# 30

# Data Collection in the Zoo Setting, Emphasizing Behavior

*Carolyn M. Crockett and Renee R. Ha*

## INTRODUCTION

Systematic observations and record keeping are essential for consistent advances in the management of zoos and related facilities. Casual observations of the outcomes of innovative exhibit modifications are of much greater value when supplemented by data collected using appropriate quantitative methods. Quantification is important because qualitative observations may provide inaccurate estimates of what is really occurring. A great deal of "success" in zoo exhibitry may be serendipity—the right combination of individual animals that happen to be of a species able to thrive in marginal conditions. Only systematic data collection can lead to the conclusion that particular management decisions had anything to do with success.

This updated chapter benefits from the expertise of a second author (RRH), who has taught a zoo behavior course incorporating new data collection technologies, and whose background includes teaching statistics. We provide an overview of techniques sufficient to allow an inexperienced researcher to design and conduct a quantified study of zoo animals. Observational research on behavior is emphasized, but we suggest ways these methods can be applied to the systematic collection of other data pertinent to zoo management. For further details on methodology, serious researchers should consult Bakeman and Gottman (1997), Martin and Bateson (1993), Altmann (1974, 1984), Lehner (1996), and Sackett (1978b).

As this chapter covers a variety of topics, we recommend that the reader skim the section headers in advance for a preview of content and organization.

## PLANNING A ZOO RESEARCH PROJECT

Most research in zoos is nonexperimental. The researcher usually is unable to manipulate environmental conditions or group membership in a well-controlled manner. Collection of physical information (e.g. measurements, urine specimens) may be too invasive to perform on a regular basis. Thus, many studies are primarily descriptive and based on observational data. Information is collected, and after some period of time, an effort is made to determine what it means. Such studies frequently remain unpublished because of their unfocused and possibly ungeneralizable conclusions. This fate can be avoided by clearly identifying research questions before beginning data collection.

### FORMULATION OF A RESEARCH QUESTION

Data collection methods are designed with respect to the question being asked, and therefore, an appropriately formulated question is the first step in research design (Altmann 1974). Research questions may develop out of interest in a particular aspect of the animal's biology or behavior. Alternatively, a management issue may have arisen that requires research to address. Identifying a research question usually requires preliminary "reconnaissance" observations (Lehner 1996). In a zoo setting, possible research questions might include the following:

1. Is visitor interest higher when animals are more active? For example, Margulis, Hoyos, and Anderson (2003) evaluated the effect of felid activity on visitor interest.
2. What steps can zoos take to reduce aggression between surplus males? For example, can endogenous levels of testosterone be suppressed using Gonadotropin-releasing hormone (GnRH) and result in reduced aggression between males, e.g. several species of ungulates (Penfold et al. 2002)?
3. What behavioral indicators of pregnancy can be identified, and are they correlated with physical characteristics, e.g. lowland gorillas, *Gorilla gorilla gorilla* (Meder 1986) (fig. 30.1)?

Fig. 30.1. Zoo research might focus on the behavioral indicators of pregnancy. Lowland gorilla Nina supports one-hour-old infant Zuri, still attached by the umbilical cord. (Photography by Carol Beach, Woodland Park Zoo. Reprinted by permission.)

4. Does pacing decrease in felids when food is presented more frequently during the day (Shepherdson et al. 1993)?

## RESEARCH DESIGN CONSIDERATIONS

**Independent and dependent variables.** After identifying the research question, the next step is to identify the relevant dependent and independent variables. A variable is any property that may take on different values at different times and may change with various conditions. The values can be one of 4 types:

1. Nominal data are on a categorical, and often qualitative, scale rather than one that is quantitative.
2. Ordinal data are on a categorical scale, in which categories can be ranked in relative order.
3. Interval data are collected in a manner that measures actual magnitude and which has equal intervals between possible scores, but does not have a meaningful absolute zero point.
4. Ratio data are collected in a manner that measures magnitude, has equal intervals between possible scores, and contains an absolute zero point (table 30.1).

The property that the researcher either manipulates experimentally or records as a naturally changing condition is described as the *independent variable*. A clear distinction between the *independent variable* and the *dependent variable* is that the independent variable is the predictor variable. The dependent variable is the response variable, or what the observer actually measures. The dependent variable is often referred to as the outcome variable (Ha and Ha, forthcoming).

**TABLE 30.1.** Summary of the properties of measurement scales

| Scale | Order | Magnitude | Equal intervals | Absolute zero |
|---|---|---|---|---|
| Nominal | No | No | No | No |
| Ordinal | Yes | Some | No | No |
| Interval | Yes | Yes | Yes | No |
| Ratio | Yes | Yes | Yes | Yes |

Some independent variables are interval variables, such as ambient temperature or time of day. Others are nominal variables, such as sex (male or female), enclosure type (naturalistic or bare concrete), or physical condition (pregnant or not pregnant). It is important to consider that interval variables can be grouped into nominal categories (e.g. morning and afternoon; hot, warm, cool, cold [could also be ordinal rank of declining temperature]). Independent variables can also include age/sex composition of groups, the rearing conditions of individuals whose behavior serves as dependent variables, food delivery schedule, size of enclosure, and many others (fig. 30.2). Thus, the importance of having accurate and systematic records available to draw on becomes obvious. Furthermore, when independent variables of particular interest are identified in advance, they can be specified and filled in on each data collection sheet.

Dependent variables can include behavioral measures such as rates of aggression, sexual behavior, or play (fig. 30.2). They can also be physical measurements such as food intake or weight. Occurrence of injuries, interbirth interval length, and infant survival rate are some dependent variables that can be derived from daily reports.

**Alternative hypotheses, confounding, and bias.** Much research in zoos is descriptive in nature (we don't know what is going on and want to find out). However, research data are most amenable to statistical analysis and interpretation when null and alternative hypotheses are specified beforehand. The null hypothesis suggests that any effect or relationship between 2 variables is due to chance factors, whereas the alternative hypothesis proposes that there is an effect or relationship between the variables of interest.

Whether or not a specific hypothesis is formulated, the methodology must be appropriate for ruling out alternative hypotheses. For example, the researcher may hypothesize that males use the top branches in an enclosure more than females do. Suppose that data are collected on males in the morning and on females in the afternoon. Further suppose that these data suggest that males do use the top branches a greater percentage of the time. Under these circumstances, one cannot rule out the alternative hypothesis that animals, regardless of sex, spend more time in the top branches in the morning. In other words, time of day and sex are confounded in this study, and we cannot determine which effect (time of day or sex) is driving the result. (In this example, the independent variables are sex *and* time of day, while the dependent variable is the percentage of time spent in the top branches.)

A common goal of zoo research is to identify changes in behavior occurring as a result of a change in the zoo envi-

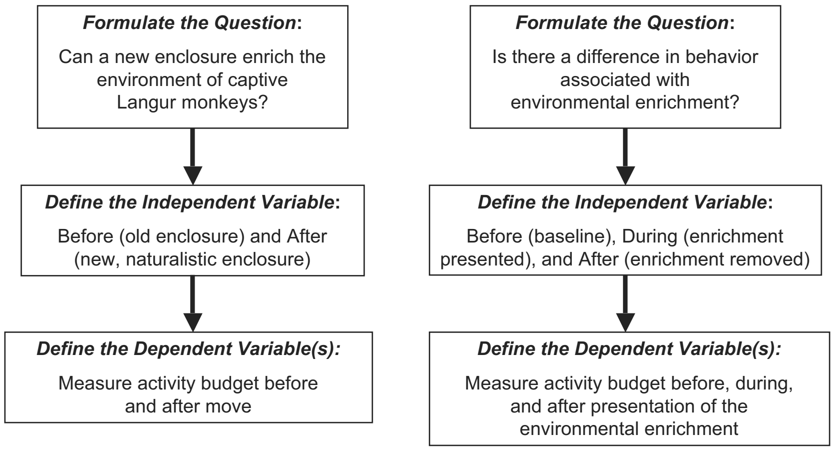


Fig. 30.2. Independent and dependent variables and 2 research designs. (Little and Sommer 2002; Young 2003.)

ronment, such as the addition of "furniture" or the introduction or loss of a group member. To assess unambiguously the effects of such a change, all other factors must be held constant. Since such control is often difficult or impossible in a zoo setting, the interpretation of results must take into account the possible effects of any extraneous, uncontrolled events. For example, if a new branch were introduced into a cage and a few days later a new infant were born, one might not be able to conclude unequivocally that changes in activity or enclosure utilization (dependent variables) were a result of one and only one of these factors (independent variables)—that is, they are "confounded." To resolve this confounding, the branch would have to be removed and reintroduced, replicating the experimental manipulation. Seasonal and weather changes may also influence the behavior of one's subjects in a manner that can confound interpretation of a project's results. These factors must be recorded systematically if their effects are to be assessed. Thus, the researcher not only needs to take into account changes that were intentionally brought about, but also must characterize factors that may represent environmental changes from the animals' point of view.

Ideally, the influence of a change, such as addition of a new form of environmental enrichment, would follow an ABA design, where A is the baseline, B is the enriched condition, and A is postbaseline, after the removal of the enrichment (Young 2003) (fig. 30.2). This type of design usually is not possible when evaluating responses to a new enclosure (Little and Sommer 2002) (fig. 30.2).

It is usually impractical and expensive to collect data 24 hours a day, every day. For this reason, sampling methods have been devised to ensure unbiased estimates of behavior based on a subset of total time. *Unbiased* means that the observations are representative of what is going on when observations are not being made, and that, when data are being collected, researchers do not inadvertently record data supporting their hypotheses at the expense of data refuting it. Observer bias will be discussed further in the section on sampling methods (see table 30.2). Lehner (1996) describes various potential sources of error in observational research, in addition to observer bias, including observer error (making recording or computational mistakes of various sorts), observer effect (affecting the behavior of the subjects by being present), and errors of apprehending (when the physical location or attributes of the subject make it more or less visible than other potential subjects).

**When and how often to collect data.** Another preliminary consideration in research design is when to observe. If the re-

**TABLE 30.2.** Summary of sampling methods

| Sampling method | Scoring basis | Mutually exclusive | Exhaustive | Comments and uses |
|---|---|---|---|---|
| Ad libitum | Behavior change | No | No | Longhand field notes. Preliminary observations; ethogram development; reconnaissance observations. |
| Continuous | Behavior change | Yes | No | For frequencies (onsets) of selected behaviors, especially infrequent behaviors of short duration. |
| | | Yes | Yes | When relative frequencies are to be calculated from onsets (table 30.3). |
| | | Yes | Yes | For transition times (to calculate durations if start and stop times are recorded during data collection). Time budgets can be calculated from mutually exclusive behaviors with start and stop times. |
| Scan/instantaneous | Time-point | Yes[1] | Yes | Especially useful for time budgets, activity patterns, group behavioral synchrony; usually produces high interobserver reliability. More appropriate for states than events (table 30.3). |
| One/zero | Time-interval | Yes[2] | Yes | Not recommended except for special circumstances (see text). |

[1]Simultaneous behaviors can be scored and later combined into mutually exclusive categories.
[2]More than one mutually exclusive category can be scored per interval.

search question focuses on diurnal variation in behavior, then all time periods of interest must be sampled (Brannian and Cloak 1985; Heymann and Smith 1999; Vickery and Mason 2004). It may be practical to eliminate the hours of darkness from the sample if preliminary observations indicate that the animals are mostly inactive then. Around-the-clock observations are essential for studies of parturition and other events whose exact timing may be impossible to predict (Robeck et al. 2005).

To study day-to-day changes in behavior, such as correlates of estrous cycles or infant development, daily or almost daily records are necessary. If the amount of time available for data collection is limited, making observations at the same time each day will eliminate the confounding factor of time of day. However, this will also sacrifice the ability to generalize to other time periods unless diurnal variation in behavior has been ruled out first. If specific behaviors are of interest, preliminary observations will determine the best times to record them. For example, preliminary observations of ibex revealed that 95% of play (the behavior of interest) occurred between 0500 and 0700 and between 1900 and 2115, so observations were done at those times (Byers 1977).

Longitudinal studies (e.g. developmental) raise the question of how often observations must be made in order to provide valid estimates and yet be practical from a time and resource point of view. Kraemer et al. (1977) suggested a method for evaluating the spacing and timing of observations to minimize sample error and cost of data collection. For physical data (e.g. weight) that cannot be taken daily, records at approximately equal intervals are desirable (e.g. once a week). Weights should be taken at approximately the same interval since last feeding (Kawata and Elsen 1984).

**TABLE 30.3.** Terms pertinent to behavioral data collection

| Term | Definition |
|---|---|
| Event | The onset or the single defining instant of any behavior; instantaneous behavior; momentary behavior (Sackett 1978a). |
| State | Behavior with appreciable duration (durational behavior), or any behavior at a given instant in time. |
| Duration | Time spent in a state. |
| Transition time | Time of onset or termination of behavior; changing from one state to another. |
| Frequency | Number of occurrences; can refer to events or states (see "bout"). Try not to be confused by the fact that in genetics gene "frequency" refers to the proportion of an allele in the population, and that in other contexts "frequency" is a "rate" (occurrence per unit time; see below), such as radio frequency. |
| Bout | One occurrence of a durational behavior or a behavior sequence (e.g. a play bout). |
| Rate | Frequency (number of occurrences) per unit time; requires knowledge of sample duration. Rates are most usefully interpretable when translated to a common time base, e.g. frequency per hour (see table 30.4). |
| Exhaustive | Behavior taxonomy is all-encompassing; subject is always recorded as doing something, even if "not visible" or "other." |
| Mutually exclusive | Recording categories do not overlap; within a given set of categories, the subject is never recorded as doing more than one thing simultaneously. |

*Note:* Several definitions are paraphrased from Altmann (1974).

