Biol 326
Animal Behavior
Austin College
Statistics for Novices

Section 1: General Principles

The purposes of science include describing and explaining "natural phenomena," which include all objects and processes in the observable, physical universe. To do this we observe the objects and processes, draw inferences from our observations, make predictions about natural phenomena based on our inferences, generate hypotheses based on our predictions, and perform additional observations or experiments to test our hypotheses. Our inferences may be supported by additional information, or they may be refuted. Over time a conceptual framework builds up that explains a range of related phenomena, and that makes predictions about phenomena that have not yet been observed.

Organisms and their characteristics are natural phenomena that we can try to understand through the process of science. A characteristic of an organism that can be observed or measured, and that differs from one individual to the next in an observable fashion, is a **variable**. There are innumerable types of observations that we may make or variables we can measure in the science of biology. One general type is qualitative observations; the characteristics that we make qualitative observations on are called **qualitative variables** or **attributes**. These variables are mutually exclusive, non-overlapping states of physiological, morphological, behavioral, or ecological categories. For instance, the distinction between male and female is a qualitative one, as is the difference between larva and adult. Another general type of variable in biology is the ranked variable, in which states or conditions (or for us behaviors) are assigned ranks in a series. The ranks do not imply an step-wise or incremental progression of intensity from one state to the next, but allow us to recognize discrete categories or conditions.

Qualitative and ranked variables do not "measure" things in the traditional sense. Measurement variables are called **quantitative variables**. There are two types of quantitative variables. For discrete, meristic, or discontinuous quantitative variables, observations can take on only whole number values; these are counts of things. Number of individuals in a group, number of ectoparasites on an individual animal, and number of eggs laid by a female are examples of discrete quantitative variables.

Continuous quantitative variables can theoretically take on any value between two fixed points. The accuracy of the measurement of the variable depends on the sensitivity of the measuring device. Here the true value of the variable cannot be known with certainty, but can only be estimated. There are obviously many examples of continuous quantitative variables in biology. Anything to do with body size or size of structures, mass of organisms or pieces of organisms, longevity of organisms, preferred activity temperature, etc. would be included in this category.

In biology, our explanations of natural phenomena often make reference to similarities and differences between various categories. For instance, we might want to understand the effect of several pesticides on two species of insect, one harmful and one beneficial. Which pesticide would be the most effective in controlling the population of the harmful insect, and which would be the least detrimental to the beneficial insect? To answer our question we would need to compare not only the effect of the various pesticides on each of the two species, but we would also need to compare the two species' resistance or susceptibility to the pesticides. We might expect that pesticide A is more effective than pesticides B, C, or D, and that insect species A is more resistant to all of the pesticides than is species B.

Biology is also a science in which absolute answers to our questions are uncommon. This is because of the variation (genotypic and phenotypic) that is inherent in all biological systems. Individual members of a species differ from one another in obvious and subtle ways, as a result of both environmental and genetic factors. Absolute answers to our questions are also uncommon because in general it is impossible to make all possible measurements on a particular variable. In the example above, it might be theoretically possible to test the effects of each pesticide on **every individual** of species A that is presently alive. The **set of all possible observations** that could be made on a particular phenomenon is called a "statistical population." In general it is impossible to work with a statistical population because of practical constraints on time, energy, and money [exceptions might be the wing length of all alala (Hawaiian Crow) or the body mass of

all Jabiru Storks]. Thus it is inherently impossible to know with absolute certainty the value of all possible observations that could be made on a particular phenomenon. This is where statistics comes in.

One purpose of statistics is to make a guess at the true value of a set of potential observations (a statistical population), based on a subset of those observations. The true value of a set of observations, which is practically "unknowable" for the reasons discussed above, is called a "population parameter." We cannot measure the effects of pesticide A on every individual of species A, but we certainly could measure the effect (for instance on longevity after exposure) of pesticide A on a single individual of species A. What would such a measurement tell us? Would it tell us the value of longevity for every possible similar measurement? Obviously not, because each individual of species A might be more or less susceptible than other individuals. We should measure the effect of the pesticide on several individuals of species A to get a reasonable estimate of the longevity after exposure for the population. In other words, we would take a **sample** of a statistical population, and use the sample to estimate the true or parametric value of the variable we are interested in (e.g. average longevity after exposure). The number of observations that we take is referred to as the **sample size**, and the estimate of the population parametric value is the "sample statistic." The average longevity of a sample of individuals of species A after exposure to pesticide A would be an estimate of the parametric value of that average if we could measure the longevity of every individual of species A.

