Hurlbert S.H. and C.M. Lombardi, 2004. Research methodology: experimental design sampling design, statistical analysis. *In* M.M. Bekoff, (ed.), Encylopedia of Animal Behavior, 2:755-762. Greenwood Press, London.

Methods
Research Methodology

Science is structured knowledge and the processes used to generate it. This knowledge facilitates explanation, prediction and control of the real world. Systematic procedures for obtaining this knowledge are known as scientific methods, and a tremendous variety of them exist, even within a single discipline such as animal behavior. Here only a few key as-pects of research methodology are discussed.

Questions and Hypotheses

While walking through a forest, one might suddenly hear loud bird vocalizations and dis-cover a group of small birds repeatedly flying toward and then away from a tree. On closer in-spection, an owl is seen resting on a branch. After learning that the observed behavior is termed *mobbing* and reading about it, one may decide to go further and carry out his own study.

Initially one might want to make more precise observations and answer questions such as: How many bird species were involved? Were same-species birds of same sex? Same age? Did all come equally close to the owl, running the same risk of being attacked? Were they all calling at same rate? Then broader questions would arise: How had the birds recognized an owl? Why did different categories of birds behave differently? What might have been the adaptive value of the behavior in this instance? Tentative answers to these questions, formu-lated within the context of existing knowledge about mobbing, would be *research hypotheses*.

It also is possible to devise research hypotheses without making one's own observations of particular phenomena. There are no rigid rules for the finding of good research hypotheses.

Three Types of Studies

There are three main complementary methodologies for conducting scientific research: theoretical, descriptive, and experimental. Investigations that combine them can be especially informative.

In following a *theoretical approach*, one might think about a subject and develop new ideas about it, and then test the ideas by referring to existing knowledge, by using logic

and, often, by using mathematical models. One could, for example, develop a mathematical model that described how mobbing behavior affected choice of nest sites by owls. The mechanisms that induce a given behavior are represented by mathematical equations. These generate predictions which can then be compared with observed behavior. Models can be very useful in suggesting ideas for further observational or experimental studies.

Observational or descriptive approaches are a second major type. They are designed to gather more information on a phenomenon, often for the purpose of testing hypotheses, and can be simple or complex. One might use time-lapse photography as an aid to quantify how close birds come to the owl and how frequently, or use recorders to measure the frequency and nature of birds' calls, and repeat these measurements at different times of day, at different seasons, or in different locations. One could also use a dummy owl placed on a branch to see how mobbing behavior varied as a function of time of day or other variables. Observational studies require careful consideration of what the sampling design and basic sampling unit should be.

Preference trials are common in animal behavior. Examples would include presenting female subjects with different types of males and then documenting which types are selected as mates; or presenting different types of food items to animals and then recording the frequency with which each item is selected. Such preference trials are a type of observational study similar to sample surveys of human populations.

Manipulative experiments are the third major methodological approach. These entail manipulation of some experimental variable (or treatment factor) by the experimenter for the purpose of measuring its effect on one or more response variables. Their particular advantage is that they allow direct determination of causal relations. For example, field observations or theory might suggest that owl size may affect the occurrence or intensity of mobbing by other birds. To test this, one could use stuffed owls of different sizes and place these in trees on alternating occasions or at different sites and measure the responses of other birds. A disadvantage of manipulative experiments is that they usually cannot be carried out on large spatial or temporal scales that are often of great interest and importance.

Specific Methodologies

In any given observational or experimental behavioral study, there are numerous other aspects of research methodology. These are the specific field and laboratory methods, equipment, techniques, and protocols involved in selecting field sites, finding and maintaining animals, applying experimental variables, measuring responses, and recording and analyzing data. These naturally will be very different for each study, so useful generalization about them is not possible.

Experimental Design

Experimental design is sometimes used to refer to all the methods, procedures and operations involved in the conduct of a *manipulative experiment*. As mentioned above, these vary so much from one field to another and from one study to another, useful generalizations about this sense of design are difficult. The more precise and useful meaning of experimental design is the logical structure of a manipulative experiment. The purpose of such an experiment is to assess the effects of one or more experimental variables or treatment factors on one or more properties of the experimental unit. This unit can be an individual organism in a cage or tank, a group of organisms, a plot of ground, an entire lake, a bird nest, or any of a variety of other systems.