**Determining what information is important.** Determining what types of information are needed to answer a research question requires a reading of the relevant literature on the topic or species in question and preliminary observations. Knowing what has been done before may suggest useful techniques and avoid unnecessary duplication. Decide what behaviors are of interest and what parameters are of biological importance (Altmann 1984). For example, is it more relevant to know how often the behaviors occur (e.g. hourly rate), how much of the time is spent in particular activities (percentage of observation time), or how long the animals tend to engage in a behavior once it begins (bout duration) (see tables 30.3 and 30.4)? Determine whether sequences of behavior are important, as in courtship interactions. Their recording and analysis greatly complicate a research design (Lehner 1996; Bakeman and Gottman 1997).

Decide whether identification of individual animals is essential, e.g. to record actors and recipients of social interactions. In some cases, subjects can be lumped into age and sex classes without loss of essential information. If identification is necessary, marking of individuals may be required (see Kalk and Rice, appendix 2, this volume). If enclosure use is a subject of study, obtain accurate maps or blueprints of the exhibit.

**Preliminary analyses.** As a final preliminary consideration, data collection methods should be planned with some thought to subsequent data analysis. A good rule is to try some preliminary analyses after some initial data collection. Determine whether all the research questions posed are indeed answerable with the method chosen. *Preliminary analyses are important.*

## GATHERING DATA FOR THE RESEARCH PROJECT

### DEFINING WHAT DATA TO RECORD

To record research data systematically, appropriate definitions of behaviors or other types of data must be developed. Precise definitions for each element to be recorded must be written out, to ensure that observers do not "drift" from the original definition and to enable other researchers to use the same recording system. Part of this task follows from prior identification of independent and dependent variables, as all must be defined in some way. In general, defining recording categories for nonbehavioral data is more straightforward than developing them for behavioral data. Catalogs of an animal's behavioral repertoire, also known as a behavioral inventory or taxonomy, are called ethograms. For behavioral and nonbehavioral categories, a thorough literature search will reveal whether adequate categories have already been defined. When preexisting categories are used, not only does the researcher avoid "reinventing the wheel," but the previous literature can also be cited, thus shortening a manuscript prepared

**TABLE 30.4.** Useful calculations for analyzing behavioral data

| Calculation | Definition |
|---|---|
| Raw scores | Unadjusted totals per observation (or focal sample) period (e.g., total occurrences per behavior, recorded with any sampling method); can be used in statistical tests if all observation periods are of equal duration. |
| Adjusted or corrected scores | Raw scores weighted so that all scores are equivalent (e.g., to adjust when observation periods are unequal across subjects or days. |
| Proportion | A fraction expressed in decimals, e.g., 5/8 = .63. |
| Probability | Expressed by a proportion; for example, if a study's results show that during the full moon an average of 5 of 8 females in a group are in estrus, one may conclude that the probability of any female being in estrus during the full moon is .63. |
| Percentage | Same calculation as proportion but multiplied by 100 so that unity = 100% (unity for proportions and probabilities = 1.0) |
| Range | Highest and lowest score (e.g., of frequencies, durations, rates, percentages). |
| Mean | The sum of the scores / sample size or number of scores (*N*). |
| Median | The midpoint of the scores (half are greater and half are smaller). |
| Variability | Measures of variation in scores about the mean; see any general statistics book for calculating standard deviations and other variability (error) measures. |
| Rate (e.g., of occurrence of solitary behavior or social interaction) | Frequency / observation time. |
| Hourly rate (frequency per hour) | Frequency / hours of observation, in decimals. |
| Relative frequency | Frequency of one behavior / total behavior changes (total number of behaviors); indicates probability of a particular behavior being observed at a randomly selected behavior change (Sackett, Ruppenthal, and Gluck 1978). |
| Mean duration per bout | Total duration of a behavior / its frequency. |
| Mean duration per hour (mean minutes per hour in a state) | Total duration in minutes / hours of observation, in decimals. |
| Mean rate (or duration or percentage) per individual (e.g., averaged across the entire group or within age/sex classes) | Sum of mean rates (or durations or percentages) for all individuals / total number of individuals in group (or subgroup). |
| Percentage of time (continuous sampling)[a] | (Total duration of behavior / total duration of observation) × 100. |
| Percentage of time (scan sampling)[a] | (Number of point samples when behavior was scored / total number of point samples) × 100. |

[a]When these percentages are expressed as proportions, they indicate the probability that a given behavior will be seen during any randomly selected moment.

for publication. This practice also facilitates direct comparisons with the results of prior research.

**Ethograms.** In the early days of ethology (the study of how natural selection shapes adaptive behavior), an ethogram was always the first step and was sometimes itself the objective of many years' study (Tinbergen 1951; Lorenz 1958). Defining behaviors is still an essential step, but the extensiveness and detail with which this needs to be done depend on the specific question at hand. One of the first tasks of a project is to formulate a list of well-named, carefully defined behaviors relevant to the research objectives. Select the behaviors essential to a study to avoid being swamped during data collection (Hinde 1973).

Behavior descriptions are of 2 basic types, empirical and functional (Lehner 1996): Empirical, objective descriptions include body parts, movements, and postures, whereas functional descriptions include interpretations as to the purpose of the behavior. In general, when formulating an ethogram, first try to use objective names and operational definitions and avoid subjective inference regarding function. For example, in describing a facial expression common to many monkeys, "open-mouth stare" is more objective than "open-mouth threat" (fig. 30.3). The function of some behaviors, such as nest building, may be readily agreed on, but still need to be described for different species (Lehner 1996).

Researchers may find, after some experience, that it is appropriate to lump behaviors into a larger functional category such as "threat" or "aggression." This may occur during, or as a result of, data analysis. A behavioral taxonomy might be restricted to discrete categories of behavior. On the other hand, researchers not especially concerned with sequences of behavior might record fairly predictable sequences, such as "copulation" and "rough and tumble play," as single units of behavior (G. P. Sackett, personal communication). If several types of behavior are included within one scored category, each type should be described in the ethogram. For some classes of behavior, observer judgment is very important. For example, in discriminating between rough play and aggression in monkeys, the ability to make reliable judgments may require many hours of observation to develop.

Some examples of ethograms for studies conducted in zoos and similar facilities are published (Byers 1977, pp. 201–2; Freeman 1983, p. 7; Kleiman 1983; Stanley and Aspey 1984, pp. 91, 94–95, 103; Traylor-Holzer and Fritz 1985, p. 119; Nash and Chilton 1986, p. 40; Tasse 1986, p. 119; Macedo-

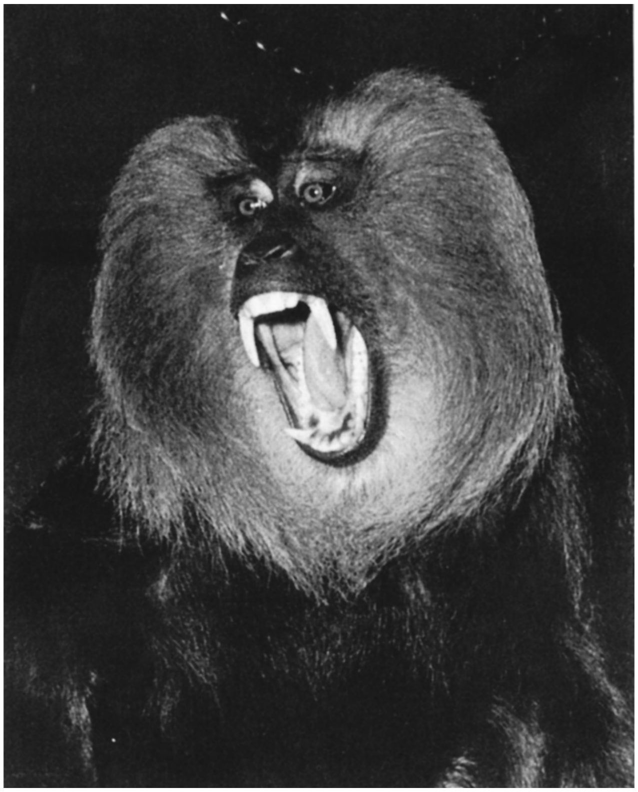

Fig. 30.3. In formulating an ethogram, use objective names and operational definitions. The function of this open-mouth expression given by an adult male lion-tailed macaque should be verified from quantitative observations. (Photography by Joy Spurr, Woodland Park Zoo. Reprinted by permission.)

nia 1987, p. 58; Merritt and King 1987; Margulis, Whitham, and Ogorzalek 2005, p. 630, including definitions for recording spatial locations in evaluating enclosure use; White et al. 2003, p. 274). The Behavioral Advisory Group of the American Zoo and Aquarium Association, facilitated by Lincoln Park Zoo, Chicago, maintains a Web site of ethograms of zoo animals: www.ethograms.org (Behavioral Advisory Group 2002).

**Exhaustive and mutually exclusive recording categories.** For purposes of data recording and analysis, it is often advantageous (and for some sampling methods, necessary) to define categories that are both exhaustive and mutually exclusive. *Exhaustive* means that the subject (S) is always recorded as doing something, even if "inactive," "other," or "not visible." *Mutually exclusive* means that the subject is never recorded as doing more than one thing simultaneously; that is, S can be "sitting" or "grooming," but not both. The recording system should include rules for establishing priorities or precedence, such as recording the "action" rather than the "posture" (Sackett 1978a). For example, a tiger, *Panthera tigris*, might be lying down but licking its paw, and this would be recorded as grooming, not lying down. Within a particular scoring system (e.g. a check sheet), more than one set of mutually exclusive and exhaustive categories can be included: e.g. the subject could be scored, simultaneously, for one behavior, one location, and one proximity relationship (e.g. nearest neighbor identity and distance).

**Codes.** Codes are useful for recording behavior in a variety of sampling schemes. Depending on the number of behaviors to be scored, one may simply code each behavior with one to 3 letters or numbers. When there are many behaviors to record and codes to memorize, reliability is improved by use of mnemonic abbreviations, such as GR = groom and AP = approach, or a dimensionalized coding scheme in which the first letter or number designates a general category and the second, the specific behavior, such as LW = locomotion-walk, LC = locomotion-climb, HG = handle-groom, HH = handle-hit (Bobbitt, Jensen, and Gordon 1964; Sackett, Stephenson, and Ruppenthal 1973; Astley et al. 1991; Lehner 1996, pp. 240–41).

Codes also can be used to identify individuals, actors and recipients, and locations. When developing codes that eventually will be analyzed by computer, keep in mind what the available computer system or existing programs can handle. If a coding system is incompatible with an analysis package, it is relatively easy to modify codes with the Find and Replace features of Microsoft Excel.

## CHOOSING SAMPLING METHODS

Sampling methods are used to make estimates about an entire population (e.g. all lions in captivity) based on a subset, or sample, of that population (e.g. the lions in one zoo observed for 200 hours). Certain methods of sampling have been devised to ensure that the estimate obtained is unbiased (Altmann 1974). Even though a research project usually has predefined categories of all the possible things to record, some behaviors, individuals, or locations might be momentarily more interesting than others. If who, what, or when to observe were entirely up to the observer's whims, his or her data recording might focus on certain events to the exclusion of others that also had been predetermined to be important. This is the essence of observer bias.

Table 30.2 summarizes the major sampling methods, table 30.3 gives some pertinent definitions, and table 30.4 presents some useful calculations.

**"Focus" of observations.** The most common focus is on a single individual ("focal animal"), and all behaviors of interest initiated by that animal are recorded. In some sampling systems, all interactions in which the subject (S) is the recipient are also recorded. Although recording S as both actor and recipient allows one to collect more complete information about interactions, this protocol requires special consideration during data analysis. If one chooses to focus on one animal at a time, then total observation time may have to be increased if each focal subject is to occur often enough in the sample to be characterized adequately. The focus can be an individual, subgroup, group, or behavior, depending on the research question and the appropriate sampling method:

1. Focal animal: selected from the total group or a subset of it. Note that what Altmann (1974) called "focal-

animal sampling," we call "continuous sampling" (see "Continuous Sampling," below, and Altmann [1984]).
2. Focal subgroup: for example, "mother-infant pair" or "all females."
3. Group or subgroup, one individual at a time (see "Instantaneous and Scan Sampling" below) (Martin and Bateson 1993).
4. All occurrences of certain behaviors (Altmann 1974) or behavior sampling (Maestripieri 1996): focusing on the total group while restricting attention to certain behaviors, such as aggression, sexual behavior, or a particular facial expression.
5. Sequences of behavior (Altmann 1974, sequence sampling): Sequence sampling was effectively used by Byers (1977).

