APPENDIX C

Introduction to Statistics

BONNIE J. PLOGER[1] AND KEN YASUKAWA[2]
[1]*Department of Biology, Hamline University, 1536 Hewitt Avenue, St. Paul, MN, PA 55104, USA*
[2]*Department of Biology, Beloit College, 700 College St, Beloit, WI 53511, USA*

INTRODUCTION

In rare cases behavioral research provides results that are clearcut and easy to interpret, but in most cases things aren't so easy. Behavior is complex and highly variable even in the simplest of situations, so people who study the behavior of animals need tools to help them with their hypothesis testing. Statistical analysis is one of the most important and most frequently used items in the behavioral toolbox. For this reason, it is important for students of animal behavior to have some knowledge and understanding of statistics.

Simply put, a **statistic** is a number that summarizes information about a group of numbers. For example, if you measured the body lengths of all the spring peeper frogs (*Hyla crucifer*) in Clear Lake, you could summarize their lengths by listing the length of each frog, or, alternatively, you could give their average length. The average (or mean) is an example of a **descriptive statistic**, a single number that summarizes information about a group of numbers and is used to describe that group. Descriptive statistics are used to summarize information about (1) the sample of data that you

ISBN 0-12-558330-3

observed and (2) the total population from which the data were drawn. Consider the following example.

Suppose there are 100 adult male spring peepers in a small lake named Clear Lake. You are interested in knowing the mean length of male frogs in this entire adult population. You *could* measure the total adult male population, but this would be impractical and probably impossible. Instead, you would have to restrict your observations to a subset (that is, a **sample**) of the total population. When you measure a sample of 18 males, the mean and other descriptive statistics that you calculate from this sample constitute **sample statistics**. Sample statistics enable you to make precise descriptions of your sample.

But the population of all adult males in the lake cannot be described precisely from the statistics that you derived from your sample. Remember that although you measured only 18 animals, you really wanted to know about the lengths of the entire population in the lake. The population of all 100 adult males in the lake is your **statistical population**, the collection of all elements about which you seek information. Had you measured the lengths of all the adult males in the lake, you could have calculated the **parametric mean**—the mean for the entire population. The parametric mean is an example of a **population parameter**, which is simply a descriptive statistic that is derived from the entire population of interest (the statistical population).

For example, the mean length from your sample of 18 males is an estimate of the mean for the 100 males in the Clear Lake population. The mean and other parameters of the total population may differ from the descriptive statistics of a sample of the population because by chance you may have sampled lots of really large males (or lots of really small ones). Although what you really want to know are the population parameters, what you usually have are sample statistics, which are only estimates of the parameters. But the sample statistics become better and better estimates of the population parameters as the sample size is increased. For example, you would get a much better estimate of the population parameters if you measured and observed 50 males instead of 18. In this case, if you took your data correctly, then your sample statistic should be a good representation of the population as a whole.

The preceding example demonstrates the two purposes for which statistics are computed from data: (1) to describe the data obtained in a sample, and (2) to make inferences about the characteristics of a population on the basis of a sample of observations drawn from that population. The calculation of the mean from a sample is an example of statistical description. The use of the sample mean as an estimate of the population mean (a parameter) is an example of statistical inference. Descriptive and inferential statistical analyses are discussed in further detail in the sections that follow. But the first step in data analysis is to summarize your data.

SUMMARIZING DATA

You can tell little from one observation (one datum) of some phenomenon. Therefore, you always want to summarize data from many different individuals or observations. You must use the right kind of summary for the data that you have collected.

Ways To Express Data

Frequency

The **absolute frequency** is the total number of times you observed some characteristic or phenomenon (e.g., number of hops or number of frogs); the **frequency** is the number in a unit of time (e.g., hops/minute; also called a **rate**) or space (e.g., frogs/puddle; also called a **density**).

Percentage

The absolute frequency divided by the total number of observations times 100 is the **percentage** at which some observation occurred. Percentages can range from 0 to 100%.

Probability

The **probability** is the absolute frequency divided by the total number of observations (it is also called the **relative frequency**). Possible probabilities range from 0 to 1. If every time your professor goes near your animal, it hides, you could say that the chances are 100% that it will happen again or that it happened 10 out of 10 times or that the probability of its happening again is 1. If 75 times out of 100, your animal turned black when you turned on the light, you could say that the chance of such an occurrence happening again is 75% or that the probability is 0.75.

Presenting Data

Tables

A table presents the detailed numerical findings of a study, but it never presents raw data (unless it is in an appendix). Every table must have (1) a descriptive title at its *top*, (2) headings to all rows and columns (including units), and (3) the summarized data (such as descriptive statistics) that make up the body of the table. Tables are numbered sequentially in the order in which you refer to them in the text. You want to present a summary that includes some measure of central tendency and the amount of variation (see the explanation that follows). A helpful "rule of thumb" is that a table is something that you could make with an ordinary typewriter.

Figures

The clearest and easiest way to get the reader to understand your important findings is to present your data in a graphical form (where you may present either raw or summarized data). In scientific writing, graphs, photographs, maps, drawings, and any other illustrations are called figures and are numbered sequentially in the order in which you refer to them in the text. Several kinds of graphs, such as bar graphs, histograms, scatterplots, and box-and-whisker plots, are commonly used in animal behavior. Others kinds, such as pie charts, are not typically used.

When you make a graph, you should follow these rules: (1) Plot the independent variable on the *x*-axis and the dependent variable on the *y*-axis. (2) Always plot your data in such a way as to maximize the chances that the reader will see the point you are trying to make. (3) Use the proper scale for the data—for example, a log scale for growth data. (4) Always include clearly labeled axes with the units you used. (5) For every figure, be sure you have an explicit and clear title placed *below* the figure.

You may be tempted to present the same data in both tables and figures, but you should avoid this temptation. Data should be reported only once (in a table or in a figure, but not in both), so you must think carefully about the *most effective* way to present them. Unless you have limited data to present, a good figure is usually more effective than a table. One guideline to consider when you design a table or a figure is to present it in such a way that it can *stand alone.* In other words, someone should be able to understand the information given in your table or figure by reading it and nothing else. This means that axes and legends must be clearly labeled and that titles may need to be fairly detailed, including names of species and sample sizes, where these are not obvious from the data presented.

DESCRIPTIVE STATISTICS

As the name implies, descriptive statistics are used simply to describe a group of numbers. Suppose that you are doing a study of courtship patterns in male spring peepers in Clear Lake. You successfully trapped, measured, and individually marked 18 of the 100 males in the lake. Say you also followed each marked male for one night and determined the number of times that each animal called during the night. These hypothetical data are presented in Table C.1.

Before calculating any statistics, it is often wise to plot the data. To plot the frequency of numbers of calls, simply plot the number of frogs that called 1, 2, 3, etc. times per night, as shown in Figure C.1 (i.e., plot the number of frogs that fall into each category of calling). This figure is called a **bar graph**. Note that the bars do not touch each other because the

Table C.1 **Hypothetical data on male spring peeper (*Hyla crucifer*) calling frequency.**

Frog I.D.	*Number of Calls*
1	9
2	10
3	6
4	7
5	5
6	11
7	8
8	10
9	9
10	6
11	7
12	8
13	8
14	9
15	9
16	7
17	10
18	9

