Chapter 3

Essentials of Standardization and Quantification

Introduction

Throughout this book, we emphasize the need for standardization of techniques across and within studies. Our rationale reflects the rationale for the book in general; if a primary goal of field inventories and monitoring studies is to provide comparative data for analyses of biodiversity and examination of population trends, local extinctions, and the impact of human activities on amphibian populations, then studies must use standard techniques. The importance of standardization cannot be overemphasized, because studies using different techniques are often simply not comparable, even at the simplest levels. Thus, if one field researcher uses visual encounter techniques to derive a species richness list for an anuran breeding area, and another researcher later uses aural transects, it is impossible to determine whether any differences in the species listed reflect real changes in species composition, different sampling biases of the two techniques, or both. If both studies used the same technique, either visual surveys or aural transects, any changes in the list over time could be attributed to a real change in species composition. Even so, certain species certainly would be missed (semifossorial leaf-litter species by visual surveys and voiceless ones by the aural transects). Obviously, the best approach in this case would be for both techniques to be used in both studies; we always advocate using a combination of techniques to survey a habitat as completely as possible.

In preparation for this volume, we attempted to review all known techniques for the inventory and monitoring of amphibians. Here, we endorse and describe 10 techniques, and we encourage their use. These techniques may not be equally effective under all conditions, but it is our collective opinion that they represent the field sampling strategies that can best be standardized over the widest range of conditions.

A number of techniques are described and recommended for inventory and monitoring, because amphibians occupy a variety of habitats. The biphasic life cycle of most amphibians also means that different techniques are needed to sample larvae and adults. Several methods may be needed to sample an assemblage of amphibians, but methods must be consistent among study areas and across years.

Study questions

The approach used depends on the questions being asked. Thus, the purpose of a study should be clearly stated before the study begins. If the purpose is to compile a list of species for a poorly studied area, then an inventory is appropriate. If comparisons of species abundance among areas or across years are desired, then more-detailed monitoring methods must be used.

An *inventory* is a study of a specific area (for example, a national park or a defined geographic region) to determine the number of species of amphibians present (species richness). Inventories produce presence-absence data for species. Inventories are most often conducted (1) in areas in which little work has been done previously, and for which an enumeration of species richness will provide a baseline for biodiversity analyses, (2) across areas or habitats in which geographic or ecological distributions of single species need to be established or verified, and (3) in regions in which point comparisons over time can document changes in species distribution (presence or absence) and habitat use.

Monitoring is used to determine species composition and abundance (numbers of individuals per species) at one or more sites through time. Because all taxa in all habitats cannot be monitored with equal success, investigators most often target a specific type of habitat or an individual species or group of species for study.

Sampling considerations

Scale

Just as it is crucial to focus on the questions being asked, it also is important to define the sampling design and methods of analysis well before the field program begins. This maximizes the utility of the information gathered in the field, the comparability of that information with information from other studies, and the extent to which the data can be used to answer the questions being posed. Design considerations (discussed in Chapter 4) are as important for a one-day survey to produce a list of species for one locality as they are for a multiyear study of species abundances across a range of habitats.

Two extremely important points that should always be addressed are (1) the goal of the study (i.e., why the study is being done) and (2) the geographic scale over which the results will ultimately apply. These factors define the spatial scale of the sampling program that will follow. For example, if one were interested in a complete enumeration of the breeding population of the toad *Melanophryniscus moreirae* in one pond in Itatiaia National Park in Brazil, a sampling strategy that counts every individual from that pond should be used, not a randomized sampling design. However, if the goal were to compile species abundance information for all amphibians in Itatiaia Park, one might divide the park into ecologically relevant habitat types and

randomly sample within them. Finally, if the goal were to compare species abundances throughout the Serra da Mantiqueira of Brazil, one might divide the mountain range into elevational zones and randomly sample one-hectare plots within each zone. All of these approaches, if properly used, would provide quantitative results comparable with other identically designed studies. Which approach is most appropriate, however, depends on the goal of the comparison. Thus, in the first case, the data could be used to track the absolute abundance of *M. moreirae* in that pond over time; they could not be used to track abundance among ponds (because there is no sampling among ponds). In the second example, inferences could be made about the amphibian fauna of Itatiaia over time or space, although information on any given species might be relatively poor. Similarly, in the final example, a point sample for comparisons of changes in amphibian diversity over one mountain range would have been established, although information on the amphibians of Itatiaia or on the species *M. moreirae* might not be available, if the randomized samples missed Itatiaia and the small area where *M. moreirae* occurs within the park.

Randomization and Bias

Studies of biodiversity can yield insights at three levels. At one level, species' presences are documented to produce a list for the study area. At another level, abundances and distributions of individual species in time and space are determined. At a third level, general patterns of diversity derived from the data, and processes that account for these patterns are deduced. Comparisons can be made at all three levels.

Whether two investigators estimate biodiversity of different areas or of the same area at different times, interpretations of the results hinge on a few fundamental elements of sampling procedure. The question or hypothesis to be pursued determines the general boundaries for sampling, but the sampling protocol within those boundaries must be selected. Because environments are never truly homogeneous, different sampling protocols can give substantially different results. In particular, if sampling points are not distributed randomly through the area of interest, then analyses of the resulting samples are likely to underestimate or overestimate biodiversity. Environmental heterogeneity can effectively be removed (1) by recognizing the variation, subsampling within different types of habitat, and then comparing resulting estimates among the habitat types (this procedure is called blocking or stratified sampling in statistics) or (2) by ignoring the heterogeneity and sampling randomly without regard to habitat type. In either case, distributing samples randomly in the study area is an excellent way to minimize the problem of sampling bias, both for internal comparisons within a study and for comparisons across studies. Chapter 4 provides an in-depth discussion of random sampling.

Replication and Assumptions

Replication is another major component of study design. It is important for two reasons. First, replication provides a basis for confidence in the estimates obtained, because the bounds within which a population estimate falls can then be determined with appropriate statistical techniques. Second, replication minimizes the effects of localized factors that can obscure the study-site-wide variables of interest.

Sampling programs should be designed such that resulting data can be subjected to objective statistical testing. Thus, it is important to understand the assumptions of the sampling program and to attempt to satisfy the assumptions imposed by particular statistical tests. Two basic assumptions for most statistical procedures in this book are (1) that sampling is randomized and (2) that observations are independent. These assumptions present real problems for some of

the techniques described in this book, but reliable statistical inferences can be expected only when the assumptions are reasonably approximated. If sample sizes are reasonably large and samples are obtained randomly, then statistical tests can be used to determine whether observed differences between sampled areas are due to chance alone (the null case) or reflect significant differences attributable to biological factors (see Chapter 4).

Reporting data

The primary data that emerge when most of the techniques discussed in this book are used are species richness and abundance information. Frequently these basic results are summarized and reported as "simpler" summary statistics, such as a diversity index for a given habitat or species assemblage; such indices can make cross-study comparisons difficult or impossible. The proper use and interpretation of diversity indices are complex and controversial topics that are discussed in Chapter 9. However, all diversity indices must be derived from accurate, statistically sound species abundance information, and presentation of those data is the most important contribution a field survey can make to biodiversity issues. We strongly encourage authors to make such data available for cross-study comparisons by including them in journal articles or depositing them in an accessible repository.