**Random sampling and balanced observations.** To avoid observer bias, the order in which focal subjects are sampled during each observation period should be randomized (fig. 30.4). Random sampling can be accomplished by using the table of random numbers found at the end of most statistics textbooks, or with the RAND() function in Microsoft Excel. An easy way is to write each subject's name on a small card. Shuffle the cards, put them in an envelope, and select one. Repeat until all the cards have been drawn and their order recorded. This is random sampling without replacement, which ensures that each subject is observed only once during an observation period. Random sampling should be repeated for each observation period. Remember that if subject A's card is drawn and A is not visible, data must still be recorded on this individual under the "not visible" category. Subject A may appear sometime during the sample period.

A methodology in which observation times were selected at random rather than being prescheduled would reduce other sources of bias. However, interobservation variability might swamp any meaningful results unless a large number of observations were made at each time of day to eliminate the potential error introduced by diurnal variation in behavior. Given the nature of the zoo setting and the schedules of observers, many of whom are zoo staff or students, observation times are unlikely to be randomized. Under such circumstances it is more important for them to be "balanced," that is, to schedule the same number of observation periods during each of several selected time blocks. If several time blocks are being sampled and observations occur only once a day, some effort should be made to avoid scheduling consecutive days' observations during the same time block; this will reduce bias imposed by abnormal streaks of weather or other factors (i.e. confounding of weather and time-of-day effects). Such potential bias is eliminated if all subjects are observed daily during all time blocks sampled. If daily observations are not possible, evenly spaced observations, such as every third day, provide "balance" as long as there are no behavioral cycles coinciding with the same interval. If at all possible, a pilot study should be conducted to determine the optimal observation schedule (Kraemer et al. 1977; Thiemann and Kraemer 1984). Scheduling observation periods well in advance will allow the project to run more smoothly, especially if arrangements for after-hours admission must be made. Times of day routinely allocated for daily husbandry activities should be avoided unless related to project goals.

**Bases for recording data.** Essentially, there are 2 kinds of events that activate the observer to record data: a change in behavior or the passage of time (Sackett 1978a). A *behavior change* scoring system, as the name implies, usually involves recording the onset of a new behavior, but it may also include recording the termination of the current behavior or the transition time between 2 behaviors. Behavior-change scoring usually is associated with continuous sampling sys-

Fig. 30.4. To avoid observer bias, observe focal subjects such as these patas monkeys, *Erythrocebus patas*, in random order. (Photography by Mark Frey, Woodland Park Zoo. Reprinted by permission.)

tems. For some behaviors the transition from one to another "bout" (see table 30.3) can be ambiguous. In such cases, the behavior taxonomy should include defining events that signal when a new behavior should be recorded: e.g. a certain number of seconds of inactivity that must elapse before a new behavior bout is recorded, or a certain critical distance that must be reached before "approach" is scored.

In a *time sampling* scoring system, the observer either scores the behavior occurring at the moment of a transition between intervals (scan, instantaneous, or point sampling), or scores the occurrence or nonoccurrence of each behavior of interest during the interval (one-zero sampling). A stopwatch or other device with a programmable alarm is used to signal the end of an interval. These methods and the factors contributing to choice of time interval length are discussed below.

**Sample period.** For ease of data analysis, it is useful to divide observation periods into equal-length sample periods. There are several types of sample periods, but generally the primary or focal sample period is considered to be the length of time during which a particular individual or behavior is the focus of observation. Individual subjects are the most common focus, so the more individual subjects there are to be observed during the observation period, the shorter the focal sample period will be, or the whole sampling period could be longer. However, increasing focal sample duration will reduce between-sample variability, which is desirable for some kinds of analysis.

A simple system is to define a basic observation period that includes a complete replication of data collection; i.e., each subject is observed once and only once in random order. Let's say that the basic observation period is one hour. If 5 subjects are to be observed, then the focal sample period ought to be 10 minutes, providing an additional 10 minutes during the basic observation period to shuffle papers and to deal with unexpected events or to record different kinds of data between focal samples. Within each focal sample period, smaller time intervals may be employed, as in all time-sampling scoring systems or to keep a time base in continuous sampling. When methodology dictates collecting more than one kind of data, define the basic observation period to allow for this. When there is only one subject, or when the whole group is observed at once, the basic observation period is synonymous with the focal sample period. The length of the basic observation period should be shorter than the "fatigue threshold," which is likely to be reached faster when a noisy public is present to distract the researcher. A focal sample period should not be less than 5 minutes, so if the group is large, it might have to be observed over more than one observation period.

Although projects by zoo staff and students may be constrained by other schedules, or by the nature of the project itself, for the sake of data analysis and statistical tests it is best for each observation day to be uniform in terms of total observation duration and the number of focal samples taken.

## SAMPLING METHODS: USES AND LIMITATIONS

**Ad-lib sampling.** Ad-lib sampling (Altmann 1974) is equivalent to traditional field notes or reconnaissance observations and generally involves nonsystematic, informal observations preliminary to quantified study. This technique is useful for recording rare, unusual events and often takes the form of a comments section on the data sheet.

**Continuous sampling.** In continuous sampling (focal-animal sampling: [Altmann 1974]; continuous real-time measurement: [Sackett 1978a]), the start time (and, for durations, ending time) of specified behaviors and interactions are recorded. This behavior-change method usually records behavior initiated by (and in some protocols, directed toward) focal subject(s), but can be modified to record focal behaviors, sequences, or use of enclosure locations.

Continuous sampling always allows for the calculation of frequency, rates, and (if stop times recorded) durations of behavior (table 30.2). Continuous sampling of a focal animal potentially allows for the most complete record of behavior and is the only way to collect data on sequences without missing anything. Analyzing continuous data can be very time-consuming if many behaviors or subjects are involved, unless electronic recording devices are used. If sequences are not important, and a computer is not to be used, a check sheet can be designed to simplify data collation and analysis. If the behaviors of the most interest are momentary or relatively infrequent, continuous sampling is the method of choice. If the frequency of behaviors is the main interest, then only the onset of behavior need be recorded, simplifying the analysis.

**Instantaneous and scan sampling.** Instantaneous and scan sampling (Altmann 1974), also known as point sampling (Dunbar 1976), are time-sampling-based systems in which the observer records the behavioral state (table 30.3) at the instant ending a predefined interval—e.g. on the minute. To avoid bias, the observer must record only what the subject is doing at that instant, whether an ongoing behavior, the onset of a new behavior of some duration, or a brief behavior that happens to coincide with the sampling instant.

One potential problem with these methods is the difficulty of identifying a particular behavior or subject at a single glance. An effective solution is to observe the subject for, say, 5 seconds after the signal and then record the behavior observed at the last instant (e.g. on the count of five) (Sackett 1978a). This "count-to-five" method worked very well in a field study of red howler monkeys, *Alouatta seniculus*, scanned at 15-minute intervals (C. M. Crockett, personal observation). When the time intervals are short ($\leq$30 seconds), the observer is likely to anticipate the next time signal so that behavior determination can be made without the counting method. Some researchers record the first behavior that lasts for a defined duration, such as 5 seconds (Mahler 1984, "sustained" behavior), but this leads to underrepresentation of instantaneous behaviors and should be avoided (Clutton-Brock 1977). If the main interest is instantaneous "events" rather than "states" (tables 30.2 and 30.3), then continuous sampling is more appropriate.

Instantaneous sampling refers to time-activated recording methods in which the focus is a single individual (the reason to avoid using Altmann's [1974] term *focal-animal sampling* to refer to the continuous sampling method). Scan sampling involves scoring an entire (sub)group, hence the observer must visually "scan" to record the behavior of all individuals. Although it takes more than an "instant" to scan a group, the observer records only the behavioral state occurring when each individual is first seen. To avoid bias, scans should be performed in a systematic manner, such as always from the left to the right of the enclosure. In principle and in common usage, "instantaneous" and "scan" sampling are equivalent.

Scan sampling provides the easiest method for estimating the percentage of time spent in specific activities or percentage usage of different enclosure locations (table 30.4). Scan sampling is thus particularly well suited to studies of activity cycles (variation in behavior as a function of time of day). It is less suitable for collecting data on specific social interactions, since they often occur in sequences that cannot be recorded using a scan sample. Infrequent behaviors of short duration are generally missed unless the interval between scan samples is very short or the total duration of observation is long. Rates and bout durations cannot be calculated with this method. The great advantage of scan sampling is its relative simplicity: naive observers can quickly learn to score clearly defined behaviors if the number to choose from is relatively small. Thus, inter- and intraobserver reliability is usually high.

The interval length chosen for scan sampling depends on various factors, such as the subject's activity level (how often it changes behavior, and how long the behaviors scored typically last), group size (how many individuals are to be scanned per interval), whether a single or a mixed sampling strategy is to be used, and whether temporal autocorrelation is an issue in statistical analysis. In general, the shorter the interval, the closer data collection approximates what can be recorded with continuous sampling. Shorter intervals, however, mean more data to analyze, since data are scored for each interval. Longer scan intervals are more practical for relatively inactive animals, especially when combined with continuous sampling of selected behaviors of brief duration (i.e. a mixed sampling strategy). Some types of information, such as food intake or animals' locations plotted on a map of the enclosure, can be recorded only once a day and can still be treated as a scan sample. For statistical purposes, once-a-day records generally avoid the problem of temporal autocorrelation.

**One-zero sampling.** In one-zero (or 1-0) sampling (Altmann 1974), also known as modified frequency (Sackett, 1978a), time intervals are established just as in scan sampling. However, each behavior category occurring during the interval is given an arbitrary score of 1 regardless of its true frequency. For example, a behavior observed 5 times during an interval is still scored as 1, and a behavior of longer duration is given a score of 1 for every interval in which it occurs, regardless of onset. Thus, more than one behavior category can be scored per interval.

Because true durations, true frequencies, and true percentages of observation time spent in different activities cannot be calculated with this method, Altmann (1974) advised that it not be used. In response, a number of studies were published comparing how estimates of rates, durations, and percentages of time varied depending on the sampling method used to score the same series of events (Dunbar 1976; Chow and Rosenblum 1977; Leger 1977; Sackett 1978a; Kraemer 1979; Tyler 1979; Rhine and Ender 1983; Suen and Ary 1984). The results indicate that, although the 2 time-sampling-based methods provide results that are generally positively correlated with one another, the degree to which they reflect the true occurrence of behavior depends a lot on the sampling interval length relative to behavior rate and bout duration (Suen and Ary 1984). Of course, average rate and duration will vary from behavior to behavior. Where bouts or flurries of specific behaviors are of greater interest than specific rates or time budgets, the simplicity of one-zero sampling might make it an acceptable choice, but be aware of its drawbacks (Bernstein 1991).

One-zero sampling should be avoided when estimates are to be compared with those of other studies using other methods. However, because one-zero is easy to score and analyze and produces high interobserver reliability, it can be employed when many observers are to be used or direct comparison with other studies is not important. Nevertheless, proper training and data collection design usually can achieve equally high interobserver reliability in studies using scan sampling.

One-zero sampling can also be used to quantify past daily reports in which the information recorded is accurate only to that level. For example, occurrence or nonoccurrence (1-0) in the written record can be scored for sexual behavior, consumption of particular foods, use of a new cage furnishing, fresh injuries, and so on for each individual present that day. Some events tend to be biologically important at the one-zero level, e.g. whether a female mates at least once during estrus or whether an animal eats at least once during a day. Such one-zero scoring of keepers' records was used effectively to supplement systematic data on proceptive calling by female lion-tailed macaques, *Macaca silenus* (see Lindburg 1990).