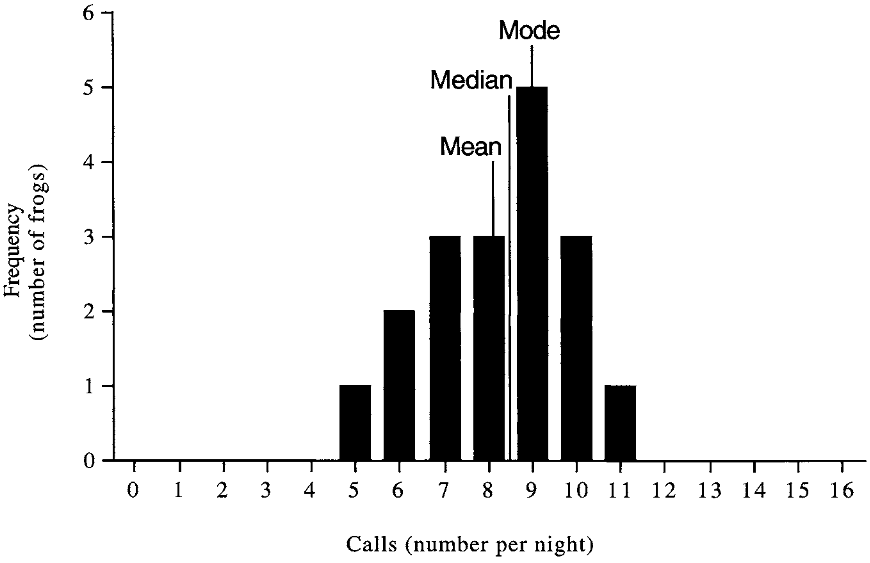

Figure C.1 **Hypothetical data for the number of calls per night by male spring peepers (*Hyla crucifer*) in Clear Lake.**

distribution is **discrete** (i.e., a frog can call 1 time or 2 times a night, but it cannot call 1.5 times in a night). If we had plotted **continuous** data (e.g., length or speed), then we would have used a **histogram** in which the bars touch each other, with each bar representing the frequency of occurrence within a range of possibilities (e.g., the number of frogs between 3 and 4 centimeters in length).

Figure C.1 is a **frequency distribution**, a graph of the distribution of data points that represents how often each value occurred in the sample. To describe your data with descriptive statistics alone, you need to find a way to capture the shape of the frequency distribution. A complete description of the frequency distribution must include a description of both the center of the curve (distribution) and its width. That is to say, when you characterize your data, you must use two descriptive statistics: one that measures the **central tendency** (location) and one that measures the **variability** (dispersion) in your data.

Measures of Central Tendency

Measures of central tendency are descriptive statistics that represent the common values in the distribution. Each statistic of central tendency is a single number that represents the value of the variable where the majority of the data lie (the center of the distribution).

Mean

The **arithmetic mean**, or average, $\bar{X}$(pronounced "ex bar"), is calculated by taking the sum of the values obtained, ΣX, and dividing by the total number of values, n. For example, the mean number of frog calls is 8.22 (from data in Table C.1; see Figure C.1). Calculate the mean as follows:

$$n = 18$$

$$\sum X = 9 + 10 + 6 + 7 + 5 + 11 + 8 + 10 + 9 + 6 + 7 + 8 + 8$$
$$+ 9 + 9 + 7 + 10 + 9 = 148$$

$$\sum X = 148$$

$$\bar{X} = \sum X/n = 148/18 = 8.22$$

Median

The **median** is the value that divides a frequency distribution into two equal halves such that the same number of items fall on each side of the median value. For example, in the series 1, 2, 3, 4, 5, the value "3" divides the data such that there are the same number of points, 2, on either side. If there are an even number of values, you must take the mean of the middle two values. Thus, in the series 3, 4, 5, 6, 7, 8, the median is 5.5.

Similarly, the median number of frog calls is 8.5 (from the series in Table C.1 of 5, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10, 11; see Figure C.1).

Mode

The **mode** is the most common value in a series. In a frequency distribution, the mode is the value of the variable at which the distribution peaks. For example, the modal number of frog calls is 9 (see Table C.1 and Figure C.1).

In a **symmetric distribution** (one in which the left and right sides are mirror images of each other), the mean, median, and mode are identical. In such cases we typically use the mean as the measure of central tendency. In contrast, the three measures of central tendency differ when a distribution is **asymmetric**. Suppose you had a distribution that was **skewed** to the right (in other words, one with a big "hump" on the left-hand side and a long "tail" on the right-hand side). What do you think would be the left-to-right order of the three measures of central tencency? It would be mode < median < mean, but can you figure out why? Here is a hint: Why is the mean to the right of the median in this case? (That is, why is it closer to the side with the long tail?) How can you remember this order? It's alphabetical from the side with the long tail, which is also the side toward which the distribution is said to be skewed.

Measures of Variability

Measures of variability or dispersion are descriptive statistics that represent the spread of values in the distribution on either side of the center.

Range

The **range** is the difference between the largest and smallest values. It represents the maximum spread in the data. For example, for the series 21, 15, 13, 24, 18, 19, the range is 24 − 13 = 11. Similarly, the number of frog calls (Table C.1) spanned from 5 to 11, so the range is 11 − 5 = 6. Note that it would be incorrect in these two examples to say that the range was 13–24 or 5–11 (that is, the range is the *difference* between the minimum and maximum values, not the minimum and maximum values themselves).

Variance

The **variance**, s^2, measures the amount of variability in your sample. Variance differs from the range in that the variance takes into account the distribution of all data points, whereas the range simply describes the distance between the lowest and highest extremes. For example, imagine that you have the following two sets of data:

Dataset A:	5, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 11
Dataset B:	5, 5, 6, 6, 6, 7, 7, 7, 8, 8, 8, 9, 9, 10, 10, 10, 11, 11

In both sets, the lowest value is 5 and the highest value is 11 (range = 6). But in Dataset A, the values are all clustered around 8 and 9, except for the two odd points, 5 and 11, whereas the data are spread widely through the entire range in Dataset B. Clearly, Dataset B with its many dissimilar values is far more variable than Dataset A, with its cluster of data between 8 and 9. Variance is a way of comparing the degree of variability among different sets of data.

To calculate variance, take the **deviation** (or difference) of each value, X, from the mean, $\bar{X}$ (that is, calculate $\bar{X} - X$ for each value). Then calculate the **squared deviation** by squaring each deviation [$(\bar{X} - X)^2$ for each value], to eliminate any negative signs. Add all the squared deviations together [that is, compute $\Sigma(\bar{X} - X)^2$] to calculate the **sum of squares**, and divide by the number of values minus one ($n - 1$), which is called the **degrees of freedom**. In other words,

$$s^2 = \frac{\Sigma(\bar{X} - X)^2}{n - 1}.$$

An easier but equivalent formula is

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n - 1}.$$

For example, the variance in the number of frog calls can be calculated from Table C.1 as follows:

$$\begin{aligned}\sum X^2 &= 9^2 + 10^2 + 6^2 + 7^2 + 5^2 + 11^2 + 8^2 + 10^2 + 9^2 + 6^2 + 7^2 \\ &\quad + 8^2 + 8^2 + 9^2 + 9^2 + 7^2 + 10^2 + 9^2 \\ &= 81 + 100 + 36 + 49 + 25 + 121 + 64 + 100 + 81 + 36 + 49 \\ &\quad + 64 + 64 + 81 + 81 + 49 + 100 + 81\end{aligned}$$

$$\sum X^2 = 1{,}262$$

$$\left(\sum X\right)^2 = (148)^2 = 21{,}904$$

$$s^2 = \frac{1{,}262 - \frac{21{,}904}{18}}{18 - 1} = 2.65$$

Similarly, Dataset A in the preceding example has $s^2 = 1.31$, whereas for Dataset B, $s^2 = 3.82$.

Standard Deviation

The variance is based on the sum of the squared deviations and so has units of measure that are squared. To convert this measure of dispersion to one that uses the original unit of measure, we could take the square root of the variance. That is how to calculate the **standard deviation**

(s, where $s = \sqrt{s^2}$). Another advantage of the standard deviation emerges if your data conform to a **normal distribution**, familiar to most as symmetric, bell-shaped curves around the mean. In a normal distribution, about 95% of the values fall within 2 standard deviations on either side of the mean.

Percentile

Percentiles are points that divide the distribution of the data into hundredths. The 95th percentile is the point on a distribution below which 95% of the data fall.