Chapter 4

Research Design for Quantitative Amphibian Studies

Lee-Ann C. Hayek

Introduction

Research on biological diversity includes description of phenomena observed in the field, as well as controlled and objective investigation of these phenomena and the relationships among them. Basic to such research is the generation of hypotheses and the formulation of plans for testing them. In this chapter I discuss procedures necessary to develop such plans.

It is impossible to carry out research or evaluate research literature in today's environment without understanding probabilistic and statistical aspects of research. Toward this end, I consider some general conditions basic to the proper application of statistical techniques. I do not list specific assumptions or tests appropriate for particular sampling techniques. Rather, I acquaint the reader with the circumstances under which certain probabilistic approaches can best be used. I also provide an overview of the connection between the biological reality being tested or examined and the conditions under which results meaningful for its evaluation can be obtained.

Project design

The Research Question

A scientific research question is one that asks about the relationship between two or more factors or variables (e.g., species, environmental

conditions) within a defined context. The question should be precise and limited and asked in definitive, quantitative terms. It must contain the basis of and clear implications for testing. An example would be, "Does the number of amphibian species increase with an increase in rainfall in rain forest habitats in western Amazonia?" Generalized or metaphysical questions, such as, "Does good weather affect catchability of frogs?" will not suffice.

Generally, research questions are formulated in terms of concepts and constructs. A *concept* is an abstraction or idea of a universal term that is developed by generalization from many individual cases. For example, the concept of "frog" is not derived from the characteristics of one specimen of *Rana ingeri* but is a generalized idea of the collection of characteristics that make up the nature of all frogs. A *construct* is an abstraction of an unobservable, postulated phenomenon (e.g., aggressiveness, dominance behavior, territoriality, conformity), the idea for whose existence is synthesized from specialized cases of behavioral observations. It is an impression of the behavior of the object of study.

FOCUS OF THE RESEARCH QUESTION

The question, once defined, should be applied to a specific group. In this context, the *population* is the set of all possible observations of the same kind that can be obtained. The *sample*, in contrast, is all the observations actually made, or all the data obtained. One population, therefore, can give rise to many different samples. The core of statistics is to determine the extent to which generalizations pertaining to the population can be developed from the sample obtained during fieldwork. Often, observational or sampling studies can be improved by more clearly specifying the population and ensuring that the samples are representative of that population.

It is necessary to distinguish between two types of populations in a statistical sense. The *target population* includes those individuals to which the researcher would like to generalize the results of her/his study. The *available population* is the actual group of individual amphibians that the investigator can reach for participation in the study; it is the group of potential subjects from which the sample is drawn. For example, the investigator may wish to generalize about all the species of *Ambystoma* in the United States (in terms of the concepts of richness or diversity); these species then would constitute the target population. The study plan, however, may call for sampling these salamanders only along a series of transects in the southeastern United States or in Virginia. The *Ambystoma* species occupying the area sampled constitute the available population. When the target and available populations are not the same, potential inferences are weakened. Whenever possible, these two populations should coincide.

DEFINITION

A *definition* is a set of terms or characteristics used to delimit the essential qualities or nature of a particular variable, procedure, or phenomenon. Definition plays a vital role in scientific inquiry—especially in formulating research questions—by setting out the specific conditions under which an observation constitutes a particular type of information and by identifying the actual and inferential targets of the research. The terms used in a scientific research question must be clear and precise. Imprecision is one reason that contradiction and confusion exist in studies of species associations and diversity. The following rules should be used in formulating definitions:

1. A definition should give the essence or nature of the thing defined and not its accidental properties. By means of a definition, one attempts to show how the object or individual belongs to a group or general classification. A definition also delimits how the

object or individual differs from all others in that group, which is, of course, the commonly accepted basis for biological systematic classification.

2. A definition should identify the specific category into which the thing falls and the differentia, as well. A frog is not defined as an animal that jumps, because many other animals also jump. An amphibian that jumps narrows the field but is not specific enough for study purposes. The importance of this rule cannot be overstated. Without a way to eliminate extraneous factors from a study design, one cannot construct valid inferences.
3. A definition should be concise but inclusive.
4. A definition should not be based on synonyms. The vagueness that characterizes spoken language has no place in science. Rather, the individualized concept or construct must be coupled with an objectively measurable phenomenon.
5. A definition should not be based on metaphor. A metaphorical definition—for example, "Calling activity at ponds sometimes is tumultuous, and the full chorus sounds like distant thunder"—has low information content and provides no objectively measurable character of the item being defined.
6. A definition should not be based on negative or correlative terms. Although such statements are acceptable and even authoritative in technical fields, they are frequently ambiguous and imprecise in nontechnical studies. For example, in mathematics, $f^{-1}(x)$ is unquestionably the inverse of the function $f(x)$. However, in ecological studies, cold weather cannot be defined as the opposite of hot or warm weather.

OPERATIONAL DEFINITION

The definition of the variable, object, state, or individual in a scientific study must be precise and reproducible; it is a rule of correspondence between a set of constructs and the observable data (Torgerson 1958). A definition thus constructed is termed an *operational definition* (Bridgeman 1927; Carnap 1936; Margenau 1950; Kerlinger 1973). The alternative to an operational definition is a *constitutive definition* in which the constructs or concepts are defined with other constructs (Margenau 1950; Kerlinger 1973)—for example, "The weight of this frog is its heaviness." Constitutive definitions are neither appropriate nor sufficiently informative for amphibian sampling.

An operational definition clarifies how the object of study functions as a result of its specialized nature, focusing on its observable characteristics (e.g., Bridgeman 1927; Kerlinger 1973). Basically, this type of definition assigns meaning to a construct or variable by specifying the activities or "operations" necessary to measure it. It connects the scientific concepts to experimental and quantitative procedures with terms that have empirical meaning. In fact, this type of definition is a manual of precise instructions to the investigator.

For example, let us say that we isolate a breeding pond with an enclosure. Operationally, we could define a breeding male as a male found within the enclosure during the relevant season. This statement could be correct, but the criterion of definition is not unique; nonbreeding males might also be found within the enclosure. To define breeding males, we would have to include as many singularly observable characteristics as possible (e.g., calling from the water, in amplexus) in order to eliminate all other males from consideration.

For amphibian research, it is useful to distinguish three types of operational definition: procedural, behavioral, and structural.

A *procedural definition* is an operational definition that sets forth the manipulations or procedures required to induce or to observe the phenomenon or state to be studied. For example, for an observational experiment to be performed on "an abundant species," *abundant* could be

defined by a certain number of specimens observed per kilometer of transect. Within the context of the study, anyone can agree whether a selected species is abundant or not by noting the numbers found per kilometer; opinions open to interpretation are avoided. This type of definition is especially useful for describing independent variables subject to causal or correlative relationships.

A *behavioral definition* is an operational definition that focuses on dynamic aspects of the behavior of the object, state, or individual, linking observable antecedent behavior with an associated change, outcome, or dependent variable. The behavioral type of definition is most useful for defining the dependent variable in investigations involving behavior and is little used in studies of biodiversity.

A *structural definition* is an operational definition in which the demonstration of a specific behavior or other character constitutes the definition. This type of definition focuses on the characteristics of the individual or object, specifying static rather than dynamic qualities. It is useful for defining any variable, independent or dependent. For example, a breeding male frog could be defined as "a male frog found calling at the breeding pond."