## DATA RECORDING SYSTEMS

There are many ways to record data, and they vary in their reliability, ease of use, cost, and time required for transcription and analysis. Audio- and video-recorded data, for example, require at least twice the time to transcribe as to record. However, video or audio recording an ongoing event that is unpredictable, such as the introduction of a new animal, may be the most successful way to preserve rapidly occurring interactions. Handycams are a good option, yielding digital files that can be coded by various methods. Transcription is easier if the observer narrates ongoing behavior using memorized codes. Laptop computers or personal digital assistants (PDAs) can be programmed to accept coded data (entered by keyboard, touch screen, voice recognition software, or a barcode reader) that can then be analyzed by the device itself or transferred to a desktop computer for analysis (Forney, Leete, and Lindburg 1991; Grasso and Grasso 1994; Paterson, Kubicek, and Tillekeratne 1994; White, King, and Duncan 2002). Commercially available products can turn a personal computer or a PDA (fig. 30.5) into a behavior coding and tabulating system. Among these are The Observer, www.noldus.com/ (Cronin et al. 2003, includes example of use; Noldus 1991, 2005), EVENT (Ha 1991; Ha and Ha 2003, includes example of use), and JWatcher, www.jwatcher.ucla.edu/ (Blumstein, Evans, and Daniel 2000). Computer tech-

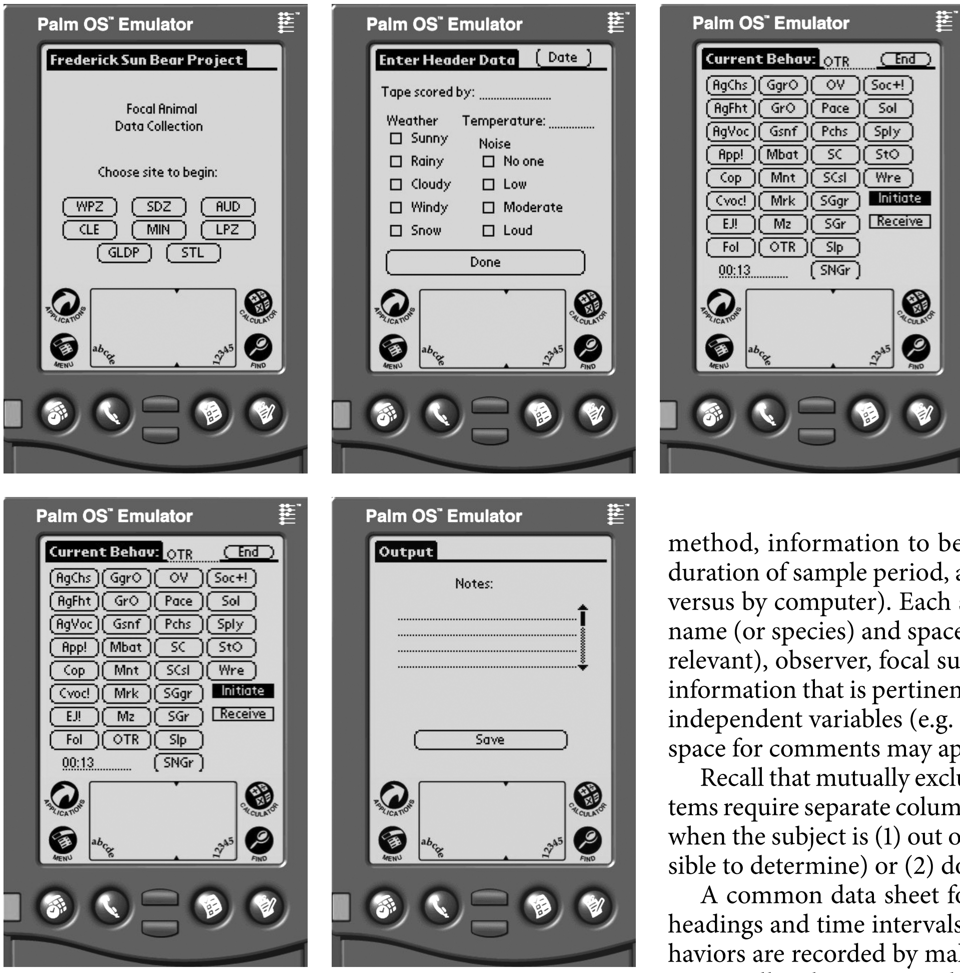


Fig. 30.5 (*A–E*). EVENT-Palm Software. Cheryl Frederick of the Woodland Park Zoo, Seattle, and University of Washington worked with James C. Ha (1991) to develop a custom PDA program to collect focal data on endangered sun bears at 6 zoos across the United States. Users touch the screen with a stylus to select coded behavior buttons, and the data are recorded into a database program for later analyses.

nology is the method of choice when large amounts of data are to be collected. The advantages of these techniques include simultaneous data entry with data collection, the potential of safeguards in the program to prevent "impossible" entries, and elimination of transcription error from data entry errors. However, for many projects, paper and pencil data sheets are perfectly adequate, are more cost effective, and are the recommended starting point for beginning observers.

## PAPER AND PENCIL METHODS

For many zoo research projects, a photocopied *data sheet* is a suitable and inexpensive method of recording data. Experiment with preliminary versions before a final version is adopted. Professionally printed NCR (no carbon required) paper is a good choice if duplicate data records are important.

Hinde (1973) gives a number of useful suggestions regarding the format of data sheets. Published papers rarely include samples of the data sheet used, but examples can be found (see Kleiman 1974; Price and Stokes 1975; Crockett and Hutchins 1978; Lehner 1996; Paterson 2001). Figures 30.6 through 30.8 present "generic" data sheets suitable for different sampling methods and purposes. The data sheet format that a researcher selects will be a function of sampling method, information to be recorded, number of subjects, duration of sample period, and method of analysis (by hand versus by computer). Each sheet should include the project name (or species) and spaces to enter date, time, weather (if relevant), observer, focal subject, location in zoo, and other information that is pertinent to the project and may serve as independent variables (e.g. phase or conditions of study). A space for comments may appear on the data sheet.

Recall that mutually exclusive and exhaustive scoring systems require separate columns, categories, or codes to record when the subject is (1) out of sight (and where, if that is possible to determine) or (2) doing something undefined.

A common data sheet format lists behaviors as column headings and time intervals as row headings (fig. 30.6). Behaviors are recorded by making a check mark in the appropriate cell or by entering the code of the recipient of social behavior or the location of the focal animal. This format is suitable for time sampling (fig. 30.6, left) and for continuous sampling of behavior frequencies (fig. 30.6, right) when sequence is not important. When a format such as that shown in figure 30.6, left, is used to scan more than one individual per interval, each individual's ID code could be entered in the appropriate cell.

To record continuous sequences, codes for actors, behaviors, and recipients can be written in the order in which they occur, using the first column of each row to enter time of onset (fig. 30.7, top). Alternatively, time intervals can be prelabeled such that behavior is recorded in the row indicating the minute period (or other time interval length) in which it occurred (fig. 30.7, bottom). Durations can be estimated if a mutually exclusive and exhaustive set of behaviors is recorded, and it is predetermined which ones are "events" (e.g. ca. one second duration) and which are "states" (variable duration). The onset of the next behavior is assumed to terminate the previous one. Transcription of data recorded with this method is tedious and time-consuming unless a computer is used.

Maps can be used to record various kinds of data. On a scale map of the enclosure, one can code each animal's location, using a scan sampling technique. Later, interindividual distances and location preferences can be calculated from map plots, as done by Kirkevold and Crockett (1987). It may

GENERIC TIME SAMPLING DATA SHEET

Date: *10/2/86* Species: *LTM* Subject: *Fred* Observer: *CMC*
Enclosure: *Out* Weather: *Sun, 75*
Start Time: *0900 h*

| Interval: | Beh.1 | Beh.2 | Beh.3 | Beh.4 | Beh.5 | Beh.6 | Other | NotVis | Comments |
|---|---|---|---|---|---|---|---|---|---|
| 0:00:15 | 1 | | | | | | | | |
| 0:00:30 | 1 | | | | | | | | |
| 0:00:45 | | 1 | | | | | | | |
| 0:01:00 | | | | | | 1 | | | |
| 0:01:15 | | | | | | | | 1 | |
| 0:01:30 | | | | | | | | 1 | |
| 0:01:45 | | | | | | | | 1 | |
| 0:02:00 | | | | 1 | | | | | |
| 0:02:15 | 1 | | | | | | | | |
| 0:02:30 | 1 | | | | | | | | |
| 0:02:45 | | | | | | | 1 | | |
| 0:03:00 | | | 1 | | | | | | |
| 0:03:15 | | | 1 | | | | | | |
| 0:03:30 | | | 1 | | | | | | |
| 0:03:45 | 1 | | | | | | | | |
| 0:04:00 | 1 | | | | | | | | |
| 0:04:15 | 1 | | | | | | | | |
| 0:04:30 | | 1 | | | | | | | |
| 0:04:45 | | 1 | | | | | | | |
| 0:05:00 | 1 | | | | | | | | |
| TOTAL: | 8 | 3 | 3 | 1 | 0 | 1 | 1 | 3 | 20 |
| PERCENT: | 40 | 15 | 15 | 5 | 0 | 5 | 5 | 15 | 100 |

Comments: *Data sheet filled in is example of scan sampling.*

GENERIC BEHAVIOR FREQUENCY DATA SHEET

Date: *10/7/86* Species: *LTM* Subject: *Fred* Observer: *CMC*
Enclosure: *Out* Weather: *Cloudy, 65*
Start Time: *1000 h*

| Minutes: | Beh.1 | Beh.2 | Beh.3 | Beh.4 | Beh.5 | Beh.6 | Beh.7 | Beh.8 | Comments |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | 11 | | | 1 | | | | | |
| 4 | | 1 | | | | | | | |
| 5 | | | | | | | | | |
| 6 | | | | | 11 | | | | |
| 7 | | | | | | | | 1 | |
| 8 | | | | | | 111 | | | |
| 9 | 1 | | 1 | | | | | | |
| 10 | | | | | | | | | |
| 11 | | | | | | | | | |
| 12 | | | 11 | | | | | | |
| 13 | | | | 1 | | | | | |
| 14 | | | | | | | | | |
| 15 | | | | | | | 1 | | |
| 16 | 1 | | | | | | | | |
| 17 | 111 | | | 1 | | | | | |
| 18 | | | | | | | | 1 | |
| 19 | | | | | | | | 11 | |
| 20 | 1 | | 11 | | | | | | |
| TOTAL: | 8 | 1 | 5 | 3 | 2 | 3 | 1 | 4 | = 27 onsets |

Comments: *Only intervals with behavior onsets have tallies.*

Fig. 30.6. (*Left*) Time-sampling data sheet for 8 mutually exclusive and exhaustive behavior categories. For scan sampling, the behavior occurring at the instant of the interval marker is checked; there is only one tally per row (interval), as shown here. For one-zero sampling, all behaviors occurring during the interval would be checked once. (*Right*) Data sheet for recording behavior frequency during continuous sampling. Behavior onsets are recorded by checking the cell corresponding to the time interval of occurrence. Multiple tallies may occur in one cell, and some rows (intervals) may have no tallies because no new behavior onsets occurred.

DATA SHEET FOR RECORDING SEQUENCES (CONTINUOUS SAMPLING)

Date: *10/7/86* Species: *LTM* Subject: *A* Observer: *CMC*
Enclosure: *Indoor* Weather: *Not applicable*
Start Time: *0900 h*

| Time: | Behaviors coded in sequence | Comments |
|---|---|---|
| 9:00:05 | A GR A | A grooms self |
| 9:00:45 | A WK | A walks |
| 9:01:00 | A AP B | A approaches B |
| 9:01:05 | A GR B | A grooms B |
| 9:03:10 | A LV B | A leaves B |
| 9:03:15 | A SI | A sits |
| 9:06:05 | A HH | A handles hay |
| 9:07:30 | A SI | |
| 9:09:10 | A AP B | |
| 9:09:15 | A GR B | |

DATA SHEET FOR RECORDING SEQUENCES (CONTINUOUS SAMPLING)

Date: *10/7/86* Species: *LTM* Subject: *A* Observer: *CMC*
Enclosure: *Indoor* Weather: *Not applicable*
Start Time: *0900 h*

| Minutes: | Behaviors coded in sequence | Comments |
|---|---|---|
| 1 | A GR A, A WK | |
| 2 | A AP B, A GR B | |
| 3 | | |
| 4 | A LV B, A SI | |
| 5 | | |
| 6 | | |
| 7 | A HH | |
| 8 | A SI | |
| 9 | | |
| 10 | A AP B, A GR B | |

Fig. 30.7. Data sheets for recording sequences of behavior using continuous sampling. (*Top*) Data sheet for recording onset time (recording the onset of behaviors is necessary for later calculation of durations of behaviors). (*Bottom*) Data sheet for recording within time intervals.

also be possible to record simple behavior categories next to the individual's identification code. The map technique is a good method to use when it is not clear from the outset of the project which location divisions might be important for analysis.

Another format for recording data is a matrix table, e.g. with columns labeled with behavior names and rows labeled with locations. Each matrix could be for a single subject for an observation of specified duration, or one matrix could be used for all animals in the enclosure if their ID codes were recorded. A matrix tally sheet could be used for scan sample data, using one tally mark per scan, or for continuous recording of frequency data (behavior by location). For recording all occurrences of one interactive behavior, a matrix could list actors as row headings and recipients as column headings; using continuous sampling, a tally mark would be made in the proper cell whenever the specified interaction occurred, e.g. supplanting (Lehner 1996).