Once we have sampled from the statistical population, we will have a set of longevity values. However, not all individuals that we measure will have the same response to our treatment; some will live longer than others. There will be some inherent variation between individual measurements within our sample. We might want to know how variable our set of observations is, so that we know what to expect if we are going to use the pesticide to try to kill individuals of species A. We might also want to know about variability if we want to compare the mean longevity of species A with that of species B.

There are two basic quantities in "parametric" statistics. The **mean** (I am going to use the symbol $\ddot{Y}$ for the mean) is a measure of "central tendency," which is calculated by summing all of the observations and dividing this sum by the number of observations [$\ddot{Y} = \sum Y/n$, where Y is any particular observation, $\sum Y$ is the sum of the observations ($\sum$ means "sum"), and n is the number of observations or sample size]. The **variance** is a measure of dispersion, or of the average difference between individual observations and the mean. It gives us an idea about how "spread out" or variable our observations are. The variance is calculated in two steps. First, we find the "sum of squares." We find the difference between each observation and the mean, square this difference, and then sum over all of the observations. The sum of squares is symbolized $\sum y^2$. In theory, $\sum y^2 = \sum(Y - \ddot{Y})^2$, but in practice, it is easier to compute it as $\sum y^2 = \sum Y^2 - (\sum Y)^2/n$ (these two equations are mathematically equivalent). In this equation, $\sum Y^2$ is the sum of the squared value of each observation, and $(\sum Y)^2$ is the sum of all of the observations, quantity squared. Modern pocket calculators will compute both $\sum Y$ and $\sum Y^2$ in the same operation (some will even calculate the mean and variance automatically). Once the sum of squares has been calculated, the variance is simply the sum of squares divided by the degrees of freedom, n-1. Symbolically,

$$s^2 = \sum y^2/n\text{-}1 = \frac{\sum Y^2 - (\sum Y)^2/n}{n\text{-}1}$$

The square root of the variance is a quantity called the standard deviation. This quantity is often reported in publications. It has the advantage of being easier to interpret than the variance, because its value is usually of the same order of magnitude as the mean. The standard deviation is another measure of dispersion. In any set of normally distributed, continuous quantitative data, approximately 68% of the observations lie within one standard deviation of the mean, and approximately 95% lie within two standard deviations of the mean.

Another purpose of statistics, indeed, for our purposes the most important one, is to make decisions about whether two or more samples could have come from the same statistical population. Without knowledge of the parametric mean of a population, we use sample means to estimate the parametric mean. We are often faced with the question "are these means different?" There are a variety of ways of answering that

question, depending on the nature of the data and on the number of means we want to compare. In statistics, we test a "null hypothesis" which is that two means are equivalent in value. The alternative hypothesis is that the means are unequal. Statistical tests are used to test the null hypothesis, and the result of the test allows us to either accept the null hypothesis (the means are equal) or reject the null hypothesis (the means are different). Because we are only estimating the values of the means (they are based on samples and not populations) there is the potential that our decision about the equality of the means is incorrect. If we reject the null hypothesis and decide that our means are not equal, it might be possible that if we took a larger sample, the value of the means would change, and they could in fact have been taken from the same statistical population. Alternatively, if we accept the null hypothesis and say that the means are the same, they might actually be different, but our sample sizes are insufficient to detect that difference. Statistics is a tool that allows us to make decisions in the face of uncertainty, but those decisions still carry a risk of being incorrect.

There are a variety of ways that we can minimize those risks. In general, a larger sample will result in a smaller variance, which will allow us to have greater confidence in our decision about either acceptance or rejection of a null hypothesis. Statisticians agree that a sample size of 30 for each mean is a minimum for parametric methods to apply with confidence. This is because most parametric statistical tests assume that the data to which they are applied are normally distributed (a "normal distribution" is the familiar bell-shaped curve). With smaller sample sizes, data sets may be badly skewed and not approach very closely a normal distribution.

Many novice students of statistics get confused by the meaning of the result of a statistical test. The most common type of result is a P (for "probability") value which theoretically ranges from 0 to 1, but which in statistical practice approaches but does not reach 0 or 1. We say this because statistics is the business of making guesses in the face of uncertainty, and probabilities of 0 and 1 imply statistical certainty (something never happens or something always happens). In biology, we almost never deal with certainties. What the P value tells us is the probability (based on our test) that two (or more) means are statistically equivalent. By convention, we reject the null hypothesis that two means are equal only if the probability that they are the same is less than 0.05 (5%). **If a statistical test results in a P value of 0.05 or less, we say that the difference between our means is "statistically significant."** This should all make more sense after a little practice.