Operational definitions set precise, concrete preconditions (e.g., operations, procedures, events, behaviors) that are observable and that lead invariably to the phenomenon under study. Such definitions satisfy the requisite of quantitative study, which is the specification and measurement of variables and their relationships. When the sex of an amphibian or the species is the variable, measurement is usually straightforward. When assemblage structure, calling activity, diversity, association, or dominance are being investigated, measurement is not so simple, and peak precision is mandatory for successful study. Although an entire theory or sampling plan cannot be laid out in operational terms, such terms must be used to define the quantitative, measurable, and testable aspects of the study.

In an inventory or monitoring study, the investigator is concerned with nonmanipulated, or naturally occurring, variables rather than with the manipulated and carefully controlled variables that characterize a laboratory study. Classical statistical experimental designs must be modified to deal with this reduced control, with minimal loss to the quality of inference. For such observational or field studies, the aim is to describe the procedures used to identify the state of a variable already characterizing each individual, locality, or habitat in the target population. Operationalism in this context simply indicates how the variable states are to be identified and the situations in which they are to be observed and recorded.

Formulating the Research Hypothesis

Development of a research project requires the statement of the research question or problem and then its restatement in terms of a testable or working *hypothesis*. The research question may first be formulated in theoretical terms that link concepts, but then it must be translated into operational terms, which requires the researcher to consider measurement in precise terms. Unless the hypothesis is stated in a testable form, one cannot distinguish positive from negative evidence. For example, an investigator may believe that "*Plethodon cinereus* prefers deep litter in eastern United States deciduous forest" or that "*Phrynohyas resinifictrix* prefers tree holes in lowland rain forest," but these ideas are not testable as stated. Is "preference" to be defined in terms of correlation, association, or dependence? And what does it mean in operational terms, only that *P. cinereus* will be found in deep leaf litter, or, in addition, that it will not be found in other substrates? Definitions that are too vague may preclude generalization because conditions cannot be replicated. In contrast, extreme

precision (e.g., patch sampling using logs only 25 cm by 1.5 m or less) may be too restrictive, limiting a study to just one log or one tree, and not allowing for any generalization.

A hypothesis should be the proposed answer to the research problem or one facet of the problem. The hypothesis must contain a statement of the relationship between two or more quantities, species, or measurable variables and must convey clear implications for testing that relationship. Legitimate hypotheses exist that do not provide criteria of relationship between variables (see below). These types of hypotheses are not appropriate for inventory and monitoring studies but may be useful for study of ancillary variables collected with the basic sampling data.

The actual purpose of the test of a hypothesis is to ascertain the probability that the hypothesis is supported by fact. Strictly speaking, empirical evidence can never be said to prove or disprove a hypothesis, but it can support or "confirm" the hypothesis under suitable conditions of repetition (e.g., Braithwaite 1955).

The object of a hypothesis test is the relationship among the variables. These relationships are identified by *propositions*, statements that can be either true or false. There are several types of propositions:

1. *Simple qualitative proposition.* **A has characteristic B.** Such a statement does not establish a relationship between the (two or more) factors or variables, and, therefore, is not testable. Nevertheless, it may contribute to the development of theory by (a) serving to specify antecedent conditions under which certain affinities among the variables may be expected; (b) suggesting other related propositions and providing a basal amount of knowledge; or (c) becoming a component of more-developed and testable propositions. An example of this type of proposition is "Bromeliads are the unique habitat of *Osteopilus brunneus* in Jamaica."
2. *Consequent proposition.* **If A, then B.** This type of statement establishes that B is always a consequence of A. Such statements can indicate a causal relationship. It is quite often possible to convert a simple qualitative proposition into a consequent proposition. For example, in order to provide testability, the statement from proposition 1 above becomes, "If *O. brunneus* is found, then the habitat is a bromeliad."
3. *Positive correlational proposition.* **The more A, the more B.** For example, "The greater the structural diversity of aquatic vegetation in ponds, the greater the species diversity of tadpoles." The obverse of this type of proposition may also be used: **The less A, the less B.**
4. *Negative correlational proposition.* **The more A, the less B.** For example, "As elevation increases from 1600 m to 2500 m in tropical Peru, the number of species of pond breeding frogs decreases."
5. *Null proposition.* **A and B are unrelated.** A null proposition indicates that no relationship between variables will be detected. For example, "There is no associative relationship or interaction (other than geographic association) between *Hyalinobatrachium valerioi* and *Smilisca sordida* along streams in lower Central America." The disconfirmation of the statement does not include a preferred direction, so this is a *nondirectional hypothesis*. In contrast, proposition types 1 through 4 are *directional hypotheses*. Directional hypotheses relay more information about the testable relationship and may form the basis for more powerful statistical tests.

Validity

TYPES AND DEFINITIONS

The most important consideration in problem formation, operational definition of variables with their attendant relationships, and formula-

tion of testable hypotheses is the maintenance of balance between specificity and generality. In this regard, the investigator must be keenly aware of the validity of the research design.

Validity is conformation to declared purpose and is used with reference to propositions, including causal propositions (e.g., see Cook and Campbell 1979). For observational studies, it is a measure of the degree to which the research design will actually produce the results or measure the variables that it says it will. High validity implies close approximation to intended purpose, and low validity suggests poor approximation or larger "error." In ecological work, *error* is used to describe all departures from representativeness regardless of their cause. Alternatively, in statistical work, *error* is used only to refer to nonrepresentativeness caused by inappropriate sampling methods; all other problems are called *biases*. The source of any compromise to the validity of a study is called a *threat*.

In designing sampling studies for amphibians, the investigator must be concerned with two major categories of validity. *Internal validity* is the extent to which differences detected in a dependent variable can be ascribed directly to changes in an independent variable in a specific sampling instance. *External validity* is the extent to which results obtained and statements inferred from a specific sampling situation apply or can be generalized to individuals, populations, objects, or settings not directly participating in the study.

THREATS TO INTERNAL VALIDITY

To achieve internal validity in a research investigation, one must be able to rule out all extraneous causative variables as explanations for the observed result. When a rival variable is eliminated, it is said to be *controlled;* uncontrolled variables are threats to internal validity.

A controlled variable is not associated with or related to the independent variable, so its effects cannot be confused or confounded with those of the independent variable. Internal validity must be ascertained for each study by asking if a given (independent) variable really has a primary associative relationship with another (dependent) variable and, if so, if the independent variable actually produced a change in the dependent variable. Before validity can be evaluated, the investigator must be sure that no extraneous variables have affected the result or been mistaken for the variable of prime interest.

When a team has only a few days to obtain a species count and list of the target fauna at a remote site, the study usually is poorly controlled and has poor internal validity. Thus, when results of such a sampling study are used to evaluate faunal change or to design a management program, conclusions may be misleading or inappropriate. Sampling in this manner often involves "misplaced precision," in which excessive care is taken in the collection of data about which conclusions can be, at best, impressionistic and imprecise and, at worst, indefensible. Because this type of one-shot study design is frequently used (particularly to evaluate the potential impact of environmental modification), we have included it as a technique (see "Complete Species Inventories" in Chapter 6), but with as much standardization as possible. Data obtained in this way may be used at least as a comparative base for more-intensive studies or as a preliminary estimate of the composition of a species assemblage.