For many projects conducted in the zoo setting, more than one type of data must be recorded. As described above, location and behavior data can be recorded at the same time using either continuous or scan sampling. However, in many cases a "mixed" sampling strategy is most appropriate. In such cases, scan data can be recorded in columns on the left side of the page and continuous data on the right (fig. 30.8). Generally, "mixed" sampling strategies record location, nearest neighbor, and general behavior category on the scan, and frequency or interaction data using continuous sampling. For example, one scan sample category might be "social behavior," whereas specific behavior, actor, and recipient would be recorded continuously. Another possibility is to observe focal subjects in random order, recording data using continuous sampling;

MIXED SAMPLING DATA SHEET WITH CONCURRENT CATEGORIES

Date: *10/8/86* Species: *Red panda* Observer: *CMC*
Enclosure: *South* Weather: *Cloudy, 55*

Start Time:
*0800 h*

| | | (Scan) | (Scan) | Scan Sample Behavior | | | | | | Continuous frequency | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Interval | Subject | LOCATION | NEAR NEIGH. | N.V. | SOC | STAT | MOVE | EAT | OTHER | GROOM | SEX | OTHER | Comments |
| 0:00 | A | *1* | *B* | | *1* | | | | | *B* | *BB* | | *2 mounts* |
| | B | *1* | *A* | | *1* | | | | | | | | |
| | C | *4* | *D* | *1* | | | | | | | | | *In den* |
| | D | *4* | *C* | *1* | | | | | | | | | *In den* |
| 0:05 | A | *2* | *B* | | | | *1* | | | | | | |
| | B | *1* | *A* | | | *1* | | | | | | | |
| | C | *3* | *B* | | | | *1* | | | | | *1* | *Climbs tree* |
| | D | *4* | *B* | *1* | | | | | | | | | *Den* |
| 0:10 | A | *2* | *D* | | *1* | | | | | | | *1* | *Bites D* |
| | B | *2* | *A* | | | *1* | | | | *A* | | *1* | *Plays w/D* |
| | C | *?* | *?* | *1* | | | | | | | | | |
| | D | *2* | *A* | | *1* | | | | | | | | |
| 0:15 | A | *1* | *B* | | | *1* | | | | | | | |
| | B | *3* | *D* | | *1* | | | | | | | | *Still play* |
| | C | *?* | *?* | *1* | | | | | | | | | |
| | D | *3* | *B* | | *1* | | | | | | | | *same bout* |

Fig. 30.8. Mixed sampling data sheet for scoring 3 concurrent scan categories as well as continuous data. Scan data are recorded at the beginning of each interval, and continuous data are recorded throughout the interval. Observation period duration for the sheet shown here is 20 minutes. NEAR NEIGH., nearest neighbor; N.V., not visible; SOC, social; STAT, stationary.

then, between focal samples, record scan data on all subjects (e.g. their locations and general activity). This method was used by Stanley and Aspey (1984).

In addition to its use in specific research projects, systematic data collection can be applied to the day-to-day management of animals. Systematic records are facilitated by using standard forms for recording information. Such forms may be a part of daily reports, or they may be designed for special events. For example, Lindburg and Robinson (1986) developed a form for systematically recording the conditions and outcome of animal introductions. Even if a PDA or laptop program is to be used, the researcher needs to think about the layout of data collection.

## DATA SHEETS AND COMPUTER ANALYSIS

When data recorded by hand are to be analyzed by a software package such as SPSS or SAS (Tabachnick and Fidell 2001), it is most appropriate for the data sheet to resemble figure 30.7, top, rather than check-sheet column formats like figures 30.6 and 30.8. This is because the computer program can use routines such as cross-tabulation to count frequencies of, e.g., coded behaviors per coded actor. Microsoft Excel has a useful feature called Pivot Table that computes cross-tabulation. New programs with more features are being released regularly, and it is worth the effort to evaluate a program's capabilities for the price before purchase. Some powerful programs are available inexpensively through site licenses to universities, such as SYSTAT version 11.0 (Wilkinson 2004). A personal favorite for the Macintosh is Data Desk (Velleman 1997, recent version 6.2 [2003]), with entering and preparing the data file in Microsoft Excel completed beforehand. Some simple statistical analyses are even built into Microsoft Excel. To view these features, select Tools, then Add-Ins, and check the Analysis ToolPak and Analysis ToolPak VBA boxes. Upon returning to Tools, a new option, called Data Analysis, should appear; it includes the ability to conduct both descriptive statistics and inferential hypothesis tests.

## REPLICATION AND INTER- AND INTRAOBSERVER RELIABILITY

The methods used in a research project should be defined clearly enough so that another researcher could use the same technique based on the written description provided in the final report or publication. Unequivocal behavior definitions are thus especially important.

An observer should be consistent in data collection from day to day (intraobserver reliability). Thus, if at all possible, preliminary data collection should be used as "practice" and either not be analyzed or be analyzed selectively (the least equivocal data being used). When more than one observer is to be used in a project, formal interobserver reliability testing is recommended. A common method involves having 2 or more persons collect data on the same subject simultaneously. The recorded data are then compared and the percentage of agreement calculated. A common calculation of agreement is

$$\% \text{ Agreement} = [\text{Agreements}/(\text{Agreements} + \text{Disagreements})] \times 100.$$

Errors can be made regarding identifications of individuals, behaviors, sequence of interaction, and so on. Depending on the methodology, reliability should be 85%–95% before a new observer's data are used in analysis.

Percentage of agreement is the easiest way to calculate reliability, but it is considered the poorest index of reliability from a statistician's point of view: it does not account for the likelihood of observers agreeing purely due to chance factors, and thus inflates the actual agreement between observers (Watkins and Pacheco 2000). On the other hand, any measure of reliability is better than none at all: observers who knew that they were being assessed showed significantly higher observer agreement scores than did uninformed observers (Hollenbeck 1978). Large projects involving many observers could use videotaped "real" sequences as a "standard" by which to measure agreement. Ideally, observers should be assessed repeatedly over time. Generally, many zoo projects are conducted by a single observer who improves in reliability over time through practice. Someone collecting data for a self-conceived, self-designed project is likely to be inherently more reliable, although the danger of observer bias—recording "predicted" behavior in ambiguous situations—may be increased. Martin and Bateson (1993), Lehner (1996), and Caro et al. (1979) discuss various factors affecting reliability and techniques for evaluating reliability.

Currently, the Kappa statistic (Cohen 1960) is the preferred measurement of interobserver reliability (Bakeman and Gottman 1997). If there are only 2 observers, it is simple to hand-calculate Kappa on the nominal categories (or number of times they both chose the same behavioral code; example adapted from Watkins and Pacheco [2000]). The 2 observers recording behavioral codes are compared by cross-tabulating one observer's recorded observations into columns and the other observer's recorded observations into rows. Sometimes the observations will be in agreement and sometimes they will not be in agreement, but we can calculate how often they are in agreement and whether that value is above chance levels.

$$\text{Kappa} = \frac{P_o - P_c}{1 - P_c}$$

$P_o$ = Observed proportion of agreement

$$= \frac{Agreements}{Agreements + Disagreements}$$

$P_c$ = Chance proportion of agreement

$$= \left(\frac{R_1 \times C_1}{N^2}\right) + \left(\frac{R_2 \times C_2}{N^2}\right) + \left(\frac{Rn \times C_n}{N^2}\right),$$

where
$R_1$ = Sum of the observations for row 1
$R_2$ = Sum of the observations for row 2
$Rn$ = Sum of the observations for the last row
$C_1$ = Sum of the observations for column 1
$C_2$ = Sum of the observations for column 2
$Cn$ = Sum of the observations for the last column

The Observer 5.0 (Noldus 2005) data coding system includes a reliability calculation, as does Systat (Wilkinson 2004) and SPSS. Online programs to calculate reliability are available; e.g. http://department.obg.cuhk/reseachsupport/Cohen_Kappa_data.asp

High observer reliability is needed only at the measurement level of analysis: if only rank orders are analyzed in statistical tests (true of most nonparametric tests, which are explained below), then observers' accuracy in recording behavior needs to be precise only at the level of rank order (Sackett, Ruppenthal, and Gluck 1978). For example, as long as the observer accurately records that male A is aggressive more often than male B, and B is aggressive more often than C, the outcome of a rank-order statistical test will not be changed if a few aggressive acts are missed.

## DATA PRESENTATION AND ANALYSIS

The purpose of this section is to introduce the reader to some considerations and techniques that are useful in the analysis of data collected in the zoo setting. It is not intended to provide all the skills needed and should be used in conjunction with the more thorough references cited. Some aspects of data analysis should be considered before a data recording method is adopted. Again, preliminary analyses are important: they may suggest a revision to the data sheet, data collection schedule, or collation protocol.

### DATA COLLATION

**General considerations and techniques.** During data collation (e.g. when the observer is totaling up a data sheet), 2 important considerations ought to be taken into account.

1. Data for each subject and/or observation session should be equivalent—based on the same amount of observation time. If observation times differ, equivalence can be achieved by converting raw scores to rates or percentages. Decide whether to use total observation time (or total number of scans) as a base, or the amount of time (or number of scans) during which the subject is visible as a base.
2. Data summaries should not be collapsed across all observation sessions until it is determined whether scores per focal sample period or some other time block will be used in statistical tests. In any event, when observation periods are not of equal length, it is often advisable for each session or day to contribute equally. Observation schedules in which each subject is observed for the same amount of time (per time block, if relevant) avoid many problems. When time "not visible" varies across subjects and observation days, this complicates analysis.

To facilitate the collating and transcribing of data from the original data sheets, some attention should be paid to the design of summary or tabulation sheets. Where possible, include summary rows on the data sheets themselves (e.g. fig. 30.6). Some tabulation sheets may be in the form of matrices. Tabulation can be facilitated by use of a spreadsheet program, such as Microsoft Excel.

**Estimates based on continuous focal-animal sampling.** When recording the interactive behaviors of a focal animal, one may decide to record all behaviors directed toward the subject, S, as well as those initiated by the subject. This method allows efficient use of observation time but requires special considerations in some data analyses. Thus, in samples in which *Si* is the focal animal and in samples in which *Sj* is the focal animal, all their interactions will be recorded. Each of the samples (*i* or *j*) or both (*i* + *j*) will give an estimate of their rate of interaction (Altmann 1974), as shown in table 30.5.

Consider the interaction data summarized in table 30.6. When the sum of observation time for subject I and subject J is used as a time base, each cell in the frequency matrix can be used to calculate a valid estimate of that dyad's hourly rate of

**TABLE 30.5.** Estimates of interaction rates

| Subject | Sample duration | Number of interactions$_{i,j}$ | Rate |
|---|---|---|---|
| i | 20 min (1/3 hr) | 5 | 15/hr |
| j | 10 min (1/6 hr) | 3 | 18/hr |
| i + j | 30 min (1/2 hr) | 8 | 16/hr |

**TABLE 30.6.** Social grooming interactions for subjects I, J, and K

| Sample duration | Focal subject | Interaction | Frequency |
|---|---|---|---|
| 60 min | I | I grooms J | 5 |
| | | J grooms I | 3 |
| 60 min | J | J grooms I | 6 |
| | | I grooms J | 4 |
| 60 min | K | K grooms I | 5 |
| | | I grooms K | 1 |
| 180 min = 3 hr | | | 24 |

Frequency matrix

| | | Groomee I | Groomee J | Groomee K | Total |
|---|---|---|---|---|---|
| Groomer | I | | 9 | 1 | 10 |
| Groomer | J | 9 | | 0 | 9 |
| Groomer | K | 5 | 0 | | 5 |
| | Total | 14 | 9 | 1 | 24 |

Hourly rate

| | | Groomee I | Groomee J | Groomee K | Total |
|---|---|---|---|---|---|
| Groomer | I | | 4.5 | .5 | 5.0 |
| Groomer | J | 4.5 | | 0 | 4.5 |
| Groomer | K | 2.5 | 0 | | 2.5 |
| | Mean grooming rate per individual: | | | | 4.0 |

interaction. In this example, subject I was observed to groom J a total of 9 times while they were focal subjects, and I groomed K once while K was the focal subject, totaling 10 grooms by I. Although I, J, and K were each focal animals for one hour of observation, one cannot divide 10 grooms by 3 hours to yield a grooming rate of 3.3 for I, because focal sampling of J does not reveal interactions between I and K (e.g. during the hour that J was the focal subject, I could have groomed K 5 times). To calculate a mean rate per individual, rates per dyad must be calculated, then summed and divided by the number of individuals. See Michener (1980) and Shapiro and Altham (1978) for other considerations in estimating interaction rates.

**The problem of visibility.** When estimates of behavioral rates or percentages are based only on the duration of the sample when the subject is visible, such as done by Ralls, Kranz, and Lundrigan (1986), it is important to consider that the animal's behavior when visible may not be a random sample of total behavior. The animal may be performing the same behaviors at different rates or may be engaging in different behaviors when out of sight. Many zoo enclosures have indoor and outdoor sections. The observer should sample both sections before concluding that behavior inside is the same as (or different from) behavior outside. If behavior is the same inside and out, then rates can be calculated using time observable as the divisor. In other situations, a subject may be unobservable because it has entered a den or nest box, where perhaps only a few behaviors are likely to occur. In such cases, total sample time should probably be the divisor, and "in den" should be considered a behavior. Similarly, some animals in naturalistic enclosures may be scored as "not visible" primarily when they are lying down, concealed by tall vegetation; in this case using observation time while "visible" as the divisor would overestimate the actual percentage of "active" behavior. The results of such a study might therefore include a category for "percentage of time not visible," which would be combined with "percentage of time inactive" for some analyses. If a large percentage of observation time occurs when the subjects are out of sight, results should be interpreted with this consideration in mind.

## STATISTICAL TESTS

All behavioral research projects will involve some descriptive statistics (e.g. table 30.4). Behavioral researchers should also use statistical tests in order to test hypotheses and draw conclusions (Lehner 1996). Otherwise, the conclusions may be unjustified. The purpose of statistical tests is to "determine how large the observed differences must be before we can have confidence that they represent real differences in the larger group from which only a few events were sampled" (Siegel 1956, 2). Statistical tests are posed in such a manner that, given a large enough difference, the null hypothesis can be rejected. For example, a null hypothesis might be that the means (averages) of 2 samples, such as mean aggression rates in 2 enclosures, do not differ. Rejection of the null hypothesis suggests that the 2 sample means are statistically significantly different.