An experimental design has four aspects: *treatment structure, treatment replication, design structure*, and *response structure*. These can be defined and illustrated with an experiment to study how a fish species changes its territorial behavior according to food availability. Different levels of food supply would constitute the experimental treatments. To set up a manipulative experiment one would supply tanks with fish, add different quantities of food to different tanks, and then record fish behavior. A minimum of two groups of tanks would be needed, one set of tanks receiving a larger food ration and the other receiving a smaller one. Let's assume that we have four tanks for each treatment, and that each tank will have three fish. Ideally we assign the food levels or treatments to the tanks at random, to avoid the possibility of experimenter bias.

Treatment structure is the set of experimental treatments or treatment combinations used and how they relate to each other. In the territoriality experiment there are two treatments or levels, low and high food availability, of one treatment factor. This is the simplest treatment structure possible, since a manipulative experiment always has at least two treatments. An example of a more complex treatment structure would be if one used light intensity as a second treatment factor, using three different intensities, with a separate set of tanks set up for each of the six food availability–light intensity combinations (3 light levels x 2 food levels). When two or more treatment factors are used, the design structure is said to be (multi)factorial.

Treatment replication refers to the number of experimental units that will be subjected to a treatment. Often, but not always, this number is the same for all treatments, as in the experiment on territoriality where four tanks were established for each food level.

Design structure refers to the manner in which treatments or treatment combinations are assigned to experimental units. There are three basic design structures. The simplest would be a *completely randomized design*, where, for example, the six light–food level combinations would be assigned at random to the, say, 18 tanks available in a single array in an aquarium room.

If we only had a total of six tanks available, one might use a *randomized block design*. This would entail setting up six tanks with fish, assigning one tank to each of the six treatment combinations, and recording their observations on fish behavior. Then one would discard the fish and water from these tanks, wash them out, set the six tanks up again with new fish, re-randomize the assignment of treatments to tanks, and repeat the imposition of treatments and recording of observations. This could be repeated any number of times; each run or set would constitute a *block*.

The third basic type of design structure would be a *split-unit design*. In such there are always two or more treatment factors and the experimental unit would actually be defined differently for different factors. For example, one might have available six chambers with light controls, two at each of the three light levels. In each chamber one could place two fish tanks and assign one to the high food and one to the low food treatment.

The *response structure* consists of the list of response variables to be measured and the sampling plan that specifies when, where, and on what components of the experimental unit one will make and record observations and measurements. Each of these individual components is an *evaluation unit*. In the territoriality experiment, our principal observations would be on individual fish that were monitored for specific periods of time on specific occasions over, say, one week. One might define specific types of behaviors and record the frequency of each, estimate the size of each fish's territory, and measure variables such as quantity of food left unconsumed, concentration of dissolved oxygen, and so on. Repeated measurement of a given response variable on each experimental unit represents a *repeated measures* response structure. The sampling plan often is quite different from one response variable to another.

Experimental Design

Stuart H. Hurlbert & Celia M. Lombardi

The term *method* is commonly used within a philosophical framework, and the term *design* refers to the actual arrangement of variables used in experiments. A *variable* is anything that can change its value, and experiments have two main sorts of variables: the independent (also called experimental variable, or treatment factor) and the dependent variable. Since the latter variable in animal behavior is invariably behavior, the independent variable is whatever a researcher does to produce an effect on behavior. Although the terms treatment and dependent variable are often used with reference to both descriptive and manipulative designs, they are only properly applied to the latter, which meet the conditions required by the most widely used types of statistical analysis.

A *manipulative experiment* aims to determine, within a certain degree of probability, the effect that one or more treatments exert on one or more properties (behaviors) of some particular system (experimental unit). The experimenter must have full control over the assignment of treatments to experimental units. A thorough description of this type of design requires specification of three aspects: the design structure, the treatment structure, and the response structure.

The procedure followed for the allocation of treatments to experimental units specifies the *design structure*. When the selection procedure is at random, the design is termed completely randomized design. In the mobbing example (see the Methods—Research Methodology essay), the only recognizable difference between experimental units (small birds in this particular case) is provided by the treatments applied. However, if the researcher has some idea about inherent variation of the selected units, s/he could control it by means of what is termed blocking. In the above example, we may suppose that males and females may make up different classes whose units may behave similarly within each class. Assigning treatments at random to units of different sex results in a randomized complete block design.

Regarding *treatment structure*, there are many different forms of goals for an experiment. The investigator studying mobbing may be interested in comparing, for example, the birds' responses to aerial predators as opposed to terrestrial predators. S/he may also desire to discern whether responses change over time, or whether different degrees of the predator's dangerousness have any effect. The many different ways in which treatments may relate to one another, would make up an experiment's treatment structure.