There are several potential threats (selectively adapted from Campbell and Stanley 1963) to internal validity for amphibian studies:

1. *Historical threat.* A historical threat is an extraneous or unexpected event occurring in the environment at the time the observations are made (especially between the first and subsequent observations) that may confound the data on the selected relationships between variables. *Confound* is used here in a statistical sense to indicate that an obtained

effect can be attributed to two or more variables, and the unique portion due to each cannot be disentangled. For example, consider a study design in which night driving along selected roads is used to sample amphibian presence and activity. The investigator carefully identifies the two samples to be compared and controls for time, date, weather conditions, seasonal factors, speed, and vehicle. On the first night several species are encountered, but the second night yields many fewer. Further checking shows that on the second night a herpetology class collected amphibians along the road 30 minutes before the investigator appeared. Comparison of sample results is impossible.

2. *Maturation threat.* Maturation threats are changes (e.g., in age, breeding status, fatigue level, seasonal activity) in the individuals in the population during the period of observation that may affect the final outcome of the study. Maturation can be a consequential threat when organisms are sampled at time intervals of considerable length relative to the length of their life cycles.
3. *Instrumentation and observer threats.* These threats result from differences in measurement calibration that may lead to differences in results. The problems encompass not only laboratory and field instruments (e.g., weak batteries can affect the quality and accuracy of frog recordings) but also differences among observers or observations, which may be even more of a threat. For example, a novice and an experienced person likely would record different values for the distance a frog call can be detected or for the number of frogs calling along a transect in a tropical forest. Likewise, a person who hears a species-specific call in two distinct microhabitats, or before and after observation of another variable of interest, may judge the calls differently because of increased experience and discrimination, increased fatigue, or plain carelessness. Such variation can be minimized with standardization of methods and checks of inter-observer uniformity prior to fieldwork.
4. *Statistical regression threat.* These threats result when amphibians or other organisms are selected for study on the basis of the extreme of any character. For example, it is inappropriate to draw a conclusion about abundance based only on visible frogs if their typical inclination is to remain hidden in leaf litter. If frogs are selected for strength of call (i.e., the louder the call, the higher the probability of selection) or other character, the effect of statistical regression can be mistaken for the effect of the variable under study. The more deviant or extreme (in either direction) the first measurement, the more the second set of measurements will vary. This phenomenon is especially pronounced with variables subject to unreliable measurement. If two large samples are drawn randomly from two different populations, or from the same population at different times, without any matching, then regression threats are not a consideration.
5. *Interaction with selection effects.* Selection effects occur when samples are selected nonrandomly and the resulting groups of amphibians differ in size, variability, or type. Such differences jeopardize group comparisons and may be a threat to internal validity. The extent to which randomization assures group equality is shown by tests on the sampling statistics (e.g., means and variances). The chance that the assumption of group equality is not tenable increases for small samples.

Interaction may occur between sample selection and any of threats 1 through 4, above. *Selection-maturation* interaction is possible when two groups (e.g., two species of anuran larvae in a pond) mature at different rates or in different

proportions (e.g., more males than females). *Selection-historical threat* interaction can result when groups have distinct local histories. Such effects are especially likely when a known source of variability is ignored. For example, if males and females of a particular species have different behaviors or habitat distributions, then random sampling may be inappropriate. Blocking or stratifying the sampling on the basis of sex, prior to random selection within each sex, will increase sampling precision and decrease the interactive threat to internal validity.

THREATS TO EXTERNAL VALIDITY

Because investigators are almost always interested in the relevance of their work beyond the confines of the selected sample (i.e., generalizability), concern about external validity is important. External validity can be strengthened by describing the population, settings, and variables to which the results will apply, before the study is initiated. The representativeness of these factors will determine how extensively the results can be applied.

The following factors may threaten external validity if they are not evaluated prior to data gathering:

1. *Interactive effect of selection.* The characteristics of the sample selected from the available population determine how extensively the findings can be generalized. External factors (e.g., weather, environmental conditions) as well as internal characteristics (e.g., breeding status, age) of the particular group selected may contribute to an atypical finding. The diversity found in artificial pools, for example, may not be representative of diversity in natural ponds.
2. *Reactive or interactive effect of preselection methods.* A treatment or method used in a study influences subsequent results through its effect on the behavior of the subject. For example, an animal captured and marked prior to a sampling effort may be more likely to avoid capture than a naive animal. Likewise, the presence of an observer in the immediate area may affect normal amphibian behavior in unregulated ways. This may not, in itself, limit generalizability, if standard preconditions are met, because all inventories and monitoring efforts to some extent are observational studies.
3. *Multiple effects interference.* Measurements taken from individuals subjected to more than one procedure may be representative only of measurements taken from other individuals that have experienced the same series of procedures.

SUMMARY

For any inventory or monitoring study, the internal and external validity of each stage of the study design must be evaluated in detail, and the presence and extent of any threat determined. The frequency and importance of each threat will vary within and among studies, and particular threats are not inevitably correlated with particular sampling designs. I have listed here only those threats that I consider to be plausible in field biodiversity studies. The list is not exhaustive (see e.g., Wright 1991), but it may help the investigator to recognize threats and to minimize or eliminate them from the study plan.

Field observation and statistical design

Data Accuracy

A fundamental assumption underlying any amphibian study is that the data have been accurately recorded and processed. As emphasized in Chapter 6, observations should be recorded in the field and later coded (if necessary) and entered into a computer (if available or desirable) for analysis, to reduce the likelihood of introduc-

ing errors. Confidence in the analytical results cannot exceed confidence in the accuracy of each observation.

If computers are used, the data should be checked as they are entered. Spreadsheet and database programs allow limits to be placed on each column of input in order to prevent errors of excess. For example, if season is coded as an integer between 1 and 4, the program can be set to prevent other digits from appearing in that column. Even with such precautions, it is wise to check the data a second time against a hard (paper) copy of the completed file.

Despite careful checking, input errors may remain undetected. One of the most common problems is that of extreme values, or *outliers*—that is, patently unrealistic observations. Both the computer data and the original field notes should be checked. Only when a source of error is clearly identified (e.g., a code of 4 for sex or a species count of 100 incorrectly copied as 1,000) should the extreme observation be corrected.

Other situations that may affect data accuracy are accidents (e.g., rain gauge tipped over) and mechanical or personnel problems (e.g., weak batteries on data recorder, investigator sick). Any decision to ignore observations because of possible data contamination must be made before the data are examined, so that the actual value of the observation cannot influence the decision. All contaminated data must be either discarded or retained. Quite frequently in field studies, values representing genuine biological effects cannot be reliably distinguished from accidental irregularities or input errors. If the investigator cannot be certain that an extreme value is the result of a problem, the value should be retained.

Even if there is an obvious reason to discard an observation, internal or external validity may be threatened. It is possible that extreme values or measurement errors are more likely to occur under one set of study conditions than another. If so, and if these values are discarded, then the remaining data may not be representative of the study population. For example, the set of quadrat observations from certain microhabitats in a study area may be faulty because of high observer error or patchiness of species occurrence. If these observations are edited or eliminated, the remaining quadrats may not be representative of the target area's true heterogeneity, and a statistical estimate then could be severely biased. Care must be taken to examine extreme observations within the frame of reference of the study before any value is discarded. This possible source of bias should be clearly mentioned in the final report or publication. It may be of value to compare the results of analyses made with and without the suspect values.