If the results of a research project are to be applied to management decisions in a zoo or aquarium, it is doubly important that the conclusions of the study have some statistical basis. However, statistical significance alone should not dictate decisions, because the magnitude of the effect, the "effect size," is really more important (Martin and Bateson 1993). Even if expensive enclosure modifications resulted in statistically significantly reduced aggression, they might not be worth applying throughout the zoo if the behavior change was small and no reduction in injuries could be demonstrated. On the other hand, behavior might be altered dramatically in some individuals but not in others, resulting in marginal statistical significance but a large average-effect size.

"Significant" differences usually cannot be eyeballed from graphed data unless error (variability) measures are included. When graphing and comparing means, it is appropriate to use the standard deviation of the mean, which is commonly called the standard error (SE) or standard error of the mean (SEM). The notation for the standard error of the mean is $\sigma_n$, where $\sigma$ is the standard deviation of the scores and $n$ is the sample size.

$$\sigma_n = \frac{\sigma}{\sqrt{n}}$$

To show significant differences that can be seen from the graphed means, simply graph the means for each group $\pm 2$ SE (Streiner 1996). Descriptive statistics (mean or median) should always include range and/or standard error or standard deviation, and sample sizes.

**Parametric versus nonparametric tests.** Parametric statistics are based on assumptions about "parameters," such as the mean (average) and variability measures (variance or its square root, the standard deviation), that describe the "population" from which the sample data have been selected. These parameters define mathematical distributions such as "the normal distribution" on which statistical equations for particular tests are based. Nonparametric tests are "distribution free" and do not require many assumptions about the "population" from which the data were drawn (Lehner 1996).

The beginning statistician should learn which statistical tests are appropriate for which comparisons or kinds of data. Gradually expand the statistical repertoire with experience. Learning about statistics is much like becoming fluent in a foreign language—familiarity comes with use. Siegel (1956) and Conover (1999) describe most nonparametric tests in detail, and Lehner (1996) provides an adequate and usable summary of the most common ones. Furthermore, Lehner (1996) uses examples that are more relevant to zoo studies (also see Brown and Downhower [1988]). Some readers may be unfamiliar with some of the statistical terminology used in this chapter. The textbook by Ha and Ha (forthcoming) is a good general introduction to descriptive, parametric, and nonparametric statistics. Some advanced statistics books emphasize biological examples (Sokal and Rohlf 1995; Zar 1999). Tabachnick and Fidell (2001) describe multivariate statistics and computer programs that calculate them. Manuals to statistical software packages can be particularly helpful in improving understanding of statistics and data analysis (Velleman 1997; Wilkinson 2004).

Table 30.7 lists a variety of nonparametric tests. Most can be done rather easily by calculator or formulas entered into an Excel spreadsheet. To become familiar with these tests, it can be useful to look at published research and see which tests were used in which situations. Try to determine what the unit of analysis was, or exactly how the data might have been set up to do the test. Be warned, however, that inappropriately applied statistics sometimes do get published.

Parametric tests (table 30.8) can be used if certain assumptions, such as homogeneity of variance and a normal distri-

**TABLE 30.7.** Summary of common nonparametric tests

| Type of data | Statistical test | Examples of use |
|---|---|---|
| Nominal—frequency | Chi-square (association and goodness-of-fit) | Byers 1977; Izard and Simons 1986; Margulis, Hoyos, and Anderson 2003; Ralls, Brugger, and Ballou 1979 |
| | G-test (multiway contingency) | Crockett and Sekulic 1984 |
| | Binomial | Izard and Simons 1986 |
| Ordinal—rank order | | |
| Two samples | | |
| Independent | Mann-Whitney *U* | Byers 1977; Freeman 1983; Kleiman 1983; Macedonia 1987; Vickery and Mason 2004 |
| Correlated (paired) | Wilcoxon signed ranks | Byers 1977; Freeman 1983; Kleiman 1980, 1983; Mallapur and Chellam 2002 |
| | Sign test | Ralls, Brugger, and Ballou 1979 |
| | Spearman's correlation | Freeman 1983; Macedonia 1987; Margulis, Hoyos, and Anderson 2003 |
| Three or more samples | | |
| Independent | Kruskal-Wallis one-way ANOVA | Margulis, Whitham, and Ogorzalek 2003; Vickery and Mason 2004 |
| Correlated | Friedman two-way ANOVA | Nash and Chilton 1986 |

*Note:* Conover (1999), Siegel (1956), Lehner (1996), Zar (1999), and Sokal and Rohlf (1995) may be consulted for details and more tests.

**TABLE 30.8.** Choosing the appropriate parametric test

| No. of groups OR conditions | Type of design* | Assumptions (see numbered text) | Type of test to use (Ha and Ha, forthcoming) |
|---|---|---|---|
| One sample | Single sample | 1, 2, 4, and 5 are all met | Single sample *z*-test |
| One sample | Single sample | 1, 2, and 4 are all met | Single sample *t*-test |
| 2 | Independent (between) groups | 1, 2, and 3 are met | Independent *t*-test |
| 2 | Dependent (within) groups | 1 and 2 are both met | Paired *t*-test (correlated *t*-test) |
| 3 or more | Independent (between) groups | 1, 2, and 3 are met | ANOVA |

*Note:* Assumptions are as follows.

1. The data must be interval or ratio.
2. The data are normally distributed, meaning (a) the population raw scores are known to be normally distributed, or (b) the sample size is $\geq 30$, or (c) the skewness and kurtosis values are approximately between $-1.0$ and $+1.0$.
3. The variances are equal between the groups, called homogeneity of variance (HOV). The variances can be up to 4 times different from each other, but no more than that, and still be considered "equal." To find HOV, divide the larger variance by the smaller variance.
4. Known population mean
5. Known population standard deviation

*A single sample test compares a sample to known population data. This might be useful if there are verified data on wild populations and you wish to compare that mean to your sample mean. A within-groups design is one in which the same subjects are measured more than once (e.g. before, during, and after for some dependent variable), and thus participate in the study as their own control. Alternatively, within-groups designs can also be pairs of associated individuals that are being compared. In other words, within-groups designs are appropriate when you cannot assume that the data are independent. In contrast, independent, or between-groups, designs are appropriate when comparing samples that are not associated by repeated measures or relatedness (Woodland Park Zoo elephant feeding behavior versus Point Defiance Zoo and Aquarium elephant feeding behavior).

bution, are met (Ha and Ha, forthcoming). It is important to recognize that both of these assumptions are robust for minor violations of the assumption (Kirk 1994; Ha and Ha, forthcoming). Parametric tests are preferable to nonparametric tests, because they have a much greater "power"; i.e. smaller differences are required to reject the null hypothesis. Power also increases as the sample size increases: for a given magnitude of difference (e.g. between 2 means), the difference is more likely to be statistically significant when the means are based on more individual data points. In some cases, a parametric test is necessary for multivariate analysis, or when unequal sample sizes make use of the Friedman ANOVA inappropriate (Lehner 1996, and table 30.7). Parametric tests can be conducted using one of the numerous statistical packages on the market (e.g. Microsoft Excel, Minitab, SPSS, STATA, Systat, Data Desk).

Whenever percentages or proportions are to be used in parametric statistics, it is recommended that the data first be arcsine-transformed to normalize the distribution (Lehner 1996, p. 378). This transformation was used by Stanley and Aspey (1984). Transformations are useful in correcting some violations of parametric assumptions, and advanced readers should consult Lehner (1996) or Zar (1999) for information on square root and logarithmic transformations.

While nonparametric tests are one alternative when the assumptions of parametric tests are not met, the reduction in power due to rank transformations is a significant disadvantage. Resampling, or randomization, tests are increasingly being used as a more powerful alternative to nonparametric tests (Adams and Anthony 1996). These tests generate probabilities based on empirical repeated sampling (resampling) of the raw data to create a randomization distribution (Hayes 2000). This technique is particularly useful when the assumption of a normal distribution is not met, but the assumption of approximately equal variances is met (ibid.). See the reviews by Adams and Anthony (1996) and Crowley (1992) for more information on the different techniques and software to derive randomization distributions. These techniques may be particularly useful when one's data are repeated samples of the same individual, a common occurrence in zoo research (e.g. Cantoni 1993).

**The unit of analysis.** To perform statistical tests, one has to decide on the unit of analysis. In experimental studies, this is usually obvious, e.g. the number of trials before a rat learns a task. In studies of observed behaviors in which the researcher defines the behaviors, the issue is more complicated. The unit of analysis might be the total number of occurrences (frequency) of a behavior, its hourly rate of occurrence, the percentage of time spent performing the behavior, the total duration of the behavior, or mean bout duration. Furthermore, the researcher must determine whether each animal's overall "score" (total frequency, mean rate, duration, or whatever) will be a data point, or whether each animal will contribute one score per observation period or designated time block (e.g. age) and thus the data points are not independent. Perhaps individuals cannot be distinguished, and each observation period contributes one score that is the average or total of all individuals. The appropriate unit of analysis will depend in part on the statistical test to be used.

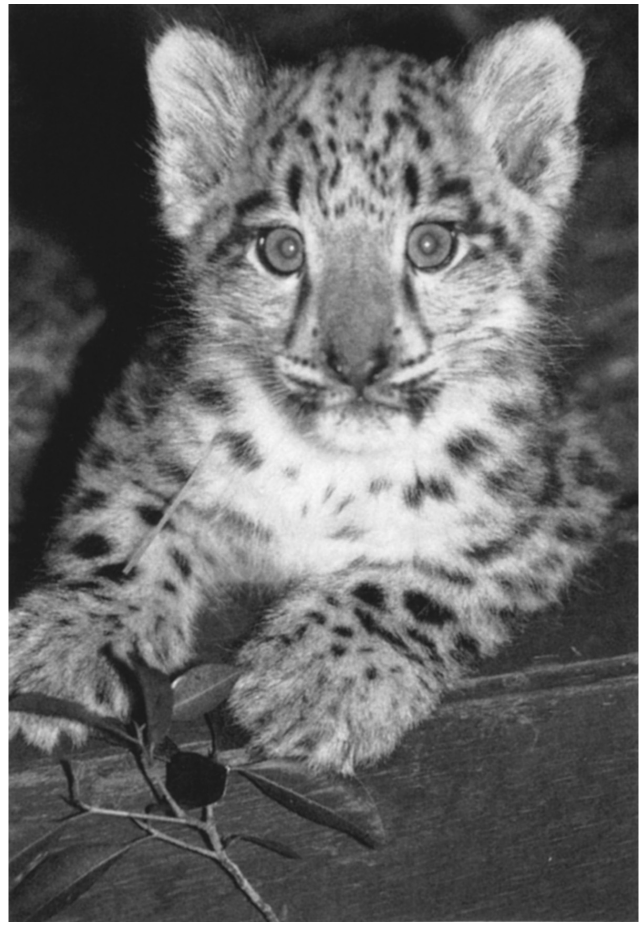

Fig. 30.9. The Mann-Whitney *U* nonparametric statistic was used to test for behavioral differences between successfully and unsuccessfully breeding snow leopard pairs. Boris is the cub of a successful pair. (Photography by Cathy Shelton, Woodland Park Zoo. Reprinted by permission.)

For some statistical tests, minimum sample sizes are required in order to demonstrate significance (Siegel 1956). Freeman (1983) used the Mann-Whitney *U* to test differences between successfully and unsuccessfully breeding snow leopard, *Uncia uncia*, pairs, analyzing data for each sex separately (fig. 30.9). For the sample sizes in that study (3 successful and 5 unsuccessful pairs), in order to achieve a 2-tailed level of significance (at a probability of 0.05 or less), there could be no reversals. In other words, significant differences could be demonstrated only if all 3 successful pairs ranked above (or below) the 5 unsuccessful pairs.

Many studies of captive animals involve small groups, in some cases too few individuals to use one data point per subject for some kinds of statistical tests. In such cases, the sample size (and statistical power) can be increased by using one score per subject per observation period or time block. These data could be used in a repeated-measures design, or in multiple tests of the null hypothesis that an individual's behavior (as opposed to the group's behavior) did not vary from one condition to another (e.g. after moving to a new enclosure). This is also a situation where the new randomization techniques dis-

Fig. 30.10. *Galago senegalensis* in the Nocturnal House at Woodland Park Zoo, Seattle. (Photography by Karen Anderson. Reprinted by permission.)

cussed earlier apply. One cannot simply lump multiple scores from one individual with those of others without the possibility of committing a type I error—i.e. rejecting the null hypothesis when it is in fact true (Machlis, Dodd, and Fentress 1985, "the pooling fallacy"). Such an error can occur when within-individual variance (i.e. between observations of the same animal) is less than between-individual variance (Leger and Didrichsons 1994). Some ways to avoid this problem while maximizing statistical power include (1) using a more complex design (e.g. a repeated measures test), (2) examining the sources of variance in detail, using the results to determine the grouping into units of analysis (Kraemer et al. 1982; Thiemann and Kraemer 1984), (3) using the mean or sum across all individuals within a basic observation period so that each individual and observation period contribute equally, and (4) testing each subject's data separately, which might be done if each individual's response to a change was of interest.