Lastly, a design's *response structure* is specified by how evaluation units are related to experimental units. It must be decided which behaviors (dependent variables) will be measured, and how this will be done. For instance, the duration of a specified behavior or its frequency per time unit may be measured. In the mobbing example, birds may be individualized, enabling various measurements to be recorded on the same individuals (several evaluation units per experimental unit). Measuring several times the number of tail flicks each bird makes represents a repeated measures response structure.

Preference experiments are seldom manipulative but rather a type of observational study, very common and useful in animal behavior, akin to sample surveys of human populations. To illustrate, they arise when one confronts a female subject with different types of males to disclose which type she selects as a mate, or when animals are exposed to different types of seeds to study a species' diet.

Further Resources

Cook, T. D. & Campbell, D. T. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company.
Cox, D. R. 1958. *Planning of Experiments*. New York: John Wiley & Sons.
McFarland, D. 1981. *Laboratory studies*. In: *The Oxford Companion to Animal Behaviour* (Ed. by D. McFarland), pp. 327–332. Oxford: Oxford University Press.
Mead, R. 1988. *The Design of Experiments*. Cambridge: Cambridge University Press.

The details of the above aspects of an experimental design are important to understand and specify clearly, because they determine the specific types of statistical procedures that would be appropriate for analyzing the data.

Sampling Design

For observational or descriptive studies, regardless of their complexity, the parallel area is that of sampling design. This may be simply defined as the logical structure of an observational study; that is, the way in which sampling units on which measurements or observations are to be made are selected from or distributed over the sampling universe of interest.

Sampling designs come in a very wide variety. This is due in part to the fact that, unlike manipulative experiments where the focus usually is on a single scale—that of the experimental unit—many observational studies have an interest in several scales simultaneously. A nesting behavior study, for example, could be aimed at studying variation among individuals within a local population, at variation among populations in a given region, and at variation among regions.

For each scale, different types of formal sampling designs are available. The three principal ones are simple random sampling, stratified random sampling, and cluster sampling. In the *simple random sampling*, one would simply locate all nests in a local population, give each a number, and then pick at random however many nests were thought to be needed for the study. In *stratified random sampling*, if one third of the nests were in tree species A and two-thirds in tree species B, then one might select the nests to be observed so that they came from the two tree species in corresponding proportions. In *cluster sampling*, one might select, perhaps for reasons of convenience, three different points at random in the forest used by the local population, and then select for observation all the nests that were present within a 100-meter radius of each point.

Of course, applying such formal sampling designs to real animal populations can be very difficult. Often one must simply make do with whatever nests, flocks, or individuals one can find. Nevertheless it is important to understand the principles of formal sampling design because they relate to how data should be analyzed and interpreted.

Statistical Analysis

Statistical methods are logical and mathematical procedures designed to help us separate the "signal" from the "noise" in data. They help us distinguish real patterns and trends from ones that are only apparent and the result of measurement or sampling error or the inherent variability of our subjects. Well carried out, statistical analyses increase the clarity, conciseness, and objectivity with which results are presented and interpreted.

So many statistical methods are used in animal behavior studies that no brief summary of them can be useful. Readers may want to browse through some elementary statistics textbooks, but should keep in mind that, unfortunately, errors abound in many of them.

Pseudoreplication

Pseudoreplication is a serious type of statistical error that is unfortunately common in all the sciences. It was originally defined in the context of manipulative experiments, but can also occur in observational studies.

With experimental data, *pseudoreplication* occurs when measurements made on multiple evaluation units, or multiple times on a single evaluation unit, in each experimental unit are treated statistically as if each represented an independent experimental unit.

How pseudoreplication might be committed in a simple behavioral study can be demonstrated by reference to the experiment on fish territorial behavior described earlier where there were four tanks under each treatment and three fish in each tank. As the measure of territorial behavior or response variable, one might record the number of aggressive acts by each individual fish over some period of time. These data would allow calculation of the mean number of aggressive acts for each tank, and the mean number of aggressive acts for each treatment or food level (i.e., for each set of four tanks).

To determine whether fish behaved differently when they were supplied with additional food, a statistical test would be applied to compare the means of the two treatments. A valid test would entail assessing whether the difference between the two treatment means was large relative to the variation among the means for individual experimental units (tanks) within treatments. If, however, one used in such a test the variation among *evaluation units* (the 12 fish in each treatment), then they would be committing *sacrificial* pseudoreplication, the commonest form of pseudoreplication: Information on variation *among* experimental units is mixed up with that on variation *within* experimental units.

If only a single tank of three fish were set up under each food level, and if one carried out a similar statistical test for a treatment effect, then he or she would be committing *simple* pseudoreplication.

The usual consequence of pseudoreplication is exaggeration of both the strength of the evidence for a real difference between treatments and of the precision with which any difference that does exist has been estimated.