Measurement Scales and Statistical Analysis

Numbers that result from an inventory or monitoring effort may have one or more of three cardinal features: intrinsic meaningful ordering; ordered differences between number pairs; or a unique "zero" point, or natural origin, indicating absence or deficiency. Stevens (1946) named four numerical scales based upon the number of these features observed and the amount of information represented about the measured property:

1. *Nominal scale.* Numbers on this scale are used to name categories in a classification, but they do not measure any property and have no intrinsic order. In this case, numerals are actually unnecessary; word descriptions such as male and female, or letters such as *m* and *f*, can be used. When the classification is subjected to statistical analysis, numerals without an underlying order relationship may be used. For example, sex may be coded 1 or 0, and a male can be designated as either value with no sacrifice in meaning.

2. *Ordinal scale.* Ordinal numbers are those assigned to the amounts of a property, so that the order of the numbers corresponds to the order of magnitude of the amounts (Torgerson 1958). Ordinal scales may have a natural origin (Torgerson 1958) or not (Stevens 1946). On an ordinal scale, objects can be arranged in a meaningful serial order with respect to some property, showing that some individuals or objects have more of a particular attribute than do others. For analytical purposes, any order-preserving (i.e., monotonically increasing) transformation of the numbers will serve as well as the original set.
3. *Interval scale.* An interval scale denotes equal incremental amounts of a property of an individual with equally valued numerical increments. In addition to the order of the numbers corresponding to the order of magnitude of the various amounts of the property, the size of the difference between the pairs of numbers corresponds to the distance (in a generalized sense) between the corresponding pairs of amounts of the property. Any set of numbers satisfying the requirements of such a scale is not affected by a linear transformation (of the familiar form $y = ax + b$) of the set. An increase in one unit from any region of the scale is identical to a unit increase from any other region.
4. *Ratio scale.* A ratio scale is formed when the requirement for a unique natural origin is appended to the rules for forming the interval scale. Any set of numbers satisfying the requirements is insensitive to a linear transformation (of the form $y = ax$).

Baker et al. (1971) provided at least a partial answer to the question of how to relate the measurement scale for data to computational procedures and statistical tests. They showed that probabilities estimated from the sampling distributions of so-called strong statistics, such as the *t*-test, are almost unaffected by the type of measurement scale used. In general, therefore, an inferential statistical test can answer the research question it was designed to answer, regardless of the original scale of measurement.

Randomness

The basic statistical tests used with observational data from inventories or monitoring studies require that the initial selection of subjects or localities or the application of field techniques (e.g., quadrat placement) be random. The term *random* is used in a technical sense in research design; it does not describe the data in the observed sample but the process by which the data were obtained. Sampling is random if each possible sample or combination of n selections of individuals in the population has the identical chance of becoming the sample actually drawn. Random sampling does not mean, as is often stated, that each individual in the population has an equal chance of being in the sample. In practice, a sample is drawn individual by individual, and each has an equal chance of selection at the time of selection. This procedure is also called *simple random sampling* in the sense that it is sampling without further restriction (Kempthorne 1955; Cochran 1963). Implicit in this discussion of procedures for simple random sampling is selection from a finite population. If the population is of infinite size, random sampling procedures do not apply, and a sample is selected randomly only by assumption.

Many people believe that any sample drawn randomly from a population is highly representative of that population or equivalent to it in all essential characteristics. A second widely accepted notion is that chance events are self-correcting. Both of these assumptions are untrue and can affect the validity of a study.

REPRESENTATIVENESS OF SAMPLES

The law of large numbers predicts that very large samples will be representative of the populations

from which they are drawn, but it does not speak to the medium and small samples obtained during the course of most fieldwork. Five tosses of a coin can yield 4 heads and 1 tail, or 80% heads, whereas the proportion of heads in 5,000 tosses is likely to be close to the theoretical value of 50%. In a similar manner, 5 quadrats randomly placed in a forest with 95% canopy cover might fall in the open over 4 small ponds and 1 trail; locations of 100 or 1,000 quadrats would likely be more representative of the area's true habitat profile.

The comparability of random assignments to groups is probabilistic, not deterministic. Individual samples can vary greatly, but the distribution of samples will depend upon the nature of the population from which they were selected. For a given sample size, one will obtain more heterogeneous samples from populations in which the individuals are more variable among themselves. Averages and proportions will vary less in large samples than in small samples from the same population. The value of random selection does not emanate from fairness or lack of bias with which any selected samples portray the population, but lack of bias is an important consideration in large samples. Among R. A. Fisher's greatest contributions was furtherance of the idea that random processes can be used to achieve group equivalence prior to an experiment or study. No test of significance requires de facto that all confounded, uncontrolled extraneous variables be removed, that is, that randomization be effected (McGinnis 1958). However, the process of randomization assures that conditions of equivalence of samples will be met within reasonable limits; failure to randomize does not guarantee, however, that such conditions will be violated.

SAMPLING METHODS

All methods of sampling have some attached pattern (distribution) of variability. Knowledge of this pattern is a basic tool of statistical inference and can be obtained only through the laws of mathematical probability, which are applicable only to samples obtained under random processes.

Other sampling methods are available. Some rely in part on random procedures, and some exploit certain features of random processes. *Convenience sampling* is a generic term for a type of sampling used for convenience rather than for formal representativeness. Sampling of a certain fauna by an expert herpetologist might well approximate the true proportion of species in the area. The numbers obtained, however, are still subject to sampling variability, but of an unspecified amount, and the possibility of scientific replication is diminished. There are three common types of convenience sampling: accidental, grab, and haphazard. *Accidental sampling* occurs when the data on a species or genus are obtained peripherally as part of a larger or unrelated study and are not randomly selected. Such sampling does not guarantee that the achieved observations are representative of the population, nor is it clear how to specify the target population.

Cochran et al. (1954) discussed *grab sampling*, in which an investigator, in effect, just "grabs a handful." Whether the sample units be individuals in the available population or cards with random numbers, the ones in the handful almost always resemble one another more, on average, than those from a simple random sample; variability is underestimated. These authors showed that even if the grabs are randomly spread such that each one has an equal chance of entering the sample, they do not share the characteristics of a randomly obtained sample.

Haphazard sampling heightens the implication of chance and, unfortunately, can pass for random sampling in some ecological applications (e.g., quadrat sampling). However, unless investigators employ a device, such as a random number table or generator, to select localities, microhabitats, or sites for the placement of quadrats or transects, they may create a *halo effect*. A

halo effect occurs when an investigator selects the "best" or the "good" sites or individuals for sampling. For example, the frog just outside the actual study area boundary is selected because the frog displays the desired characteristic. Such a selection method certainly may allow for larger samples or species abundance limits. However, haphazard sampling threatens external validity and can influence the results of subsequent statistical testing.

In contrast to convenience sampling, *probability sampling* is a process wherein randomness is a requisite. Randomness can enter the sampling procedure at any stage. A few such sampling methods are of interest for inventory and monitoring studies. With *stratification* (*stratified random sampling*) a target population is divided into relatively homogeneous groups or strata prior to sampling, based on factors (e.g., sex, size, breeding condition, life history state) known to influence variability in that population. Subsequent selection within each stratum is random. Factors are selected on the advice of amphibian experts. Such expert judgment may be inappropriate as a basis for statistical sampling, but it is vital for controlling variables extraneous to the phenomenon being studied and, thereby, for increasing internal validity. Stratified random sampling is also a useful operational strategy for screening individual specimens, events, or microhabitats for possible exclusion from the study and for reducing study variability.