Some examples from the literature illustrate different units of analysis. Byers (1977, figures 4 and 5) used a Wilcoxon matched pairs test to determine whether play events occurred at different rates on different substrates. For example, for each individual ibex, *Capra sibirica*, kid, the total number of "butt" play events that occurred on sloped surfaces was paired (matched) with the same kid's total number of "butt" play events on flat surfaces. Sloped and flat areas each made up about half the enclosure; otherwise the play events per S would have been multiplied by the proportion of the enclosure made up of the surface type on which they occurred to correct for differences in "available" area. To compare sex differences in behavior, Freeman (1983) matched male and female percentages of time spent in selected behaviors (calculated from scan samples) for members of 8 mated snow leopard pairs. Since pairs were studied for different numbers of years, "cat-mean" data (mean percentage per leopard across years of study) were used in statistical tests. In Kleiman's (1980) figure 50.4, the total amount of time that the sexually active male golden lion tamarin, *Leontopithecus rosalia*, spent grooming the female was matched with the total duration of grooming by the sexually inactive male for each observation period. Thus, each observation session contributed one score per male, and the data from each trio (2 males and one female housed together) were statistically tested separately. In Nash and Chilton's (1986) study, each galago, *Galago senegalensis* (fig. 30.10), was observed for the same amount of time for each of 3 "phases," except that infants' observation sessions were twice as long. The analyzed data for each behavior scored consisted of total frequency per individual per phase (i.e. "raw" scores), except for infants, whose frequencies were halved, i.e. "corrected" or "adjusted," to make them equivalent. Alternatively, raw frequencies could have been converted to hourly rates. In a longitudinal study of chimpanzee, *Pan troglodytes*, development, all observations—made 3 days a week—over a 3-month seasonal period for a single subject were combined into a single data point for analysis (Kraemer et al. 1982).

## OTHER STATISTICAL CONSIDERATIONS

**The problem of independence.** Theoretically, for purposes of statistical analysis, data points (e.g. the units of analysis described above) should be independent. For example, one individual's rate of performing a given behavior should be unrelated to another individual's rate, or the occurrence of one behavior type should not influence the probability of occurrence of another. In reality, the independence assumption is often violated in the case of interactive social behaviors (most zoo studies), which usually influence the behavior of other group members and thus may be inherently correlated (G. P. Sackett, personal communication). Furthermore, when more than one of a mutually exclusive and exhaustive set of behaviors is tested, the outcome of one statistical test is not independent of the outcome of the other: if behaviors are categorized as either "social" or "nonsocial," rejecting the null hypothesis that social behaviors did not differ between conditions guarantees that the difference in nonsocial behaviors will also be statistically significant (Sackett, Ruppenthal, and Gluck 1978). For this reason, adjustments to probability levels are sometimes applied to make tests more conservative (Stanley and Aspey 1984). Fortunately, new techniques are quickly being developed to eliminate this problem. Advanced readers should explore the topics of Monte Carlo Simulations, Modeling, and Resampling Techniques for more information on how to deal with violations of the assumption of independence (Crowley 1992; Todman and Dugard 2000).

**Temporal autocorrelation.** Another aspect of independence is temporal autocorrelation, or the probability that the occurrence of a behavior at one point in time will affect its likelihood of being observed at the next point in time. Obviously, the shorter the time interval between successive "points," the more likely that temporal autocorrelation will occur. For scan or instantaneous samples that are converted to percentages, this poses no problem; shorter intervals generally produce more accurate estimates of true percentages of time spent performing the behavior in question. However, contingency analyses (chi-square, goodness-of-fit tests) require independent data points (Siegel 1956). If, for example, one wanted to compare the use of several different enclosure locations, one possibility would be to count the number of times that the subject was scored in each location. However, these counts could not be used in a chi-square test if the points in time were temporally autocorrelated—that is, if the animal's location on a particular branch was not independent of the fact that it was found there in the previous interval.

The interval at which independence can be assumed varies with behavior, species, and so forth, so no general rule can be stated; the appropriate interval must be determined from the data. For example, Janson (1984) found that nearest neighbors of wild brown capuchins, *Cebus apella*, usually were temporally autocorrelated at 5-minute intervals, rarely were at 10-minute intervals, and never were at 15-minute intervals. Thus, only records at 15-minute intervals were used for analyses requiring independence. A pilot study using continuous sampling could be used to choose the appropriate scan interval. In this manner, Slatkin (1975) computed the autocorrelation time for adult male geladas, *Theropithecus gelada*, and yellow baboons, *Papio cynocephalus*, and found the correlation time to be about one minute for the geladas and 4–5 minutes for the yellow baboons.

Ketchum (1985) studied enclosure utilization by snow leopards at Woodland Park Zoo, Seattle. Scan samples were taken every 20 seconds, an interval likely to be highly autocorrelated. The enclosure was divided into 4 location categories (based on visibility to the public and distance that the cats could visualize), and the percentage of scan samples spent in each area was calculated. To analyze these data with a chi-square goodness-of-fit test, which requires independence as well as frequency (i.e. not percentage) data, the percentages were multiplied by the number of focal sample periods. This calculation produces adjusted frequencies approximately equivalent to randomly sampling the location of the subject once per period. Since the sample periods were at least 2 hours apart, and often more than a day apart, these adjusted frequencies were accepted as independent. The expected frequencies were calculated by multiplying the number of sample periods by the percentage of the enclosure area that each location category constituted. (Expected frequencies in this test are the values that we would "expect" if the snow leopards were using the locations in proportion to their availability, i.e. showing no preference.)

Lehner (1996) describes a test for comparing 2 percentages; however, if this test is used on scan sample data, the scan intervals must not be temporally autocorrelated. If there is reason to believe that they are, a simple but statistically conservative solution is to use the number of observation periods as $n$ in the equation.

The logic of the independence requirement is simple: Recall that the power of the statistical test improves with sample size. Obviously, the closer the scan samples, the more samples there will be in a given observation period. An inflated sample size will increase the likelihood of refuting the null hypothesis (and committing a type I error), and scan sample interval length will be inversely related to achieving statistical significance. Clearly, it is not valid to pick a sampling interval that would guarantee significance. On the other hand, using the technique of multiplying percentages by the number of observation periods makes the test unnecessarily conservative when the true interval of independence is less than the sample duration.

Whenever each focal sample period contributes a data point, the underlying assumption is that each session is an independent estimate of the animal's behavior. This further stresses the importance of scheduling balanced or randomized observation periods so as not to introduce systematic bias.

The violation of the independence assumption restricts the number of conventional statistical tests that can be applied to certain behavioral data. Dunbar and Dunbar (1975) describe some considerations and solutions with respect to the independence assumption. Also, see the section on randomization tests, mentioned previously.

## CONCLUSION

Data collection in the zoo setting can provide answers to management questions as well as basic information about the biology of captive animals. Research is now being recognized as important and is expanding in many zoos (Finlay and Maple 1986; Leong, Terrell, and Savage 2004; Maple and Bashaw, chap. 24, this volume). For example, the benefits of environmental enrichment are being evaluated (Mellen and MacPhee 2001; Mellen and MacPhee, chap. 26, this volume; Young 2003). To be most useful, data should be quantified in a manner amenable to statistical analysis, whether it is statistical testing or straightforward description. Furthermore, proper sampling methods should be used so as to avoid observer bias and other sorts of sampling error. This chapter has summarized the major sampling methods and has provided some hints for data analysis.

Systematic data collection is not difficult and mostly requires systematic thinking ahead of time. A project is more likely to be successful if these guidelines are followed:

1. Formulate a specific research question.
2. Keep data collection simple.
3. Perform preliminary analyses on some sample data before finalizing the data collection design.
4. Collate and begin to analyze data while data collection is in progress.
5. Finally, if the results of the study seem to be of general interest, publish them.

## ACKNOWLEDGMENTS

The 1996 version of this chapter was based on the workshop "Applying Behavioral Research to Zoo Animal Management," Woodland Park Zoo (WPZ), Seattle, 1986, partially funded by a Conservation Grant from the Institute of Museum Services to WPZ. C. Kline assisted with the literature search; M. Hutchins had the inspiration to develop the workshop, with assistance from W. Karesh and CMC. The 1996 chapter benefited from comments from J. Altmann, S. Lumpkin, and R. Baldwin, and conversations with G. Sackett and C. Janson. We thank D. Kleiman and K. Thompson for the opportunity to update this chapter, and appreciate the constructive comments of 2 reviewers. The current version incorporates suggestions from students in RRH's psychology class, Behavioral Studies of Zoo Animals, and J. C. Ha. The National Primate Research Center at the University of Washington provided support to CMC (NIH RR00166). RRH was supported by the Department of Psychology, University of Washington.

## REFERENCES

Adams, D. C., and Anthony, C. D. 1996. Using randomization techniques to analyse behavioural data. *Anim. Behav.* 51:733–38.

Altmann, J. 1974. Observational study of behavior: Sampling methods. *Behaviour* 49:227–67.

———. 1984. Observational sampling methods for insect behavioral ecology. *Fla. Entomol.* 67 (1): 50–56.

Astley, C. A., Smith, O. A., Ray, R. D., Golanov, E. V., Chesney, M. A., Chalyan, V. G., Taylor, D. J., and Bowden, D. M. 1991. Integrating behavior and cardiovascular responses: The code. *Am. J. Physiol.* 261:R172–R181.

BAG (Behavioral Advisory Group, American Zoo and Aquarium Association). 2002. Ethograms, ethograms.org/. Silver Spring, MD: American Zoo and Aquarium Association; Chicago: Lincoln Park Zoo.

Bakeman, R., and Gottman, J. M. 1997. *Observing interaction: An introduction to sequential analysis*. New York: Cambridge University Press.

Bernstein, I. S. 1991. An empirical comparison of focal and ad libitum scoring with commentary on instantaneous scans, all occurrence and one-zero techniques. *Anim. Behav.* 42:721–28.

Blumstein, D. T., Evans, C. S., and Daniel, J. C. 2000. JWatcher. www.jwatcher.ucla.edu/. Los Angeles: UCLA; Sydney: Macquarie University.

Bobbitt, R. A., Jensen, G. D., and Gordon, B. N. 1964. Behavioral elements (taxonomy) for observing mother-infant-peer interaction in *Macaca nemestrina*. *Primates* 5:71–80.

Brannian, J., and Cloak, C. 1985. Observations of daily activity patterns in two captive short-nosed echidnas, *Tachyglossus aculeatus*. *Zoo Biol.* 4:75–81.

Brown, L., and Downhower, J. F. 1988. *Analyses in behavioral ecology: A manual for lab and field*. Sunderland, MA: Sinauer Associates.

Byers, J. A. 1977. Terrain preferences in the play behavior of Siberian ibex kids (*Capra ibex sibirica*). *Z. Tierpsychol.* 45:199–209.

Cantoni, D. 1993. Social and spatial organization of free-ranging shrews, *Sorex coronatus* and *Neomys fodiens* (Insectivora, Mammalia). *Anim. Behav.* 45:975–95.

Caro, T. M., Roper, R., Young, M., and Dank, G. R. 1979. Interobserver reliability. *Behaviour* 69:303–15.

Chow, I. A., and Rosenblum, L. A. 1977. A statistical investigation of the time-sampling methods in studying primate behavior. *Primates* 18:555–63.

Clutton-Brock, T. H. 1977. Appendix I: Methodology and measurement. In *Primate ecology*, ed. T. H. Clutton-Brock, 585–90. London: Academic Press.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20 (1):37–46.

Conover, W. J. 1999. *Practical nonparametric statistics*. 3rd ed. New York: Wiley.

Crockett, C., and Hutchins, M., eds. 1978. *Applied behavioral research at the Woodland Park Zoological Gardens*. Seattle: Pika Press.

Crockett, C. M., and Sekulic, R. 1984. Infanticide in red howler monkeys (*Alouatta seniculus*). In *Infanticide: Comparative and evolutionary perspectives*, ed. G. Hausfater and S. B. Hrdy, 173–91. New York: Aldine.

Cronin, G. M., Dunshea, F. R., Butler, K. L., McCauley, I., Barnett, J. L., and Hemsworth, P. H. 2003. The effects of immuno- and surgical-castration on the behaviour and consequently growth of group-housed, male finisher pigs. *Appl. Anim. Behav. Sci.* 81 (2): 111–26.

Crowley, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annu. Rev. Ecol. Syst.* 23: 405–47.

Dunbar, R. I. M. 1976. Some aspects of research design and their implications in the observational study of behavior. *Behaviour* 58:78–98.

Dunbar, R. I. M., and Dunbar, P. 1975. *Social dynamics of gelada baboons*. Basel: Karger.

Finlay, T. W., and Maple, T. L. 1986. A survey of research in American zoos and aquariums. *Zoo Biol.* 5:261–68.

Forney, K. A., Leete, A. J., and Lindburg, D. G. 1991. A bar code scoring system for behavioral research. *Am. J. Primatol.* 23:127–35.

Freeman, H. 1983. Behavior in adult pairs of captive snow leopards (*Panthera uncia*). *Zoo Biol.* 2:1–22.