INFERENTIAL STATISTICS

Experimental Design

Sampling

According to Webster, *to infer* means "to derive as a conclusion from facts or premises." For example, when you see a person driving a police car, you may infer that the person is a police officer. In this example, you made conclusions about a person's employment and lifestyle on the basis of one observation of the vehicle being driven. In statistical inference, you draw conclusions about a large number of events on the basis of your observations of a subset of them, your **sample**.

Suppose that you sampled two sets of 18 male spring peepers in Clear Lake rather than just one as described earlier. Let's say that one sample had a mean of 8.2 calls and the second sample had a mean of 8.8. Both samples came from the same population of adult male frogs in Clear Lake, so why do the two samples differ? There are several possible reasons.

Biased Samples

One reason for the difference might be that different selection criteria were used to choose study animals for the two samples. An example of such **biased sampling** would be if, for your first sample, you selected small males that call rarely, whereas for your second sample, you selected large males that call frequently. In this example, rather than being estimates of the total adult male population, your first sample would be an estimate of the population of small adult males, and your second sample an estimate of the population of large adult males, in Clear Lake. Had you selected small males for both samples, you would still not have a good estimate for the total male population, but only an estimate for small males. Why would someone choose to gather biased samples? It usually occurs by accident. For example, perhaps the first sample was gathered near shore and the second from deeper water. It just so happened that small, quiet males were hanging out in shallow water and big noisy males in deeper water. Unintended bias in sampling is a serious problem. Proper sampling methods are designed to avoid it, among other things.

Random Samples

Another reason for the differences between your two samples might be that *by chance alone*, your first sample happened to consist of animals that called infrequently, whereas your second sample happened to consist of animals that called frequently. This would be an unlikely outcome if chance differences were the only differences between your samples. If you looked at more samples, and the samples were random, differing only by chance, then the calling rates in most samples would be fairly similar, although a few would be quite different. When you do a study, what you want is to have **random samples**. In your study of frog calls, to make sure that the differences between the means of the two samples are the result of chance alone, you must choose the sample males randomly by applying objective, preset criteria for selecting the animals (rather than choosing the easiest ones to catch or the loudest callers).

To do any type of statistical inference, it is *essential* that the sample be random so that it will be truly representative of the population. To ensure randomness in the sampling procedure, make all decisions about the experiment before the experiment begins. Before beginning the experiment, you must decide (1) what individual animals you will use for the experimental and control groups, and (2) what data you will take and how you will take them. You must be very careful *not* to say, "Oh, this looks like a nice aggressive animal, so I'll test this one" or "I didn't quite see what happened that time, so I'll use it as a control."

How do you randomize the collection of data? You have to apply preset, objective, and clearly specified criteria to obtain a random sample. This may involve assigning numbers to your animals and treatment groups and using a table of random numbers to assign the animals to the different treatment groups.

Data Independence

Independent data are those in which the presence of one value or data point has not influenced the presence of another value in your sample. Within any sample, the data must always be independent. For an example of nonindependent data, imagine that you watched a frog in an aquarium and recorded at 15-second intervals where the animal is located. Your data were not independent because the position of the animal in the previous 15 seconds drastically influenced where it was in the following 15 seconds. Similarly, if you always present the animal with a red stimulus after a blue one, then your data will not be independent because the blue stimulus may influence the response to the red one. Your data are useful only when they are independent.

The most common way in which data independence is a problem is through the **pooling fallacy**, which is sometimes called **pseudoreplication**. We often treat repeated observations of the same individual as though they were independent. This is incorrect. To make repeated observations on one individual is not a substitute for making observations on many individuals. All that these repeated observations of one individual do is increase your knowledge about that one individual. In other words, one individual is one data point, and by observing the same individual over and over, you simply increase the reliability of your estimate of that one data point. For example, suppose you want to know if ponds next to highways have fewer frogs than ponds far from highways. You could pick two ponds, one by a highway and one far from any highways, and count the number of frogs in the two ponds once a month for 6 months. This is an example of pseudoreplication. You would generate lots of data, but at the end of the study, all you would know would be whether these two particular ponds differed. You would not know whether the differences had anything to do with being near or far from highways. To avoid pseudoreplication, a better experimental design would be to pick a bunch of different ponds, half near highways and half far from highways. You would want to pick ponds that were as similar as possible (in size and vegetation type, for example) and differed mainly in their proximity to highways. You could then count the frogs in each of these ponds and compare the counts in ponds near highways to those in ponds far from highways. This experimental design would enable you to draw conclusions about the relationship between proximity to highways and number of frogs in ponds.

Independent and Dependent Variables

The goal of many scientific investigations is to determine whether changes in some condition(s) (the **independent variable**) result in different effects (the **dependent variable**). The dependent variable is the score that you measure for each of your study subjects. For example, you might want to know whether frog calling rate (dependent variable) depends on distance from highways (independent variable). You might start by simply measuring the calling rates of frogs at ponds adjacent to highways vs those 4 kilometers from highways. Here, your independent variable would include two **comparison ("treatment") groups**, adjacent to highways and 4 kilometers from highways. Such an **observational comparison** would not involve any experimental manipulation. Next, you might conduct a **controlled experiment**, in which you manipulate the location of the frogs. For example, you might randomly assign equal numbers of frogs to the following three treatment groups: aquaria adjacent to a highway, aquaria 4 kilometers from a highway, and "control" aquaria at the same distance (such as 2 kilometers) from a highway as was the pond from which all the frogs originally were captured.

Observational Comparisons

To make an observational comparison, a researcher simply compares two or more groups that already exist naturally and measures some dependent variable to see whether it differs among these comparison groups. In other words, these are observations that do not involve any experimental manipulation of the independent variables. Examples include comparing the calling rates of frogs in ponds near and far from highways, or calling rates by frogs with different skin colors (certain colors may provide better camouflage so that cryptic frogs may call more because of lower risk of being killed by predators). Observational comparisons are often made at the beginning of a study to detect possible relationships. With such observations you can detect potentially interesting relationships for further study, but they do not enable you to make conclusions about what *caused* an observed relationship. To identify causes requires controlled experiments.

Consider the following example of an uncontrolled observation. Let's say that from your initial investigations of the effects of highways on frogs, you hypothesized that there were fewer frogs near highways because oil from the highways was affecting the hatching success of the frogs' eggs. You then revisited your bunch of ponds that were near and far from highways, this time measuring the amount of oil on the water and the hatching success of frog eggs. If you found that fewer eggs hatched in areas with high amounts of oil than in areas with low amounts of oil, you might be tempted to conclude that the oil *caused* low hatching success. But such a conclusion would be valid only if you had controlled for all other factors. In the case just described, you did not control for all other factors. Thus, although it might be that oil caused low hatching, the low hatching in oily ponds could just as easily have been caused by some other factor (a **confounding factor**) that you did not measure. For example, it might be that the ponds that had more oil in them also happened to have higher levels of lead in them. Perhaps the lead was what was killing the eggs. Or perhaps the ponds that had more oil in them happened to be in more open areas such that the water received more sunlight. High light levels and/or higher water temperatures, rather than oil, might have been responsible for lower hatching success in these ponds. Your observation that fewer eggs hatched in ponds with more oil is still useful in that it indicates that there might be a relationship between oil and hatching success. Such uncontrolled observations are a good first step in figuring out what is going on. But in order to find out what actually *caused* the reduction in hatching success in oily ponds, you must do a controlled experiment.

Controlled Experiments

To do a controlled experiment, the researcher manipulates a factor of interest and holds *all* other factors constant (the **rule of one variable** or one difference among treatment groups). Often, these independent treatment groups include a **control group** that experiences normal conditions, plus one (or more) **experimental group**(s) that receive(s) some manipulation.