Randomization within stratified groups adds precision to a study by ensuring that the sample contains the same proportional distribution of amphibians, events, or microhabitats as in the target population. It increases the likelihood that the research will be representative of the population. Stratified random sampling is appropriate for populations that tend to be patchy (Seber 1986). In such a case, each sample of size n in each stratum has a known probability of being selected for the study. The actual probability of selection does not have to be known; only the relation of this probability to the proportion of such samples in the population is necessary.

In sampling, a major effort must be made to reduce any large or important threat and any random sampling error. After this reduction, the easiest way to increase sample accuracy is to increase sample size. Other things being equal (Yates 1981), the random sampling error is approximately inversely proportional to the square root of the number of units included in the sample. The accuracy attained will depend on the sample size as well as on the variability in the population of subjects that contributes to the sampling error.

Methods other than simple random sampling and stratification that do not introduce further bias but substantially reduce variability and, in turn, reduce the sample size required to attain the level of accuracy desired are of considerable benefit. One method is *systematic sampling* with a random start point in which a sample is obtained by a systematic, not random, method (e.g., sampling at equal intervals in space or time). For example, one might choose among localities on a list by selecting every fifth entry. Another method, *cluster sampling*, uses groups or clusters (e.g., ponds, tree holes), not individuals, as the basic (multistage) sampling unit; this approach may be preferred in ecological situations that require even area coverage (Scherba and Gallucci 1976; Seber 1986). Many authors (e.g., Seber 1986) have noted the need for more sample designs that use sampling approaches with higher *efficiency*, that is, that provide for reduced variability. Both Yates (1981) and Krebs (1989) provided readable discussions of this problem.

It is important to realize that random assignment of individuals or objects to groups only minimizes but does not eliminate all threats to validity. Random assignment does not guarantee a productive research design or a testable hypothesis. It provides a proper environment for

inference but does not guarantee the infallibility of the inferences; assumptions of comparability are merely assumptions and must be examined for plausibility. Randomization is the best way to avoid accidental bias, but if group sizes are small, then large intrinsic differences that bias the estimated effects can still occur between groups by chance (Gilbert 1989). Alternatively, the judicious selection of control variables (McGinnis 1958) and the use of the more sophisticated sampling plans discussed above can reduce required sample sizes and add to both the internal and the external validity of the research.

USE OF A RANDOM NUMBER TABLE

We include in this book a table of edited random numbers (Appendix 7). Use of such a table in designing sampling studies is probably the most widely accepted method of obtaining random samples and is recommended in many of the techniques, but it must be properly used.

For purposes of inventory and monitoring studies, the first step is to list the objects, habitats, quadrats, or transects to be subjected to the random assignment process. Individuals, although the primary objects of concern, are not randomized in a field study; the sites at which they will be studied or trapped are. Each item on the list is assigned a unique numeral. The next step is the selection of the tabled numbers for use in selecting the sample. Let us assume we want to place 10 one-meter-square quadrats randomly throughout a specified habitat that is 25 × 25 m, or 625 m^2, in area. A map of the area could be subdivided into 25 equal-sized (5 × 5 m) plots, with each assigned a unique consecutive numeral, from 1 to 25. We then would need to select, at random, the 10 plots in which to place one each of the 10 quadrats. To do so, we would use the table of random numbers (Appendix 7) to select 10 numbers, each representing a specific plot in the set of 25. Use of this table, which was devised for this book, is explained in detail in Appendix 7.

At times a second application of the random sample procedure may be required to ensure that the selected sites are observed in random order. We also could select a "control" group in the same manner. The purpose would be to create two groups that are equivalent in a probabilistic sense. The two sets of quadrats would be located in areas representative of the total 625 m^2 of area selected for study. The use of such a control group does not necessarily mean that the target population under study has some unifying characteristic or forms a biological population of interest.

In some circumstances it may be desirable to employ an alternative form of probability sampling. For example, if the target population were stratified by microhabitat before sampling, plots within each of the strata would be selected separately using the random number table.

Stratifying, or blocking, prior to study is preferable to adjusting initial random assignment to groups (Cook and Campbell 1979). Nevertheless, if new information on the fauna indicates that microhabitat differences may be important in the study, the location of each randomly placed quadrat can be checked for noncomparability. The investigator might find, for instance, that 7 quadrats include parts of streams, 2 include small ponds, and 1 is covered with a deep layer of dry litter. If such a bias were noted prior to actual field observation, rerandomization (the selection of 10 new values) would be a possibility. However, in this case, stratification or blocking should be given serious consideration. To be maximally effective, the research design should determine the point at which randomization should enter.

Another common procedure for making random assignments of sampling units is to write numbers representing the sampling units available (e.g., 1, 2, . . . 25) on slips of paper, put them into a container, shake it thoroughly, and select slips until the sample size is reached (e.g., 10). This method can involve unforeseen bias (e.g., the slips of paper stick to the container or each

other). Computer-generated tables of random numbers can also be used, but each program must be checked because many authors have identified inaccuracies with particular generators.

Connor (1977) suggested the following procedures as aids to reliable randomization:

1. The individual or individuals who design the study and best understand the rationale for sampling make the random assignment.
2. The investigators control the assignment, not an agency's operating personnel.
3. One person makes the assignments, not a group.
4. Investigators do not use "loopholes," in which random assignment is circumvented for a small number of individuals or objects (e.g., the investigator includes a subject slightly outside the quadrat because of desirable characteristics).

Another point may be added to this list:

5. Randomization is carried out before entering the field. This practice eliminates unintentional as well as conscious bias in selection and prevents direct substitution of one unit, specimen, or plot for another that is less convenient or less apt to provide information.

A sample will be sufficiently representative of a population only if errors introduced by the sampling process in the field are adequately minimized. Even so, bias that affects wholly objective conclusions does not necessarily invalidate the total study. Many times, constant or small biases are inconsequential. For example, when limited bias is constant from species to species, inventories designed to look across species at one site are little affected. Investigators must avoid attaching exaggerated importance to minor sources of bias that, in fact, can only produce errors that are trivial relative to random sampling error (Yates 1981). Ascertaining the relative importance of bias and sampling error is a vital concern in statistical inference.

Independence

Statistical tests may lack validity if research events are not independent. Events are said to be *statistically independent* when the probability of occurrence of one event remains constant regardless of the occurrence of another. Successive samples from a population are independent if the probability of selecting any one sample is independent of the selection of the others (Marriott 1990). Samples of amphibians removed from a plot one by one (see "Quantitative Sampling of Amphibian Larvae" in Chapter 6 and "Removal Sampling," in Chapter 8) can be assumed to be independent if the population is large. Observations of amphibians along a transect are also independent as long as the individuals are not highly mobile and not likely to be recorded in more than one sample. In contrast, removal of males from a chorus probably affects the calling activity of other males and, thereby, makes their location and removal difficult. Likewise, samples from night driving at short time intervals may lack independence.