Grasso, M. A., and Grasso, C. T. 1994. Feasibility study of voice-driven data collection in animal drug toxicology studies. *Comput. Biol. Med.* 24:289–94.

Ha, J. C. 1991. *EVENT-PC and EVENT-Mac Software*. Seattle: University of Washington Regional Primate Research Center.

Ha, R. R., and Ha, J. C. 2003. Effects of prey type, prey density and energy requirements on the use of alternative foraging tactics in crows. *Anim. Behav.* 66:309–16.

———. Forthcoming. *Integrated statistics for behavioral science*. Thousand Oaks, CA: Sage Publications.

Hayes, A. F. 2000. Randomization tests and the equality of variance assumption when comparing group means. *Anim. Behav.* 59:653–56.

Heymann, E. W., and Smith, A. C. 1999. When to feed on gums: Temporal patterns of gummivory in wild tamarins, *Saguinus mystax* and *Saguinus fuscicollis* (Callitrichinae). *Zoo Biol.* 18: 459–71.

Hinde, R. A. 1973. On the design of check sheets. *Primates* 14:393–406.

Hollenbeck, A. R. 1978. Problems of reliability in observational research. In *Observing behavior*, vol. 2: *Data collection and analysis methods*, ed. G. P. Sackett, 79–98. Baltimore: University Park Press.

Izard, M. K., and Simons, E. L. 1986. Isolation of females prior to parturition reduces neonatal mortality in *Galago*. *Am. J. Primatol.* 10:249–55.

Janson, C. H. 1984. Female choice and mating system of the brown capuchin monkey *Cebus apella* (Primates: Cebidae). *Z. Tierpsychol.* 65:177–200.

Kawata, K., and Elsen, K. M. 1984. Growth and feeding relationships of a hand-reared lowland gorilla infant (*Gorilla g. gorilla*). *Zoo Biol.* 3:151–57.

Ketchum, M. H. 1985. Activity patterns and enclosure utilization in the snow leopard, *Panthera uncia*. Master's thesis, University of Washington.

Kirk, R. E. 1994. *Experimental design: Procedures for the behavioral sciences*. 3rd ed. Belmont, CA: Brooks/Cole.

Kirkevold, B. C., and Crockett, C. M. 1987. Behavioral development and proximity patterns in captive DeBrazza's monkeys. In *Comparative behavior of African monkeys*, ed. E. L. Zucker, 39–65. New York: A. R. Liss.

Kleiman, D. G. 1974. Activity rhythms in the giant panda. *Int. Zoo Yearb.* 14:165–69.

———. 1980. The sociobiology of captive propagation. In *Conservation biology: An evolutionary-ecological approach*, ed. M.E. Soulé and B.A. Wilcox, 243–61. Sunderland, MA: Sinauer Associates.

———. 1983. Ethology and reproduction of captive giant pandas (*Ailuropoda melanoleuca*). *Z. Tierpsychol.* 62:1–46.

Kraemer, H. C. 1979. One-zero sampling in the study of primate behavior. *Primates* 20:237–44.

Kraemer, H. C., Alexander, B., Clark, C., Busse, C., and Riss, D. 1977. Empirical choice of sampling procedures for optimal research design in the longitudinal study of primate behavior. *Primates* 18:825–33.

Kraemer, H. C., Horvat, J. R., Doering, C., and McGinnis, P. R. 1982. Male chimpanzee development focusing on adolescence: Integration of behavioral with physiological changes. *Primates* 23 (3): 393–405.

Leger, D. W. 1977. An empirical evaluation of instantaneous and one-zero sampling of chimpanzee behavior. *Primates* 18:387–93.

Leger, D. W., and Didrichsons, I. A. 1994. An assessment of data pooling and some alternatives. *Anim. Behav.* 48 (4): 823–32.

Lehner, P. N. 1996. *Handbook of ethological methods*. Cambridge: Cambridge University Press.

Leong, K. M., Terrell, S. P., and Savage, A. S. 2004. Causes of mortality in captive cotton-top tamarins (*Saguinus oedipus*). *Zoo Biol.* 23:127–37.

Lindburg, D. G. 1990. Proceptive calling by female lion-tailed macaques. *Zoo Biol.* 9:437–46.

Lindburg, D. G., and Robinson, P. 1986. Animal introductions: Some suggestions for easing the trauma. *Anim. Keep. Forum* (January): 8–11.

Little, K. A., and Sommer, V. 2002. Change of enclosure in langur monkeys: Implications for the evaluation of environmental enrichment. *Zoo Biol.* 21:549–59.

Lorenz, K. 1958. The evolution of behavior. *Sci. Am.* 199 (December): 67–74.

Macedonia, J. M. 1987. Effects of housing differences upon activity budgets in captive sifakas (*Propithecus verreauxi*). *Zoo Biol.* 6:55–67.

Machlis, L., Dodd, P. W. D., and Fentress, J. C. 1985. The pooling fallacy: Problems arising when individuals contribute more than one observation to the data set. *Z. Tierpsychol.* 68:201–14.

Maestripieri, D. 1996. Gestural communication and its cognitive implications in pigtail macaques (*Macaca nemestrina*). *Behaviour* 133:997–1022.

Mahler, A. E. 1984. Activity budgets and use of space by South American tapir (*Tapiris terrestris*) in a zoological park setting. *Zoo Biol.* 3:35–46.

Mallapur, A., and Chellam, R. 2002. Environmental influences on stereotypy and the activity budget of Indian leopards (*Panthera pardus*) in four zoos in southern India. *Zoo Biol.* 21:585–95.

Margulis, S. W., Hoyos, C., and Anderson, M. 2003. Effect of felid activity on zoo visitor interest. *Zoo Biol.* 22:587–99.

Margulis, S. W., Nabong, M., Alaks, G., Walsh, A., and Lacy, R. C. 2005. Effects of early experience on subsequent parental behaviour and reproductive success in oldfield mice, *Peromyscus polionotus*. *Anim. Behav.* 69:627–34.

Margulis, S. W., Whitham, J. C., and Ogorzalek, K. 2003. Silverback presence and group stability in gorillas (*Gorilla gorilla gorilla*). *Folia Primatol.* 74:92–96.

Martin, P., and Bateson, P. 1993. *Measuring behaviour: An introductory guide*. 2nd ed. Cambridge: Cambridge University Press.

Meder, A. 1986. Physical and activity changes associated with pregnancy in captive lowland gorillas (*Gorilla gorilla gorilla*). *Am. J. Primatol.* 11:111–16.

Mellen, J., and MacPhee, M. S. 2001. Philosophy of environmental enrichment: Past, present, and future. *Zoo Biol.* 20 (3): 211–26.

Merritt, K., and King, N. E. 1987. Behavioral sex differences and activity patterns of captive humboldt penguins (*Spheniscus humboldti*). *Zoo Biol.* 6:129–38.

Michener, G. R. 1980. The measurement and interpretation of interaction rates: An example with adult Richardson's ground squirrels. *Biol. Behav.* 5:371–84.

Nash, L. T., and Chilton, S.-M. 1986. Space or novelty? Effects of altered cage size on *Galago* behavior. *Am. J. Primatol.* 10:37–49.

Noldus, L. P. J. J. 1991. The Observer: A software system for collection and analysis of observational data. *Behav. Res. Methods Instrum. Comput.* 23 (3): 415–29.

———. 2005. *The Observer*. Version 5.0. Wageningen, The Netherlands: Noldus Information Technology.

Paterson, J. D. 2001. *Primate behavior: An exercise workbook*. 2nd ed. Prospect Heights, IL: Waveland Press.

Paterson, J. D., Kubicek, P., and Tillekeratne, S. 1994. Computer data recording and DATAC 6, a BASIC program for continuous and interval sampling studies. *Int. J. Primatol.* 15 (2): 303–15.

Penfold, L. M., Ball, R., Burden, I., Jochle, W., Citino, S. B., Monfort, S. L., and Wielebnowski, N. 2002. Case studies in antelope aggression control using a GnRH agonist. *Zoo Biol.* 21:435–48.

Price, E. O., and Stokes, A. W. 1975. *Animal behavior in laboratory and field*. San Francisco: Freeman.

Ralls, K., Brugger, K., and Ballou, J. 1979. Inbreeding and juvenile mortality in small populations of ungulates. *Science* 206: 1101–3.

Ralls, K., Kranz, K., and Lundrigan, B. 1986. Mother-young relationships in captive ungulates: Variability and clustering. *Anim. Behav.* 34:134–45.

Rhine, R. J., and Ender, P. B. 1983. Comparability of methods used in the sampling of primate behavior. *Am. J. Primatol.* 5:1–15.

Robeck, T. R., Monfort, S. L., Calle, P. P., Dunn, J. L., Jensen, E., Boehm, J. R., Young, S., and Clark, S. T. 2005. Reproduction, growth and development in captive Beluga (*Delphinapterus leucas*). *Zoo Biol.* 24:29–49.

Sackett, G. P. 1978a. Measurement in observational research. In *Observing behavior*, vol. 2: *Data collection and analysis methods*, ed. G. P. Sackett, 25–43. Baltimore: University Park Press.

———, ed. 1978b. *Observing behavior*, vol. 2: *Data collection and analysis methods*. Baltimore: University Park Press.

Sackett, G. P., Ruppenthal, G. C., and Gluck, J. 1978. Introduction: An overview of methodological and statistical problems in observational research. In *Observing behavior*, vol. 2: *Data collection and analysis methods*, ed. G. P. Sackett, 1–14. Baltimore: University Park Press.

Sackett, G. P., Stephenson, E., and Ruppenthal, G. C. 1973. Digital data acquisition systems for observing behavior in laboratory and field settings. *Behav. Res. Methods Instrum. Comput.* 5 (4): 344–48.

Shapiro, D. Y., and Altham, P. M. E. 1978. Testing assumptions of data selection in focal animal sampling. *Behaviour* 67:115–33.

Shepherdson, D., Carlstead, K., Mellen, J. M., and Seidensticker, J. 1993. The influence of food presentation on the behavior of small cats in confined environments. *Zoo Biol.* 12:203–16.

Siegel, S. 1956. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Slatkin, M. 1975. A report on the feeding behavior of two East African baboon species. In *Contemporary Primatology*, ed. S. Kondo, M. Kawai, and A. Ehara, 418–22. Basel: Karger.

Sokal, R. R., and Rohlf, F. J. 1995. *Biometry: The principles and practice of statistics in biological research*. 3rd ed. New York: Freeman.

Stanley, M. E., and Aspey, W. P. 1984. An ethometric analysis in a zoological garden: Modification of ungulate behavior by the visual presence of a predator. *Zoo Biol.* 3:89–109.

Streiner, D. L. 1996. Maintaining standards: Differences between the standard deviation and standard error, and when to use each. *Can. J. Psychiatry* 41:498–502.

Suen, H. K., and Ary, D. 1984. Variables influencing one-zero and instantaneous time sampling outcomes. *Primates* 25:89–94.

Tabachnick, B. G., and Fidell, L. S. 2001. *Using multivariate statistics*. 4th ed. New York: Allyn and Bacon.

Tasse, J. 1986. Maternal and paternal care in the rock cavy, *Kerodon rupestris*, a South American Hystricomorph rodent. *Zoo Biol.* 3:89–109.

Thiemann, S., and Kraemer, H. C. 1984. Sources of behavioral variance: Implications for sample size decisions. *Am. J. Primatol.* 7: 367–75.

Tinbergen, N. 1951. *The study of instinct*. Oxford: Oxford University Press.

Todman, J. B., and Dugard, P. 2000. *Single-case and small-n experimental designs: A practical guide to randomization tests*. Lawrence Erlbaum Associates.

Traylor-Holzer, K., and Fritz, P. 1985. Utilization of space by adult and juvenile groups of captive chimpanzees (*Pan troglodytes*). *Zoo Biol.* 4:115–27.

Tyler, S. 1979. Time-sampling: A matter of convention. *Anim. Behav.* 27:801–10.

Velleman, P. F. 1997. *Data desk: The new power of statistical vision*. Ithaca, NY: Data Description.

Vickery, S., and Mason, G. 2004. Stereotypic behavior in Asiatic black and Malayan sun bears. *Zoo Biology* 23:409–30.

Watkins, M. W., and Pacheco, M. 2000. Interobserver agreement in behavioral research: Importance and calculation. *J. Behav. Educ.* 10:205–12.

White, B. C., Houser, L. A., Fuller, J. A., Taylor, S., and Elliott, J. L. L. 2003. Activity-based exhibition of five mammalian species: Evaluation of behavioral changes. *Zoo Biol.* 22:269–85.

White, D. J., King, A. P., and Duncan, S. D. 2002. Voice recognition technology as a tool for behavioral research. *Behav. Res. Methods Instrum*. 34:1–5.

Wilkinson, L. 2004. *SYSTAT: The system for statistics*. Evanston, IL: SYSTAT.

Young, R. J. 2003. *Environmental enrichment for captive animals*. Oxford: Blackwell Science.

Zar, J. H. 1999. *Biostatistical analysis*. 4th ed. Upper Saddle River, NJ: Prentice-Hall.