To keep all other factors as constant as possible, the researcher should conduct a **laboratory experiment**, which results in high **internal validity** (i.e., if there is an effect, you know exactly what *caused* it). But laboratory conditions can be so artificial that their results may not be generalized to the real world (a lab experiment has low **external validity**). Clearly, then, a researcher cannot achieve both high internal validity and high external validity—these 2 characteristics of experimental design "trade-off" against one another. As a compromise, a researcher may choose to do a **field experiment** by observing organisms in their natural habitat while manipulating some factor of interest (such an experiment would have moderate internal and external validity). The following are examples of a field experiment and a laboratory experiment that you might choose to do to find out whether oil really caused the reduction in hatching success of frog eggs that you observed in the uncontrolled observation described above.

As a field experiment, you could pick a set of ponds that have either no oil or very similar, low levels of oil. The ponds should also be very similar in other potentially important respects, such as size, water temperature and chemistry, and vegetation along shore. Best would be ponds that also had similar hatching success of frog eggs. You could then *randomly* assign half of the ponds to be the experimental group, which would receive a certain dose of oil, and assign the other ponds to the control group. Control ponds could receive pond water instead of oil, to control for any effects of disturbance by experimenters. If you then found that more eggs hatched in the ponds with no oil (control group) than in the ponds that received oil (experimental group), you could conclude that oil *caused* the reduction in hatching success. Concluding that oil caused the difference is reasonable because other differences between the two sets of ponds were controlled by random assignment of treatments. If the two sets of ponds were otherwise identical, then the conclusion that oil *caused* the difference is valid. Keep in mind that if you actually planned to conduct such an experiment that involves adding a pollutant like oil, you would need to use the lowest appropriate dose, for example, similar to that released by roadsides. By doing so, you will not only make your results applicable to your question about the effects of roadside oil, but also minimize harm to the pond organisms. Exposing organisms to deleterious conditions like these may be necessary to advance knowledge needed to reduce harm from pollution, but must also minimize suffering in experimental animals. Of course you would not conduct a field study like this in an ecologically sensitive area that contained locally rare organisms!

The difficulty with field experiments is that you can never find situations in the field where all other factors are truly identical. There are sure to be some variations among your ponds, some of which might have influenced hatching success. If, despite random assignment, the experimental group happened to have more ponds with higher temperatures than the control ponds, then your conclusion that oil caused lower hatching would be questionable; high temperatures might have been the cause. But if you pick large enough samples—that is, if you do the experiment with a large enough number of different ponds—and if you assign ponds randomly to treatment groups, the chances are that differences among the ponds will appear equally in both the control group and the experimental group. For example, if you picked enough ponds and if your assignment of treatments was unbiased, then there would be about the same number of warm and cold ponds in both the experimental group and the control group. In this situation, average temperatures would be the same in both groups, so temperature differences would not be a possible explanation for lower hatching success in the experimental group. With large enough samples, you could reasonably conclude that oil caused the low hatching success. But what if getting large samples is impractical or you cannot make your control and experimental groups similar to each other in all factors other than the one you are manipulating? In this situation, you need a laboratory experiment.

As a laboratory experiment, you could set up in your lab a bunch of aquaria that all had identical conditions. You could then put an equal number of fertilized, healthy frog eggs into each aquarium. You could then randomly assign half of the aquaria to be the experimental group, giving each a prescribed dose of oil. The other aquaria would be assigned to be the control group and would receive no oil (but you could add an equal volume of aquarium water instead). Because you were doing this in your laboratory, you would have (at least theoretically) complete control over all of the variables, so you could set up the experiment in such a way as to be sure that the only difference between the control and experimental aquaria was the presence or absence of oil. If you found that hatching success was higher in the control aquaria than in the experimental aquaria to which oil was added, you could conclude that the oil *caused* the observed reduction in hatching success. However, it is risky to assume that you have controlled all the factors except the one of interest. Even in the laboratory, you might have overlooked a factor that might affect the outcome of your experiment. For example, if you put all your experimental tanks along the window and all your control tanks along the wall, you would have uncontrolled differences in light levels and temperature (confounding factors) between the two groups. The higher light or temperature levels in the tanks along the windows might have caused lower hatching success in the experimental tanks. In a room without windows, tanks might still experience differences in air currents or noise levels if some tanks were placed near air ducts or doors and others were far from such passageways. To avoid these problems, you must be sure that an equal number of tanks in the experimental and control groups are near ducts or doors and that an equal number are far from ducts and doors. In short, the mere fact that you are working in a laboratory does not mean you automatically have controlled conditions. Be aware of easily overlooked differences in variables such as light levels, air currents, noise levels (some walls may be near machine or construction noises), background complexity or color (which might affect animal behavior), and temperature differences along inner and outer walls. The easiest way to avoid these confounding factors is to assign aquaria randomly to control and experimental groups.

The Logic of Hypothesis Testing

Statistical inference is used most frequently to test **hypotheses**. When studying animal behavior, we seek explanations for the patterns that we observe. We learn about the patterns by forming hypotheses and testing specific **predictions** that are based on the hypotheses. For example, suppose we are interested in finding out why male frogs call. We may hypothesize that one **function** of calling by male frogs is to attract females. From this general, **biological hypothesis**, we can deduce specific, testable predictions. Some of these predictions may involve observations of undisturbed animals, whereas others may involve field or laboratory experiments. For example, one prediction of this hypothesis might be that among undisturbed animals in the field, more females will be found near males that called than near males that did not call. Another prediction might be that fewer females will be found near experimentally muted males than near sham-operated, unmuted males (the control treatment). (Similar field experiments involving temporary muting are relatively easy to conduct with songbirds but may be impossible to do with frogs. Keep in mind that the frog study in this handout involves hypothetical examples, not real data.)

To discuss the validity of our biological hypothesis (here, that male calling attracts females), we must first determine which (if any) of the predictions are correct. To test each prediction, we must collect data and determine whether the data fit the prediction. If all the predictions are correct, we may conclude that the data support the hypothesis. (Note that we can never say that the data *prove* the hypothesis—we encourage you to declare a moratorium on use of the word *prove*.) If the data do not support one or more of the predictions, then we must conclude that the data do not support the hypothesis, and thus we reject it. To understand the phenomenon, we will have to form a new hypothesis or modify the old one and then start all over again by designing new experiments and collecting new data to test the predictions of the new or revised hypothesis.

How do we decide whether our data meet our prediction? How can we tell whether the experimental differs from the control? For example, say that in testing the hypothesis that male calling attracts female frogs, you operated on 20 frogs. You muted 10 so that they could not call for a few weeks. You did a sham operation on the other 10, your control, so that they experienced surgery but were still able to call following the procedure. After you released the animals, you found that there were an average of 1.3 ± 0.24 (mean ± 1 standard deviation) female frogs within 5 meters of muted males, whereas there were an average of 3.2 ± 1.4 females within 5 meters of unmuted males.

From these results, you might be tempted to say that there were more females around males that could call than around those that could not. But remember that two groups can differ by chance alone. How can you tell whether the apparent differences between two groups are due to real differences rather than to chance alone? How can we tell whether a mean of 1.3 females is really different from a mean of 3.2 females? This is where statistics are used in testing hypotheses. Statistical inference tests determine how large the observed differences must be before we can be reasonably sure that they represent real differences in the populations from which only a few events were sampled. We can never be *certain* that two groups differ, but we can use inferential statistics to find out how likely (how probable) it is that the differences represent real differences between the groups rather than differences based on chance alone.

Null and Alternative Hypotheses

All statistical tests involve discriminating between pairs of alternative hypotheses (note that these **statistical hypotheses** are distinct from our original, biological hypothesis, which we are attempting to test). The **null hypothesis** is that there are no differences among groups or no effects—that is, any apparent differences are the result of chance alone. The **alternative hypothesis** (the alternative to the null) is that there *are* differences or effects. The alternative hypothesis includes all possible alternatives to the null. Because only one of the two hypotheses can be true, we call these hypotheses **mutually exclusive**. When testing these hypotheses, we accept the simpler, null hypothesis unless there is good reason to reject it. Therefore, when we are doing a statistical test, we usually say that we are testing the null hypothesis. The goal of such testing is to figure out how likely it is that our study would produce our results when the null hypothesis is true.