Transects at a study site must be placed far enough apart to make overlap of aural or visual encounters unlikely. When working with a small population, for example, larvae from a tree hole or a very small pond (e.g., one dip with a net), the initial sampling affects subsequent sampling by drastically reducing numbers. In this instance, the remaining larval population may take on a character different from those in the initial intact population. Another example in which dependence is possible involves sampling frogs from ponds within easy walking distance on successive days. Unless the frogs are marked, at least a few, and possibly many, individuals may be included in both samples.

Sample Size

In the design stage, it is common to raise questions about the sample size necessary for both testability and generalizability. Factors of time, money, and personnel act to keep the size small, whereas statistical and biological considerations call for larger samples. The prime concern is determining the minimal biological sample size needed to provide statistically credible findings.

Consideration of statistical tests, degree of sample comparability, and representativeness achieved by randomization persuades most investigators that bigger is always better. Explanations and apologies for small samples abound in the ecological literature. However, "the bigger the better" as a maxim is neither invariably true nor always a mandate for statistical analysis.

For inventory and monitoring projects we need to consider at least two aspects of sample size: (1) the numbers of quadrats, transects, or trips to the site and (2) the numbers of specimens collected and species sampled. Suggestions for the former are provided in the sections on sampling techniques, but with two caveats. First, for simple random sampling of quadrats, transects, or patches, suggested sample sizes are based on the number that experts have found necessary to achieve biological or ecological representativeness of the target area. Second, if stratification is involved, it is optimal to select judiciously one to three variables upon which to stratify and to achieve comparability of groups through random sampling within strata. This methodology will eliminate problems of missing or unattainable specimens in a large, multifaceted array.

Determining numbers of replicates in fieldwork is troublesome, especially because cost may dictate numbers of site visits. Although an investigator can specify a model, formulae to predetermine sample sizes usually are at best asymptotic rather than exact, and at worst impossible to achieve under study limitations. For our purposes, expert guidance should prove more reliable. In fact, revisiting a site according to a standard timetable is probably more vital to success than making a predetermined number of visits.

The number of specimens collected or species to be sampled presents a different problem. In field observation studies of the type we discuss in this book, predetermined numbers of specimens are not the norm; investigators find what they can. Likewise, in the realistic biological situation of inventories and monitoring, specimens and species are the topics of study but not the *sampling unit* and, unlike the transect or quadrat, they are not the direct object of the random selection. Nevertheless, it is possible to ascertain after sampling has been completed whether the sample size gives the null hypothesis a reasonable chance of being confirmed.

Investigators often are pleased to obtain a small sample and delighted with a moderate one. Field studies do not require the a priori use of formulae to determine sample sizes but rather the post hoc determination of the observed power of the fieldwork. Some authors (e.g., Rotenberry and Wiens 1985) have discussed the use of power and sample size formulae in the design of field studies. I focus on use of the concepts of power and sample size as they relate to interpretability.

Testing Errors

Both the *null hypothesis,* that the variable under investigation has no effect or the relationship has no meaning, and the *alternative hypothesis,* that the variable does have an effect or the relationship is meaningful, are under consideration when a statistical test is run. Statistical work in biology focuses on the null hypothesis. It also emphasizes the *type 1,* or *alpha, error.* The alpha error is usually called the *significance level* chosen for the study; it indicates how likely one is to reject a hypothesis when it is true and should not be rejected. This level, when allowed to range

over the interval (0, 1), actually may be seen as a random variable or as a statistic that measures the consistency of the data under the null hypothesis. The *type 2*, or *beta, error* of a particular test refers to how likely one is not to reject a hypothesis when it is false and should be rejected (a null hypothesis cannot be "accepted"); this type of error cannot be controlled simply by selecting a significance level. It is usual to preset the level of type 1 errors and to minimize the probability of type 2 errors.

Either of the two errors can be costly; circumstances of testing determine which has the more deleterious effect. It is simple to conjure up examples involving life or death in which the cost of committing a type 1 error is decidedly more than the cost of making a type 2 error. In this case, the significance level (alpha) could be set at 0.01 or even 0.001 for decision making to offset the seriousness of the possible error. When there is no life-threatening aspect to the research, a type 1 error is usually less costly than a type 2. This is especially true in exploratory studies or studies involving innovative features or elusive species effects. Toft and Shea (1983), however, have pointed out circumstances in basic research in which the cost of a type 1 error could exceed that of a type 2 error.

THE 0.05 CONVENTION

In order to make a rational choice for the levels of error, there should be some specification of the loss involved. Such specification is practically impossible in a general inventory or in a monitoring effort, which is a problem when testing hypotheses with such amphibian data. Generally, biologists have resolved the problem by adopting the conventional but arbitrary level of 0.05 (or occasionally 0.01) for alpha for all research, thereby ignoring effectively the second type of error, *and* the *power* of the test. The 0.05 level of alpha, which is listed as if it were indisputable truth (and even may determine publishability), is but convention. Use of a constant value (0.05) for alpha introduces an impartiality into the test procedure, but it can be a serious impediment to interpretation because it is a convenience that ignores other important aspects of the inference process. In addition, it addresses the question of error, not the utility or importance of the obtained result. A decision to use a fixed alpha error is not always the best strategy for observational work. The selection of paired values for alpha and beta errors, based upon the complexity of the sampling design, may well serve biodiversity purposes better. Alpha should equal beta to provide an equal chance of detection as a standard for observational fieldwork, unless circumstances make one error more costly or more difficult to detect than the other.

POWER, EFFECT SIZE, AND SAMPLE SIZE

The probability of making a type 2, or beta, error (failing to reject a false hypothesis) and the probability of correctly rejecting a hypothesis (power) are necessarily related (1 – beta error = power). Therefore, a *powerful test* is one that allows for a high probability of claiming there is a real difference when such a difference actually exists in the population. Summaries of research findings commonly report sample sizes as well as alpha error or observed probability (p) level but not type 2 error or power. In addition, the observed p value is often the only criterion used in making decisions about the correctness of a hypothesis, with no reference to sample size, research design, or the potential costs of the decision (Yoccoz 1991).

Interpretative remarks in amphibian literature reveal the attitude that significant results acquired with large samples are more compelling and meaningful than those based on small samples. In addition, when an investigator concludes that a statistically significant relationship exists, generally he or she is confident not only that a biological or ecological relationship exists in the target population, but also that the degree or size of that affinity or effect is worthy of further

consideration. However, the size of the effect (relationship or difference) is not a product of the size of the sample.

It is certainly true that the smaller the values of alpha and the observed probability, p, the surer one can be that the obtained result is not attributable to sampling error. However, neither alpha nor p indicates how far apart the parameters (e.g., means) are or how large an effect actually is being discussed. The *effect size* (Cohen 1977) is a relative measure, in population standard deviation units, of the difference the investigator desires to detect. It is not possible in many observational studies, however, to specify the effect size a priori. This difference may be estimated after the data have been obtained; the *observed effect size* (i.e., a standardized difference between the two observed parameter estimates) estimates this separation for a given procedure or test (Cohen 1977; Lipsey 1990). If the null hypothesis is rejected, then some real differences may exist between the situations specified by the two hypotheses. The magnitude of this difference will have a considerable influence on the likelihood of attaining significance. For equivalent-sized heterogeneous samples, the larger the effect (relationship), the more likely it will be determined statistically significant, and the greater the statistical power of the test. Likewise, any variable, regardless of how inconsequential, will manifest a statistically significant difference in large enough samples. But would one really accept that a statistically significant difference between means of 0.343 frogs/km transect and 0.341 frogs/km indicates a real change in species abundance? This type of issue must be resolved. With large samples the difference being tested may be tiny and still be called statistically significant; with small samples this difference would have to be quite large (but, as we shall discuss, still not necessarily substantive) to be detectable. For a reported difference to be evaluated, not only its significance, but also its size must be known.