Another example shows the form that pseudoreplication might take in an observational study. A researcher wishes to estimate for a 2 km^2 (.77 mi^2) lake the mean density of nests of a fish that creates conspicuous nests as depressions on the lake bottom in shallow water. She selects one 100 m (328 ft) section of shoreline, randomly select six points along it, and establishes six band transects each 1 m (3.3 ft) in width and extending to deep water. She then swims along each of these with scuba gear, counts nests, estimates nest density for each of the six transects, and then calculates mean nest density and its standard error. This would constitute pseudoreplication if she claimed or implied that the standard error so calculated estimated the precision of her estimate of *lakewide* nest density when in fact it only reflects the precision of mean nest density estimated for the one 100 m (328 ft) section of shoreline used. One might say she had treated replicate subsamples (transects) as if they could serve as substitutes for replicate sampling units (shoreline sections) of the sort appropriate to the stated objective. To calculate the standard error appropriate to an estimate of lakewide nest density she would need to swim transects established at two or more portions of shoreline randomly selected from the lake's entire shoreline.

Pseudoreplication

Stuart H. Hurlbert & Celia M. Lombardi

Pseudoreplication is a serious type of statistical error. It was originally defined in the context of manipulative experiments, but can also occur in observational studies. With experimental data it occurs when measurements made on multiple evaluation units, or multiple times on a single evaluation unit, in each experimental unit are treated statistically as if each represented an independent experimental unit. An experimental unit is the smallest system or entity to which a single treatment is assigned and applied by the experimenter independently of other such systems. An evaluation unit is the specific component of an experimental unit on which an individual measurement is made.

Let us consider how pseudoreplication might be committed in a simple behavioral study. We wish to study how a fish species changes its territorial behavior according to food availability. Different levels of food supply would constitute the experimental treatments. To set up a manipulative experiment, we would supply tanks with fishes, add different quantities of food to different tanks, and then record fish behavior. We would need a minimum of two groups of tanks, one set of tanks receiving a larger food ration and the other receiving a smaller one. Let us assume that we have four tanks for each treatment and that each tank will have three fish. Ideally we assign the food levels or treatments to the tanks at random, to avoid the possibility of experimenter bias.

As our measure of territorial behavior (response variable), we might record the number of aggressive acts by each individual fish over some period of time. These data would allow to calculate the mean number of aggressive acts for each tank and the mean number of aggressive acts for each treatment or food level, for each set of four tanks.

To determine whether the fish behaved differently when they were supplied with additional food, we would apply a statistical test to compare the means of the two treatments. A valid test would entail assessing whether the difference between the two treatment means was large relative to the variation among the means for individual experimental units (tanks) within treatments. If, however, we used in such a test the variation among *evaluation units* (the 12 fish in each treatment), then we would be committing *sacrificial* pseudoreplication, the commonest form of this error. Information on variation among experimental units is mixed up with that on variation within experimental units. If we set up only a single tank of three fish under each food level and carried out a similar statistical test for a treatment effect, then we would be committing *simple* pseudoreplication.

The usual consequence of pseudoreplication is exaggeration of both the strength of the evidence for a true difference between treatments and of the precision with which any difference that does exist has been estimated.

Further Resources

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

Jenkins, S. H. 2002. *Data pooling and type I errors: A comment on Leger & Didrichsons*. Animal Behaviour, 63, F9–F11: http://www.academicpress.com/anbehav

Kroodsma, D. E. 1989. *Suggested experimental designs for song playbacks*. Animal Behaviour, 37, 600–609.

Mead, R. 1988. *The Design of Experiments*. Cambridge: Cambridge University Press.

Further Resources

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

Hurlbert, S. H. & White, M. D. 1993. *Experiments with freshwater invertebrate zooplanktivores: Quality of statistical analyses*. Bulletin of Marine Science, 53, 128–153.

Kroodsma, D. E. 1989. *Suggested experimental designs for song playbacks*. Animal Behaviour, 37, 600–609.

Lehner, P. N. 1979. *Handbook of Ethological Methods*. New York: Garland STPM Press.

Martin, P. & Bateson, P. 1986. *Measuring Behaviour*. Cambridge: Cambridge University Press.

McFarland, D. 1981. *Classification of behaviour*. In: *The Oxford Companion to Animal Behaviour* (Ed. by D. McFarland), pp. 63–64. Oxford: Oxford University Press.

Mead, R. 1988. *The Design of Experiments*. Cambridge: Cambridge University Press.

Stuart H. Hurlbert & Celia M. Lombardi