How are the null and alternative hypotheses related to our original biological hypothesis and its predictions? The null and alternative hypotheses are simply ways of restating one prediction of a biological hypothesis. There are separate null and alternative hypotheses for each test of each prediction. For example, in testing the biological hypothesis that male calling attracts female frogs, several experiments might be conducted. For each experiment, we can make one or more prediction(s) about what the outcome would be if the biological hypothesis were correct. To do a statistical test of each prediction, we first must restate each prediction in the form of a null and alternative hypothesis. We then conduct the statistical test, which is actually a test of whether we should accept or reject the null hypothesis.

The following examples are null (**H_0**) and alternative (**H_a**) hypotheses for the prediction that more females will be near unmuted than will be near muted male frogs. There are two forms for these hypotheses: **one-tailed** and **two-tailed**.

Two-Tailed Hypotheses

H_0: There is no difference between the numbers of females near unmuted and muted male frogs.

H_a: There is a difference between the numbers of females near unmuted and muted male frogs.

A two-tailed hypothesis does *not* specify the direction of the difference; thus a difference toward either tail of the distribution means H_0 is rejected. Rejection of the null hypothesis simply means that two groups differ. Two-tailed hypotheses are the most appropriate to use when you have reason to expect groups to differ but have no reason to expect the difference to be in a particular direction.

One-Tailed Hypotheses

H_0: There are not more females near unmuted than near muted male frogs.

H_a: There are more females near unmuted than near muted male frogs.

When you have good reason to expect groups to differ in a particular direction, a one-tailed hypothesis is appropriate. In our example, rejection of the 1-tailed null hypothesis means that there are more females near unmuted than near muted male frogs. Failure to reject the null means either that (1) there were similar numbers of females near unmuted and near muted males or that (2) there were more females near muted than near unmuted males. We cannot distinguish between these two possibilities if we failed to reject the one-tailed null hypothesis. Had we done a two-tailed test, we would have detected a difference if there had been more females near muted than near unmuted males. The selection of a one-tailed test must be based on a good reason for expecting differences in a particular direction, such as past studies of the same or related species showing differences in one direction. (A one-tailed test may also be appropriate if you are testing a prediction of a theoretical model that exists in the literature and the predicted differences must be in one direction for the model to be valid.)

Suppose that you tested the foregoing one-tailed hypothesis and discovered that there appeared to be more females near muted than near unmuted males. Could you switch to a two-tailed hypothesis "after the fact"? No, absolutely not. The null and alternative hypothesis and whether they involve one or two tails must be stated *before* doing the statistical test. In other words, just as in a conversation, you must ask the question before you can answer it. The set of statistical hypotheses to be tested must be chosen before, not after, the analysis and final decision.

Significance Level

Remember that when we have not actually measured the entire population, we do not know which is true, H_0 or H_a. We may decide to accept H_0 when it is true (a correct decision) or when it is false (an incorrect decision). Alternatively, we may decide to reject H_0 when it is true (an incorrect decision) or when it is false (a correct decision). These possibilities are listed in Table C.2.

Table C.2 illustrates the two types of mistakes that we might make when we decide whether to accept or reject H_0. Accepting a null hypothesis when it is actually false is a **type II error**; rejecting a null hypothesis when it is really true is a **type I error**. We try to minimize both types of errors. Type II errors are minimized by increasing our sample size. When we fail to reject the null hypothesis but have only a small sample, we must consider the possibility that a larger sample would have caused H_0 to be rejected (statisticians call it a lack of **statistical power**).

Statistical inference tests are designed to calculate the probability (**P**) that chance alone produced your results if the null hypothesis is true. A P-value equal to 0.05 means that the likelihood of a type I error is 5%. In other words, if you took 100 samples, in 5 of the 100 you might, by chance alone, incorrectly reject H_0 even though H_0 is actually correct.

Table C.2 **Results of decisions to accept or reject the null (H_0) and alternative (H_a) hypotheses.**

	Actual Condition in Nature	
Your Decision	*H_0 is Really True*	*H_a is Really True*
Do not reject H_0 (i.e., accept H_0)	Correct decision	Type II error
Reject H_0 (i.e., accept H_a)	Type I error (α level)	Correct decision

To decide whether a particular result supports H_0, the calculated P-value (type I error) is compared with a predetermined maximal level called the **significance level** or **alpha (α) level.** Typically in animal behavior, $\alpha = 0.05$; we will use $\alpha = 0.05$ in this manual.

The significance level is what you use to assess how confident you can be that you are making a correct decision when you reject H_0. With an α of 0.05, you can have 95% confidence that your decision is correct. Similarly, with $\alpha = 0.01$, you can have 99% confidence that your decision is the correct one. Because 99% confidence seems better than 95%, and 99.9% is better still, why use 95%? It's because of another trade-off, this time between type I and type II error (β). The higher our confidence that we will not incorrectly *reject* the null hypothesis (that we will not make a type I error), the more likely it is that we will incorrectly *accept* the null hypothesis (make a type II error). In other words, you can't have it both ways. An α of 0.05 is a good compromise between the two kinds of error.

Test Statistic

After stating our hypotheses and selecting an α level, we must select and carry out the appropriate statistical test (see the next section). By plugging the values of our sample into a formula for the test statistic (some common ones are t, F, and χ^2), we end up with one number that summarizes the sampled data. To make our decision, we need the P-value associated with this number (such as the value of t, F, or χ^2). If we knew how, we could use integral calculus to figure out the P-value. Luckily for those of us who are somewhat calculus-impaired, we can also look up the P-value in a published table. There is one slight difficulty, however. For any test statistic that we may wish to look up in a table of P-values, there are an infinite number of test statistic values and P-values. Unfortunately, there aren't any publishers who are willing to print a table of infinite length. Instead of *all* of the values, only a few, representative ones are tabulated. These tabulated values are called **critical values.**

For a particular test statistic, one critical value is associated with a particular P-value and sample size. Thus, to compare the numbers of females near muted and near unmuted male frogs, you would look up the critical value for the test statistic when P = 0.05 and sample size (n) = 10. Why look for the critical value corresponding to P = 0.05? Because your level of significance (α) was 0.05, and you are tying to decide whether your results are significant or not significant at that level.

Some test statistics make use of degrees of freedom (abbreviated "d.f.") instead of sample size (n). Degrees of freedom vary with sample size. Critical values for these test statistics are uniquely associated with a particular P-value and d.f. To calculate degrees of freedom, see your instructor or a statistics book.

To say that there is only one critical value for each P-level is not quite correct. Although there is only one value per P-level listed in a statistical table, for most test statistics, critical values actually come in pairs that have the same absolute value but are opposite in sign. Thus, even though only the absolute value is listed in a table of statistics, there are really two critical values, one positive and one negative, for each combination of P-level and sample size.

The absolute value of the critical value is the largest value of the test statistic that you should expect to observe if H_0 is true. Observed values larger than the critical value mean that H_a is probably true.

Decision Rule

We are finally ready to make a decision about whether to accept or reject H_0. To do this, we use the following **decision rule:**

> If the observed test statistic ≥ the critical test statistic, then you should reject H_0, accept H_a, and conclude that your results are significant at the $\alpha = 0.05$ level.
>
> If the observed test statistic < the critical test statistic, then you have failed to reject H_0 and should either (1) conclude that the results were no different from random, or (2) suspend judgment because your sample sizes were too small to reject H_0.
>
> *Exceptions*: In some statistics books (e.g., Seigel 1956 but not Sokal & Rohlf 1981), the instructions for the Wilcoxon matched-pairs test and the Mann–Whitney test (see below), call for rejecting H_0 when the observed test statistic is *less than* or equal to the critical value.
>
> When calculating correlations, in addition to determining the statistical significance, you must look at the value of the correlation coefficient (r), which describes the strength of the association. You can conclude that you have a high correlation (a strong association) if the correlation is statistically significant *and* $r > 0.7$. You should be aware that if you have a sufficiently large sample size, you might get statistical significance even if $r < 0.2$, which you should interpret as a negligable relationship (Martin & Bateson 1993).