What follows is a set of descriptions of several common statistical tests that we will use in this course, and a description of the proper way to present results of statistical tests. The following summary is designed to expedite your use of the full descriptions.

Summary of statistical tests used commonly in animal behavior studies

***t*-test:** Used to compare the value of two means. A special case of ANOVA, and can be used with equal or unequal sample sizes. Means and their associated variances must be calculated to use the *t*-test. We will use this test often, because many of our lab exercises make simple quantitative comparisons of two sets of data.

ANOVA (analysis of variance): Used to compare the value of three or more means. One need not calculate either means or variances in order to use ANOVA, as the important values come out of the computations automatically. This is a powerful statistical method, and drives much of the design of experiments. We will barely scratch the surface of the complexity of analysis of variance.

correlation: Used to determine to what extent two variables vary together. An increase in the value of one variable may be associated with an increase in the value of the other variable (a positive correlation); alternatively, and increase in the value of one variable may be associated with a decrease in the value of the other variable (a negative correlation). We will use this technique occasionally, for instance to compare body masses of worker ants with the mass of the cargo they carry.

regression: Used to determine to what extent an experimentally manipulated independent variable determines the value of an unmanipulated dependent variable. This is appropriate for experimental situations, while correlation is more appropriate for observational data.

analysis of frequencies: Used to detect similarities or differences in the frequency of occurrence of events; to detect patterns in such occurrences. Familiar tests such as chi square, which we will use occasionally, fall into this category.

Section 2: Specific statistical tests:

At this point, you probably have enough information to begin learning some actual statistical techniques. There are really only four or five commonly used statistical tests in animal behavior, and these are not really excruciatingly complicated. In fact, you do not have to remember how to actually calculate any of these statistics, because you have the luxury of a computer software package called Minitab that will run these tests (and many more) for you. The reason for going through the rigmarole below is because you will have a greater understanding of the meaning of the results of the tests if you know where they come from. The tests described below are the t-test, analysis of variance, correlation, and chi-square.

The ***t*-test** is used to compare the values of two sample means. It is necessary to compute variances as well as means to calculate t. This is a very commonly used statistic. The initial analysis involves the calculation of two quantities for each set of observations: the mean and the variance. These two quantities are used in the "*t*-test", which is a test for the difference between two means. When sample sizes are unequal, *t* is calculated as:

$$t_s = \frac{(\ddot{Y}_1 - \ddot{Y}_2)}{\sqrt{\frac{s^2_1}{n_2} + \frac{s^2_2}{n_1}}}$$

where $\ddot{Y}_1$ and $\ddot{Y}_2$ are the two means whose difference we wish to test, n_1 and n_2 are the sample sizes of $\ddot{Y}_1$ and $\ddot{Y}_2$, and s^2_1 and s^2_2 are the variances associated with $\ddot{Y}_1$ and $\ddot{Y}_2$. Our calculated value of *t* is compared with values in tables available in any statistics text. Each tabulated value of *t* is associated with a probability value, P. **These probabilities tell us how likely it is that our two means were sampled from the same statistical population**. In essence, what the *t*-test does is to express the difference between two means as a function of their variances. As the difference in two means becomes larger, or as the variances become smaller, *t* increases. Larger *t* values are associated with smaller probabilities. The smaller the probability, the more confident we are that our means are truly different. The "degrees of freedom" listed in the table is a value based on the sample sizes of the means and variances. In general, $\nu = n - 1$; more complicated formulae are used when the sample sizes are unequal.

We will not calculate $\ddot{Y}$, s^2, or *t* by hand, but it is good to know where these values come from. We will use Minitab to compute these values and to perform the *t*-test. When reporting the results of a *t*-test, one presents the means, associated standard deviations or variances, sample sizes or degrees of freedom, the value of *t*, and the value of P. See below for an example.

Analysis of variance (ANOVA) is used to compare values of three or more means. This technique tells us whether means in a series are different from one another, but does not necessarily tell which means are different from which others. There are more involved procedures to assess this.

Analysis of variance can be used with observational data (as we will do with our ant body masses) or with experimental data (for instance comparing the means of observations from treatment and control groups). Analysis of variance is a very powerful statistical tool that drives the design of experiments as well as the analysis of data.

