used increasingly often by wildlife scientists (Conroy et al. 2002, Johnson et al. 2002, Dorazio and Johnson 2003, Regan et al. 2005, Moore and Conroy 2006, McCarthy and Possingham 2007, Martin et al. 2009). Both these approaches differ from **scenario planning** (Kahn 1965, Chermack et al. 2001), practiced in business and other organizations to make flexible long-term plans based on considering multiple assumptions about the future. Such future assumptions are developed from a combination of established facts and multiple plausible forecasts of future changes, especially social changes. Scenario planning by the U.S. National Park and Fish and Wildlife Services in crisis situations, such as British Petroleum's Deepwater Horizon spill of 4.9 million barrels of oil into the Gulf of Mexico in 2010, should provide a foundation for a more measured adaptive management process to restore the damaged wetlands and marine ecosystems.

If wildlife agencies have the responsibility for management of wildlife populations and their habitats, they also have the responsibility to conduct research on the **effectiveness** of management programs. Wildlife agency administrators should strive to develop well-designed, long-term management-research programs as a basic component of annual agency operations.

SUMMARY

Carefully designed wildlife research improves the reliability of knowledge that is the basis of wildlife management. Research biologists must rigorously apply the scientific method and make use of powerful techniques in survey sampling, experimental design, and information theory. Modeling is an effective tool for predicting the consequences of management choices, especially when it is based on carefully designed field studies, long-term monitoring, and management experiments designed to increase understanding. More effort should be dedicated to the design phase of research, including obtaining critiques from other biologists and statisticians and avoiding common problems, such as insufficient sample sizes, procedural inconsistencies, nonuniform treatments, and pseudoreplication. When possible, we must move from observational to experimental studies that provide a more reliable basis for interpretation and conclusions; these studies need to be replicated across space and time. Wildlife biologists have a tremendous responsibility associated with management of animal species experiencing increasing environmental degradation, loss of habitat, and declining populations. We must face these problems armed with knowledge from quality scientific investigations.