These days most students have access to computer programs that do statistical analyses. Some examples of such statistical software are SPSS, SAS, SYSTAT, Statview, and JMP. The common spreadsheet program EXCEL also does statistical analyses, although recent reports claim that the results are sometimes incorrect. It's a good idea to try an example with a known answer (such as a worked-out example from a textbook) to test a particular calculation. The great advantage of most computerized statistical tests is that they automatically compare the observed value of the test statistic to the critical value. In other words, you don't have to use a table to figure out the P-value.

The computer program reports as the P-value the probability that the observed value of the test statistic will lead you to reject H_0 when H_0 is actually true (type I error). You want to keep your chances of making this type of mistake low, so you want $P \leq 0.05$, which allows you to be at least 95% certain that your decision to reject H_0 is correct (i.e., $\alpha = 0.05$). Thus, when using a computer, we use the following decision rule:

> If the computer produces from your data an observed $P \leq 0.05$, then you should reject H_0, accept H_a, and conclude that your results are significant at the $\alpha = 0.05$ level.
>
> If the computer produces from your data an observed $P > 0.05$, then you have failed to reject H_0 and should either (1) conclude that the results were no different from random, or (2) suspend judgment because your sample sizes were too small to reject H_0.

Choosing the Appropriate Statistical Test

Although they all do basically the same thing, which is help you decide whether to accept or reject H_0, there are many different kinds of test statistics. Each is created by a mathematical formula that produces one number from the set of values in your sample. Each test (with rare exceptions) uses the logic outlined in the preceding sections. The steps that you follow, from forming your general hypothesis through the final decision whether to accept or reject H_0, will be the same for all the tests that you will normally encounter. All the tests provide a way of deciding whether differences in samples result from real differences or from chance alone, on the basis of how likely it is that the value of the test statistic that you observed from your sample(s) could have been produced by chance.

With so many tests to choose from, how can you decide which test is the most appropriate for your data? The choice depends on the type of question you are asking and on the way your data will be measured. These topics are discussed in the sections that follow.

Type of Question

Statistical questions can be divided into four basic groups: (1) questions about one sample, (2) questions about two or more related samples, (3) questions about two or more unrelated ("independent") samples, and (4) questions about correlation and regression.

Questions about a single sample concern whether a particular sample could have come from some specified population. One-sample statistical tests answer questions such as the following: Is it likely that the sample was drawn from a population with a particular distribution (e.g., normal, Poisson, binomial). Is there a significant difference between the observed frequencies and the frequencies that we would expect on the basis of some principle, such as expectations from transmission genetics or from events occurring at random? For example, you might ask whether the sex ratio of frogs in Clear Lake is 50:50, as expected by chance. To find the answer, you could capture a single sample of 30 adult frogs, count the males and females in the sample, and use a statistical test to decide whether the difference in the numbers of males and females in your sample was significantly different from 50:50.

In contrast, we might ask questions about differences between two or more comparison groups: Is there a difference between the effects on the treatment and control groups? Is one treatment better than the other(s)? Does the effect differ among different types of observational groups? To answer these questions, we must sample from different groups, so we refer to each group as a different *sample*. The samples may be related to each other (dependent) or unrelated to each other (independent).

Questions about two or more related samples arise from experimental designs in which the same individuals are measured more than once. When the same individual is exposed to two treatments at different times, we say that the two samples are **related** or **matched**. For example, you would have two related samples if you compared the calling rates of 20 male frogs before and after they were confined in buckets. The two samples would be (1) before confinement and (2) after confinement. The samples are related because the same individuals are used in both samples. If the same individuals are measured more than twice, the design is usually called a **repeated-measures design**. Occasionally, samples are matched by other criteria, such as body length, past experience, or age.

When different, unrelated individuals are used in each comparison group, we say that the two samples are **independent**. Such designs lead us to ask questions about two or more unrelated ("independent") samples. For example, all the unmated male frogs and all the unmated female frogs in Lake Minnetonka belong to two independent groups because different, unrelated individuals are in each group. This example specifies unmated frogs because, for some questions, members of mated pairs would have to be considered related, not independent.

Clearly, *samples* (the treatment or comparison groups that are the independent variables) can be related to or independent of one another. But remember that *within* any one sample, each *data* point that you measure must always be independent of every other data point (see the foregoing "Experimental Design" section for a discussion of independent *data*).

Questions about **correlation** and **regression** concern whether one type of score that you measured varies with another type of score. These questions ask whether there is some sort of relationship between the two types of scores, which must both vary continuously rather than being measured in discrete categories. Although they seem quite similar, the questions asked by correlation and regression analyses are actually very different, and you should be very careful in choosing one or the other.

We may ask whether two types of scores vary together in a systematic way. In other words, are two variables **correlated** (associated)? For example, you might compare the body sizes of male frogs to their calling rates. If you found a significant **positive correlation** between body size and rate of calling, you could conclude that big frogs call more than little frogs. Or, if you found a significant **negative correlation**, you could conclude that big frogs call less than little ones. It is very important to note here, however, that *correlation does not imply causation*. In other words, you could *not* say that large size *causes* high calling rates; there may be a third factor that affects both size and calling. If calling takes a lot of energy, then frogs that are eating well might be able to call more often, *and* would attain larger size, than those catching fewer, poorer-quality prey. In this example, the third, causal factor would be caloric intake. For another example, a positive correlation between countries in which food is spicy and countries where there are frequent earthquakes would not mean that eating spicy food causes (external) earthquakes! You may laugh, but there are examples almost as naïve in the biological literature, some current. So *be careful* about what you infer from correlation.

In contrast to correlation, regression analysis tests for a functional relationship between an independent variable, x, and a dependent variable, y (i.e., $y = f(x)$). Note that in both correlation and regression analysis, we are measuring a pair of continuous variables to determine whether they vary together in some way. But in correlation analysis, we cannot say which variable is dependent and which is independent. By contrast, in regression analysis we explicitly test whether one type of score, the dependent variable, depends on the other, independent variable.

In regression analysis, functional relationships can take on many shapes. In this manual, we introduce the simplest form of this type of analysis by asking whether the relationship is linear. You may remember the following equation from high school: $y = mx + b$, where m is the slope (rise/run) and b is the y-intercept. For some reason unknown to us, however, statisticians write the same equation as: $y = bx + a$, with b for the slope and a for the y-intercept. For example, we might shine light of 10 different intensities on 10 different tadpoles (larval frogs), measure the amount of growth of each tadpole, and ask whether tadpole growth is a function of light intensity (it doesn't make sense to say it the other way around). Here, light intensity is the independent variable, plotted on the x-axis, and tadpole growth is the dependent variable, plotted on the y-axis. If we were to find a significant regression, we would say that tadpole growth is a function of light intensity. The relationship could be positive (as indicated by a positive slope), meaning that tadpoles grow more rapidly with brighter light, or negative (negative slope), meaning that tadpoles grow more slowly with brighter light. Note again that saying that y is a function of x almost seems like saying x *causes* y, but we are *not* necessarily justified in inferring a causal relationship. We would need to do a controlled experiment to test for causation.

Levels of Measurement

All kinds of (independent and randomly obtained) data are useful and interesting, but different kinds of data require different techniques for presentation and analysis.

Nominal data result when the dependent variable is a measure that simply classifies objects or characteristics into separate categories, but the categories cannot be ranked in any particular order. For example, the sex of a spring peeper and the kind of marsh plant on which these frogs can be found are nominal variables. The categories male and female cannot be ranked, and cattails, reeds, and sedges could be placed in any order with equal validity. Nominal measurement variables are somewhat uncommon in animal behavior, but we often use nominal categories as a **classification variable** (that is, to identify comparison groups).