The results of ANOVA are often reported as an "ANOVA table," with F_s and P being the most pertinent values. The means themselves do not automatically come out of the ANOVA computations, but these are typically reported as well (along with sample sizes and variances or standard deviations).

Correlation is used to assess the degree to which two variables change with respect to one another; how do the observations in two data sets vary together? This procedure is used for data sets in which we have neither controlled nor manipulated either of the variables. We might collect data on body size of female fruit flies and on the body size of the males that succeed in mating with them. If larger female fruit flies mate with larger male fruit flies, the variables "female body size" and "male body size" change in the same way, and would be "positively correlated". In contrast, if we collect data on mating frequency of individual male fruit flies and on population size or population density of males fruit flies, we might find that as population size increases, mating frequency decreases. In this case, there would be a negative correlation between population density and mating frequency. Correlation is generally used for observational data rather than experimental data. Regression (see below) is more appropriate for experimental data in which one variable is controlled or manipulated by the observer.

The "correlation coefficient" is a single value that describes the degree to which two variables vary together. It ranges from a value of +1 (which means perfect positive correlation) to -1 (perfect negative correlation). Values near zero indicate that two variables do not affect one another. The value of the correlation coefficient is compared with a set of tabulated values for different degrees of freedom. Usually such a table only includes the "critical values" of r associated with P values of 0.05 and 0.01. Thus we can determine whether our value of r indicates that there is a significant positive or negative correlation between our two sets of observations. In correlation, one reports r, N, and P.

Analysis of frequencies: The tests described above are what we call "parametric" statistical tests and are generally used with continuous quantitative data. However, not all quantitative data are continuous; some are meristic or discontinuous. If you have had a genetics course, you counted fruit flies (probably until your eyes were bugging out) to detect the frequencies of various mutant forms produced by different types of crosses. To determine whether patterns exist in this type of data, we need a tool to detect differences in frequencies of events or observations. Chi-square is one of those tools; it is used to assess whether a particular set of observations conforms to an externally generated set of expected values.

The computation of Chi-square (symbolized χ^2) is quite simple. We expect a certain frequency of observations to fall into each of our categories, and we have a certain number of actual observations in each category. Chi-square is computed as:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o is the frequency of actual observations in each category, and f_e is the frequency of observations that we expect in each category. Chi-square is the sum over all categories of "observed minus expected squared over expected". The value of chi-square increases with increasing difference between observed and expected values, and the associated values of P decrease with increasing χ^2. Thus the difference between categories becomes more significant with increasing value of χ^2.

One of the common uses of Chi-square is to assess whether the observed sex ratio of a population is male-biased or female-biased. For instance, censuses of a population of the cottonwood borer (a large and spectacular beetle) resulted in 230 sightings of males and 359 females. With a total of 592 individuals we would expect there to be 296 males and 296 females (this is our externally generated expectation). We calculate chi-square as $(230 - 296)^2/296 + (359 - 296)^2/296 = 28.125$. This value of χ^2 is associated with a probability value of less than 0.001, which means there is a very small chance that our sample comes from a population with equal numbers of males and females.

Section 3: Presenting and interpreting results:

The result of each of the tests described above is a value of the test statistic and a probability value. The test statistic is a value that is compared with a set of values (generally in a table) that are associated with specific probability values. The probability values (often from 0.9 in increments to 0.001) refer to the probability that the two means (or other statistics) you are interested in come from the same statistical population. In other words, a probability value of 0.25 means that there is a 25% chance that your two sample means are equal. Would we be comfortable rejecting the null hypothesis that two such means were sampled from the same statistical population? Statisticians feel that such a high probability of making an erroneous decision is unacceptable. By convention, we require a probability value of 0.05 (the two means have a 5% chance of being equal) before we are willing to accept the decision that they are different. This probability is referred to as the statistical "significance" level, and we say that two means are "statistically significantly different" only if the probability that they are equal is 5% or less.

When reporting the results of a statistical analysis, the standard procedure is to report the values of the means, the associated variances (or standard deviations) and sample sizes (or degrees of freedom), the value of the test statistic, and the probability (P) value. This information allows your readers to evaluate the statistic for themselves, to check the validity of your result. Believe me, reviewers, editors, and instructors do check!