When sample sizes are about the same, the maximum observed difference (in terms of an appropriate measure of effect size) between the groups of results termed nonsignificant and the difference between those that have been called statistically significant should be evaluated. The observed effect size for the first group should never be larger than any effect size in the second group. A moment's reflection should reveal this standard to be the basic minimum for avoiding problems of illogical summary interpretation.

In experimental settings and even some survey studies it is possible to set a minimum effect size or smallest detectable difference before the work begins. This value, determined a priori either by expert opinion (Cohen 1977) or from previous research results (e.g., Ferrari and Hayek 1990), is used in formulae to make power and sample size determinations.

An opinion not often presented is that tests of null hypotheses are not actually the most appropriate for fieldwork. However, in many instances interpretative problems and ambiguities could well be relieved or eliminated by an alternative approach. If the size of the effect of interest were specified in the statement of the scientific hypothesis, simple statistical tests would become only one aspect of the total statistical inferential picture. For example, a researcher could test whether the population sizes of *Ambystoma tigrinum* dropped by at least 10% over time, or whether the SVL (snout-vent length) of *Rana limnocharis* differed by less than one standard unit across two localities. In this way, the accumulated knowledge of the expert is drawn into play, and the exact situation to be tested has biological significance. The level of statistical significance clearly would refer to the confidence a person has in the final decision, and confidence interval estimation would play a more integral role. More important, the result would no longer be confused with the size of the effect itself, and an unrealistic effect would not be tested.

In correlational studies, workers commonly report a measure of effect size called the coefficient of determination (square of the correlation coefficient). In this setting, researchers apparently recognize that the correlation coefficient, its significance, and the probability level do not provide a measure of the size of the relationship under study. The value of a measure of the size of a relationship goes unappreciated in observational (and experimental) studies when tests of null hypotheses are the sole method used for statistical inference.

When the question of how large a sample is necessary arises, it is commonly stated that 25 (or 30 or 50 or 100) is the correct number to use to provide for statistically reasonable results. Usually the number is chosen to provide for relatively narrow bounds on the error about the mean; it is not related to any power considerations. Choosing a number on that basis can be a serious problem.

Because inventory and monitoring efforts are concerned primarily with uncovering important changes (particularly declines) in biodiversity over time, the most powerful tests possible should be used. Recommendations for additional study of areas, faunas, or species will depend upon the reported power levels. Nonsignificant findings (possibly contrary to informed perceptions) from a well-designed study with large sample sizes that minimize threats to internal and external validity do not necessarily require that the study be terminated. The best recommendation would be to revisit and resample with a higher-powered study before the investigative avenue is abandoned. Borenstein and Cohen (1988) provided a program for calculating power and determining requisite sample size for increasing power.

If the observed effect size is large and the sample size is sufficient to detect it, then confidence in nonsignificant results (based on alpha level) may well be justified. If the observed effect size is small and/or the sample size is insufficient to detect such a value, then investigation should be continued. If both power and observed effect size are calculated, the reader can make an informed decision about the population, and the investigator will know what limits to place on interpretations and recommendations.

Any statistical test procedure involves considerable subjectivity that usually is ignored by conventional methods of amphibian data analysis. Consideration and reporting of the values of observed power, observed effect size, and alpha level, as well as sample size, should lessen the tendency to accept the result of a statistical test of a hypothesis as a definitive research conclusion. The subjectivity inherent in statistical inferential procedures demands that the investigator consider whether the biological story that the statistics tell makes sense. Gilbert (1989) emphasized that the size of the biological effect must be worth bothering about or the story worth pursuing. A probability level indicating hypothesis rejection cannot provide that information. A statistical test answers the question asked; it is up to the investigator to be sure that this question bears a relationship to biological reality.

Statistical versus Substantive Significance

Strictly speaking, classical statistical inference provides valid answers in the context of long-term outcomes only, but individuals investigating amphibian biodiversity often need an answer as soon as the monitoring is complete. For example, when the findings, with 95% confidence, are that one habitat harbors significantly fewer species than another, statistical inference allows the researcher to say only that if the same habitats were randomly sampled over and over, in only 5 times out of 100, on the average, would differences as large or larger than those actually observed occur purely as a result of chance. Unfortunately, there is no way of knowing whether or not a particular set of results represents one of those cases.

Exclusive reliance on tests of significance without incorporation of other forms of inference (e.g., confidence interval estimation) obscures the relationship between the observations themselves and the magnitude of the effects to be examined. Null hypotheses of no difference are usually known to be false before the data are collected (see e.g., Savage 1957). No amphibian worker could actually believe in the possibility of a *sharp null hypothesis*—that is, that two means are absolutely equal. In field biology, systems are too noisy to allow for such absolute equality. Even though biologists know this intuitively, they still treat the test of such a null hypothesis as if it expressed a realistic and meaningful difference (zero), and many books present the null test as the only choice (see e.g., Siegel 1989).

A statistical test reflects only the size of the sample and the power of the test, not the biological question raised by the hypothesis. The existence of a specific effect must be demonstrated across settings or times to be biologically significant; it must be large enough to matter and therefore must be examined with a test powerful enough for detection. Mere graphical procedures often can show the proposed relationships to be less than meaningful. It is interesting that most people would accept that about 5 (the 0.05 level) of 100 coin-tossing experiments would show that the proportion of heads is significantly different from 0.50. Turn this situation into a test of the existence of an interesting ecological effect, and most would interpret those few of the 100 tested showing significance as affirmation of the original hypothesis and publication. Among the alternative hypotheses in any study is that of having discovered an improbable random event through sheer diligence—that is, if you look hard enough for a difference, you will find it.

Consider the problem when four articles on the same frog species indicate a relationship between certain microhabitat conditions and frog abundance, and two articles report no relationship. How can the results be evaluated? This is an example of a situation in which statistical and biological significance must be distinguished. The statistical test questions whether the variability in the sample indicates that one can place confidence in the result. That is not the primary interest of the amphibian biologists conducting inventory or monitoring projects. Rather, they wish to know whether the relationship shown is of biological importance because of its size and its intrinsic nature. Investigators who use statistical tests must keep in mind that the test itself merely asks if the relationship is large enough to require explanation (because it is not chance fluctuation).

Simple tests of significance should be deemphasized in favor of examination of the magnitudes of effects in all tests of hypotheses. Doing so would help eliminate noncomparability of results. The size of an effect can be measured as a function of the difference between means or the proportion of variance explained, or it can be measured by, for example, a biserial correlation (Cohen 1977; Lipsey 1990). For interpretation of the results, investigators should always publish the sample size, significance level, and power (see, e.g., Cohen 1977) of the specific statistical test used (and the observed probability level, if desired), and the size of the effect encountered in the variable studied. The types 1 and 2 error rates and calculated measures will serve as a basis for comparison of study outcomes across samples of different sizes.