In the two-sample example that follows, imagine that you wanted to compare two groups to find out whether frogs belonging to one group generally occurred on different types of marsh plants than frogs belonging to the other group. To answer this question, you recorded the capture location (plant type) of eight frogs belonging to two different groups, where group 1 and group 2 could be males and females (or drug A vs drug B, or large frogs vs small frogs). You recorded the following data:

Group 1	*Group 2*
cattails	cattails
reeds	cattails
sedges	cattails
sedges	cattails

Here the independent (classification) variable is group (1 vs 2) and the dependent (measurement) variable is plant type (cattails, reeds, sedges). Once again, because the scores for the dependent variables cannot be ordered or ranked, these are nominal data.

Ordinal data are collected when the dependent variable is a measure that classifies objects or characteristics into mutually exclusive categories, *and* these categories can be put in a ranked order. You can think of *ordinal* as meaning "ordered" (ranked). For a one-sample example, imagine that you wanted to find out how attractive spring peeper calls are to females. You captured a bunch of females and observed their responses to calls played from a tape recorder. Some females showed no reaction, and a few tried to climb onto the tape recorder. Some others turned toward the sound but did not approach it, whereas others hopped toward the source of the calls. It would be reasonable to rank these responses in the following order: (1) no reaction, (2) orient toward call, (3) approach call, (4) mount speaker. Although we gave these categories numbers (1–4), they are really ordinal values (first, second, third, fourth) rather than integers. Because the categories can be ranked, these are called ranked, or ordinal, data.

In the two-sample example that follows, imagine that you wanted to compare 2 types of spring peeper calls to find out whether one is more attractive to females than the other. To answer this question, you observe the responses of five females to call type 1 and of five others to call type 2. You recorded the following data:

Call Type 1	*Call Type 2*
1	3
1	4
2	4
3	3
1	2

In this case the independent (classification) variable is call type (1 vs 2) and the dependent (measurement) variable is female response (1 = no reaction, 2 = orient toward call, 3 = approach call, 4 = mount speaker). Once again, because the scores for the dependent variable can be ordered, these are ordinal data.

Although there are more precise mathematical definitions of **interval data** and **ratio data**, you will have these types of data when the dependent variable consists of numbers that can be ranked (such as ordinal data) and when, in addition, the distance between each number and the next is of known size. With interval and ratio data, you can associate each object with a unique number along some continuous measurement scale. Measurements such as length, height, weight, volume, temperature, and time are continuous variables because theoretically, if you could measure accurately enough, an infinite number of measurements would be possible between any two measurements. Rates, such as the number of events per unit time or per bout of behavior, can also be treated as interval/ratio data.

For example, imagine that you wanted to find out whether certain sizes of male spring peepers were more common than other sizes. For a one-sample example, let's say you measured the body lengths of nine frogs as follows: 2.42, 2.43, 2.50, 2.51, 2.52, 2.55, 2.57, 2.60, and 2.65 centimeters.

Had you measured less accurately, or if you combined these lengths into categories, your data table might look like this:

Frog Body Length (cm)	*Number of Frogs of that Length*
2.4	2
2.5	3
2.6	4

Note that when you measure less accurately or when you combine more accurate measurements into categories, it may not look as though there is a unique number associated with each object (here, each frog). But in theory you could always associate a unique number with each object.

In the two-sample example that follows, imagine that you wanted to compare two groups to find out whether frogs belonging to one group had a different body length than frogs belonging to the other group. To answer this question, you recorded the body lengths of 10 frogs belonging to two different groups, where group 1 and group 2 could be males and females (or drug A vs drug B, or heavy frogs vs light frogs). You recorded the following data:

Group 1	*Group 2*
1.80	1.98
1.90	2.08
2.40	2.19
2.38	2.28
2.30	2.42

Here, the independent variable is group, and the dependent variable is body length. Once again, because the scores for the dependent variables can be ordered along a continuous scale, with known distances between each number and the next, these are interval/ratio data.

CHANGING NOMINAL DATA INTO ORDINAL DATA

When you have data that are counts of how many times each subject reacted to mutually exclusive treatments, you may be tempted just to count how many individuals reacted to each treatment. You would then have **enumeration data** (counts) that you would analyze with a chi-square type of test (see below). But by just counting how many individuals reacted, you will lose considerable information. Martin & Bateson (1993 pp. 72–73) suggest two more powerful ways to handle these types of data—methods that you can use whenever you can associate specific reactions with specific individuals. The following example illustrates these different ways.

Example

You put peanuts without shells and M&Ms in a pile and use the raw data table that follows to record each peanut and M&M taken by each individually identifiable squirrel. You want to know whether squirrels show a preference for one of these food types.

Squirrel Name	*Peanuts*	*M&Ms*
Unmarked A	0	1
Two-nicks	1	0
Two-nicks	1	0
Two-nicks	0	1
Broken-paw	1	0
Scar-face	1	0
Scar-face	1	0
Broken-paw	1	0
Two-nicks	1	0
Two-nicks	0	1
Broken-paw	1	0
Unmarked B	0	1
Unmarked B	0	1
Unmarked B	1	0
Unmarked C	1	0
Unmarked C	1	0
TOTALS	11	5

Possible Analyses

You could analyze these data with a chi-square goodness-of-fit test by pooling all the squirrels, but with this approach, information about the behavior of individuals is lost and you face problems of pseudoreplication. Because you know the number of nuts taken by each individual squirrel, you can use the following methods to explore the data more thoroughly.

Absolute Differences

In this method, for each subject you subtract the response to treatment 2 (M&Ms) from the response to treatment 1 (peanuts). To do this for the raw data given earlier, you would need to sum the scores for each individual squirrel (see below) and then calculate d_i, the difference of the total M&Ms – total peanuts taken by each squirrel.

Squirrel Name	*Peanuts*	*M&Ms*	d_i	*Unsigned Rank*	*Signed Rank*
Unmarked A	0	1	−1	2	−2
Two-nicks	3	2	1	2	+2
Broken-paw	3	0	3	6	+6
Scar-face	2	0	2	4.5	+4.5
Unmarked B	1	2	−1	2	−2
Unmarked C	2	0	2	4.5	+4.5

Analyze these data with the Wilcoxon matched-pairs signed-ranks test by ranking the differences (d_i) and following the instructions given for this test.

For example, for these data: sum + = 17 (that is, the sum of all ranks with positive signs is 17), sum − = 4 (the sum of all ranks with negative signs is 4), and T = smaller of these sums of like-signed ranks, so $T = 4$. For $n = 6$, $P < 0.05$, T(critical) = 0. Because T(observed) > T(critical), conclude that the differences are not significant (NS). (In the statistical tables from Siegel 1956 used for this example, Wilcoxon has the atypical decision rule: reject H_0 when T(observed) < T(critical).

Response Ratios

In this method, first sum the scores for each individual squirrel, as you did above. Then let the response ratio = response to treatment 1/(response to treatment 1 + treatment 2) as follows:

Squirrel Name	*Peanuts*	*M&Ms*	*Response Ratio*
Unmarked A	0	1	0
Two-nicks	3	2	0.6
Broken-paw	3	0	1
Scar-face	2	0	1
Unmarked B	1	2	0.33
Unmarked C	2	0	1

To analyze the data when you have just two treatment groups, as here, you would use the Wilcoxon matched-pairs signed-ranks test to compare the observed response ratio to the chance level of response of 0.5. Be careful, especially if you do your analysis on computer, that you analyze the *response ratios* compared to the chance response values, as given below. Do *not analyze* the actual scores of number of peanuts and M&Ms taken.