Some kinds of results are best reported in text rather than in a table or figure. For results such as the sex ratio values described above, you would simply write a sentence presenting the numbers, the sex ratio, the chi-square value, and the P value:

"We counted 592 beetles (230 males and 359 females). The population sex ratio was 0.65, which is significantly different from unity ($\chi^2 = 28.1$, $P < 0.001$)".

For relatively simple cases of two means compared with a t test, results may be most simply reported in text. For example, in a study of territorial behavior in a damselfly, one could compare mean perch times of individual males who leave their perches spontaneously versus those that are disturbed by conspecific males. In presenting the results you would simply say “Mean perch time did not differ significantly between males who left their perches spontaneously and those that were disturbed by conspecifics (spontaneous mean = 118.6 sec, N = 37, s = 15.6; disturbed mean = 111.4 sec, N = 42, s = 19.5; t = 1.81, P = 0.074).”

More complex or extensive results are often reported in a table. For instance, if you wanted to report several pairs of means and associated variances, along with results of t-tests comparing the means, a table would be appropriate. Data on sexual size dimorphism in the cottonwood borer might be presented in a table like this:

Comparison of elytron length (EL) and body mass (BM) of males and females of *Plectrodera scalator*. [a]Mean ± SD; [b] (sample size)

	males	females	t_s	P
1991 EL(mm)	22.49 ±2.21[a] (190)[b]	25.76 ±1.31 (203)	17.98	<0.001
1992 EL(mm)	22.01 ±1.16 (205)	24.91 ±1.42 (197)	22.44	<0.001
1992 BM(g)	1.44 ±0.21 (123)	2.19 ±0.89 (127)	7.26	<0.001

The layout of a table is variable, and sometimes dictated by personal preferences of the author (or editor), but in general it is good to have the means that are being compared (with associated variances and sample sizes) arranged so that they are read across the table rather than up and down, and so that the sample statistics and probability values are placed in the same row as the means. Some editors use asterisks to denote P values, with one asterisk being $P < 0.05$, two asterisks meaning $P < 0.01$, and three meaning $P < 0.001$. The sample statistics are not always reported (due to space considerations), but the means, variances, and sample sizes must be so that the reader can reconstruct the statistical test.

When presenting results such as these in a table, the tendency of inexperienced scientific writers is to begin a paragraph about the results with a statement like "The data are summarized in Table 1." **DO NOT DO THIS!** Always describe (clearly and forcefully) in writing the important results, and refer the reader to the table. For instance, when I describe these data, I would (and did) write "Females were significantly larger than males, both in elytron length and in body mass (Table 1)."

Some types of results are best presented graphically. Histograms and line graphs are the most common in the animal behavior literature. A histogram is most appropriate when the data set is a series of means from categories (such as different species) rather than a series of means from treatment groups (such as activity levels at different temperatures). Line graphs are more appropriate for the latter.

A histogram typically consists of the set of categories along the abscissa and values of means or of frequencies along the ordinate. For instance, if we wanted to visualize the average distance between foraging individuals of several shorebirds, we might line up the species along the abscissa, and put the distance (in meters) on the ordinate. Such a histogram would allow the reader to see both the differences (or similarities) between species and the extent of those differences. It is also customary in histograms to include "error bars" at the tops of the columns that indicate the variances of the means, and to include the sample sizes along the abscissa or within the columns. A fairly recent trend in some publications (although not usually original scientific ones) is to use fancy graphic programs that generate histograms that look three-dimensional. This is unnecessary and is in fact misleading, because it implies that there is a third variable presented in the graph (thickness in addition to height and width of the bars), which is not actually present.

In a line graph, we usually place the treatment groups (or values of the independent variable) along the abscissa (for instance temperatures) and the values of the dependent variable along the ordinate. Running velocity of foraging seed harvester ants could be measured at $10^{o}C$, $15^{o}C$, $20^{o}C$, $25^{o}C$, and $30^{o}C$. These would be our treatments or independent variables, and mean running velocities would be plotted in the body of the graph. Error bars are often placed with the means in this type of graph as well. Sample sizes may be reported in the figure caption or placed in the body of the graph if there is room.

If we were to collect data on running velocity of several ant species at our various temperature values, we could place all means on the same graph, using different symbols (circles, squares, triangles, etc.) for the different species. Such a graph allows the reader to easily compare our data sets. For complex data sets, graphs are often very "busy". Be careful not to try to present too much information in a single figure. Include data that need to be compared, but separate different data sets into their own figures. With practice, authors can strike an appropriate balance of results in text, in tables, and in figures.