Squirrel Name	*Response Ratio*	*Chance Response*	d_i	*Unsigned Rank*	*Signed Rank*
Unmarked A	0	0.5	−0.5	4.5	−4.5
Two-nicks	0.6	0.5	+0.1	1	+1
Broken-paw	1	0.5	+0.5	4.5	+4.5
Scar-face	1	0.5	+0.5	4.5	+4.5
Unmarked B	0.33	0.5	−0.2	2	−2
Unmarked C	1	0.5	+0.5	4.5	+4.5

In this case, sum + = 14.5, sum − = 6.5, T = smaller of these sums = 6.5. For $n = 6$, $P < 0.05$, T(critical) = 0. Because T(observed) > T(critical), conclude that the differences are NS. (Remember, in the tables used for this example, Wilcoxon has the odd decision rule: reject H_0 when T(observed) < T(critical).

Because results obtained using absolute differences may differ from those obtained using response ratios, you should use both methods to analyze your data (Martin & Bateson 1993). If results are contradictory, be sure to discuss reasons that could explain the contradiction. Note also that analysis of absolute differences is more sensitive to individuals who respond strongly to a treatment. Analysis of ratios is more sensitive to variation within individuals (Martin & Bateson 1993). These sensitivity differences might help explain contradictory results of these tests.

SUMMARY OF STATISTICAL PROCEDURES

Before Data Collection

1. State the biological question or hypothesis.
2. Make specific predictions of what will happen if this hypothesis is correct.
3. Design experimental or observational comparisons to test each prediction. You must collect separate data and do a separate statistical test for each prediction. Remember that for each test, the *data must be independent and sampling must be random*. The steps that follow assume you are testing a single prediction.
4. Set up the H_0 (null) and H_a (alternative) hypotheses for your prediction.
5. Decide whether the test will be one-tailed or two-tailed. (Make H_0 and H_a consistent with your decision about the number of tails. That is, if you decide to do a two–tailed test, make sure that both H_0 and H_a are phrased as nondirectional, two-tailed hypotheses.)
6. Determine whether your question involves (a) one sample, (b) two related samples, (c) two unrelated samples, (d) *k* related samples, (e) *k* unrelated samples, (f) an association (correlation), or (g) a regression.
7. Determine what level of measurement you will use. That is, decide whether your data are nominal, ordinal, or interval/ratio.
8. Decide what statistical test you will use. Use Table C.3 to pick possible tests, and then determine which is most appropriate. Where more than two tests are listed for the same type of data, the more powerful (and therefore preferable) test is underlined. Additional nonparametric tests are discussed in Siegel & Castellan (1988) and in Conover (1980). If your data are interval/ratio, you must decide whether to use a parametric or a nonparametric test. You should use a parametric test if the data meet the assumptions of such tests; otherwise, use a nonparametric test. Additional parametric tests are presented in Sokal & Rohlf (1981).
9. Specify a significance (α) level. Published studies of animal behavior generally use $\alpha = 0.05$.

***Table C.3* Choosing a statistical test.**

Type of Data	*Statistical Test*
One sample	
Nominal	χ^2 goodness-of-fit test, binomial test
Ordinal or interval/ratio	Kolmogorov–Smirnov one-sample test
Two related samples	
Nominal	McNemar test for significance of changes
Ordinal or interval/ratio	Sign test, Wilcoxon matched-pairs test
Interval/ratio only*	Student's *t* test for matched samples
Two unrelated samples	
Nominal	χ^2 test of Independence (of two samples) Fisher exact test
Ordinal or interval/ratio	Mann–Whitney *U* Test (to detect differences in central tendency such as means, modes) Kolmogorov–Smirnov two-sample test (to detect *any* differences, including difference in variability)
Interval/ratio only*	Student's *t* test for independent samples
k related samples	
Nominal	Cochran Q test
Ordinal or interval/ratio*	Friedman two-way analysis of variance
Interval/ratio only*	Two-way analysis of variance without replication
k unrelated samples	
Nominal	χ^2 test of independence (of *k* samples)
Ordinal or interval/ratio	Kruskal–Wallis test
Interval/ratio only*	One-way analysis of variance (ANOVA)
Association (correlation)	
Nominal	χ^2 test of independence (of *k* samples)
Ordinal	Spearman rank correlation
Interval/ratio*	Pearson correlation coefficient
Functional relationship	
Interval/ratio only*	Least-squares regression

*Tests marked "Interval/ratio data only" are parametric tests, which should be used only if sample sizes are large ($n > 30$) and meet the assumptions of parametric tests, especially the assumption that the data are normally distributed. All other tests in this table are nonparametric tests, which are more appropriate than parametric tests when sample sizes are small and avoid the restrictive assumptions of parametric tests. Note that χ^2 tests for nominal data can be used for all other types of data, but they are not as powerful and so are less likely to detect significant differences when such differences are real.

During Data Collection

Be sure *data are independent* and *sampling is random.*

After Data Collection

1. Summarize the data by computing descriptive statistics (e.g., mode, mean, standard deviation) and graphing. The appropriate descriptive statistics and graphical methods for various levels of measurement of data are as follows:

Type of Data	*Measure of Central Tendency*	*Measure of Dispersion*	*Type of Graph*
Two or more samples of data that are			
Nominal	Mode	None	Bar graph (frequency or %)
Ordinal	Median	Percentiles Range	Bar graph (frequency or %)
Interval or ratio	Mean	Standard deviation	Frequency distribution
		Variance (percentile, range)	Probability distribution Box-and-whisker plot
Measures of association			Scatter plot (of data points) or point graph (medians or means)
Measures of functional relationship			Scatterplot (of data points) with or without regression line

2. Do the statistical test (compute the value of the test statistic).
3. Decide whether to accept or reject the null hypothesis (H_0). To make this decision, use the decision rule that is associated with the statistical test that you used.

   In most but not all tests, the decision rule is

   If the observed test statistic ≥ the critical test statistic, then reject H_0, accept H_a, and conclude that your *results are significant.*

   If the observed test statistic < the critical test statistic, then either (1) accept H_0 (reject H_a) and conclude that your *results are not significant* or (2) suspend judgment if you believe that the *sample sizes were too small.*

   Two exceptions are the Wilcoxon matched-pairs test and the Mann–Whitney *U* test, for which some statistics books call for rejecting H_0 when the observed test statistic is *less* than or equal to the critical value.

When calculating correlations, you can conclude that you have a high correlation (a strong association) if the correlation is statistically significant *and* the correlation coefficient (*r*) is greater than 0.7. Beware: with large sample size, you might get statistical significance even if $r < 0.2$, which you should interpret as a negligible relationship (Martin & Bateson 1993).

For *computer-based analyses*, the decision rule is

If the computer gives a P-value ≤ 0.05, then reject H_0, accept H_a, and conclude that your *results are significant.*

If the computer gives a P-value > 0.05, then either (1) accept H_0 (reject H_a) and conclude that the *results are not significant*, or (2) suspend judgment if you believe that the *sample sizes were too small.*

4. Report the results of your statistical test. Use an appropriate standard format, which includes a verbal description of the observed pattern (in the past tense) and a parenthetical statement that includes the value of the test statistic that you calculated from your data, the sample size or degrees of freedom, and the P-value associated with your test statistic. For example, you might write, "Tadpole growth rates increased significantly with increasing light intensity (Spearman $r = 0.79$, $n = 68$, $P < 0.0001$)" or "Female frogs were significantly larger than male frogs (Mann–Whitney $U = 130.5$, $n = 49$, $P = 0.001$)."

LITERATURE CITED

Conover, W. J. 1980. *Practical Nonparametric Statistics.* 2nd ed. New York: Wiley.

Martin, P. & Bateson, P. 1993. *Measuring Behaviour: An Introductory Guide.* 2nd ed. Cambridge: Cambridge University Press.

Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill.

Siegel, S. & Castellan, N. J. 1988. *Nonparametric Statistics for the Behavioral Sciences.* 2nd ed. New York: McGraw-Hill.

Sokal, R. R. & Rohlf, F. J. 1981. *Biometry.* 2nd ed. New York: Freeman.