Data Presentation and Analysis
Adapted by Anne F. Maben from *"Field & Laboratory Methods for General Ecology"* by Zar, Brower & von Ende

Introduction
The whole point of statistics is to analyze data. Statistical methods allow a scientist to:
1. *Quantitatively* describe and summarize data;
2. Draw conclusions about large sets of data (from habitats, communities, or biological populations) by sampling only small portions of them; and
3. *Objectively* measure differences and relationships between sets of data.

Basic to statistical procedures are the concepts of a statistical *population* and a statistical *sample*. A statistical population is the entire set of data about which we wish to draw conclusions, and a statistical sample is only a portion of the statistical population. If a sample is taken at *random* from an entire population, then conclusions may be drawn about the population with a known chance of error, even though only a small portion of it was measured.

Averages
A very useful measure of the average of a population is the **mean**, and a *population* mean may be estimated by taking the mean of a *random sample* from the population (n). To find the mean ($\bar{\chi}$), just add up (Σ) the values (x = a grasshopper weight, a tree height) and divide by the number of values (n):

$$\bar{\chi} = \frac{\Sigma n}{N}$$

Measures of Variability
Calculating a mean gives only a partial description of a set of data. Notice that each of the following two samples of data has the same mean (namely 11): 1, 6, 11, 16, 21 and 10, H, 11, 11, 12. So, to help describe these samples we also need a measure of how variable, or how dispersed, the data are. One measure of data dispersion is the **range**, simply the difference between the largest and smallest values.

Of special importance and use in statistical analysis are those measures of data dispersion based on the deviation of data from their mean, the **standard deviation** (abbreviated **SD**).

The sample mean and standard deviation are very useful in describing the population of data from which our sample came. Most electronic calculators can compute SD .

Confidence in Estimating Population Means
When we calculate a sample mean, we may wonder how precise the estimate of the population mean is. We know that repeated samples from the same population will each have a somewhat different mean. The variability among these possible sample means is a very important statistic known as the standard error (SE).

Using the standard error, one can express a **confidence interval**, an interval that, with a stated level of confidence, may be said to include the population mean. For normally-distributed data, a specific kind of symmetrical, bell-shaped distribution of measurements, we will use a t-test to calculate our level of confidence in our data. A significance level of 5% is most frequently used in biological research, for this convention allows for a reasonable balance between the kinds of errors inherent in statistical testing (although significance levels of 1% and 10% are occasionally employed).

Comparing Statistical Populations
One of the most common of biostatistical procedures is drawing conclusions about the similarity or difference between the means of sampled populations of data. For example, and ecologist might wonder whether on the average the plant biomass is the same in two different geographical areas (or in two different seasons, or under two different experimental regimes). The question refers to two statistical populations, and a completely confident answer would require the impractical and probably impossible measurement of the biomass of all plant material in each area. Therefore, you take a sample from each of the two populations and then infer from the two sample means and variability whether those populations have the same or different means.

Two-Sample "t" tests
Statistical analysis for a two-sample experimental design is commonly done by a type of "t-testing," where the statistic "t" draws conclusions about the similarities or differences between two means. We wish to conclude whether the mean plant biomass in one geographic location is the same as the mean in the second location. Small "t" values indicate high probability that the two population means are the same; by contrast, large "t" values

imply low probability. "T" values are most easily calculated by a statistical calculator or in an MS Excel spreadsheet. You use degrees of freedom to compare the experimental value you calculate from your data comparison (t_{calc}) with the critical value (t_{crit}) found on a table similar to a Chi square table.

In more formal terms, the statistician's **null hypothesis** (H_O) is a statement that the means of the two populations are the same and the **alternate hypothesis** (H_A) is that the two population means are <u>not</u> the same. If **t_{calc}** is <u>*at least*</u> as large as **t_{crit}**, then the null hypothesis is rejected, and the alternate hypothesis is considered to be true.

Nonparametric Testing
There is statistical testing for populations that do NOT follow a normal distribution (and includes most populations in the wild.) These are called nonparametric statistics. One of the most commonly used is the ***Mann-Whitney test***, which tests for differences between two populations of data by examining a sample of data from each population. Nonparametric methods like the Mann-Whitney test can be used in instances where t-testing is inappropriate. Consider the following hypothetical data: five bottom grabs from a pond result in 11, 14, 15, 11, and 8 worms/ 0. 1 M2; seven bottom hauls from a second pond revealed 9, 13, 7, 10, 6, 7, and 11 worms/0.1 M2 . The Mann-Whitney test allows us to ask whether both areas have the same worm density; the null hypothesis is "H_o: The two populations of worms have the same density," and the alternate hypothesis is "H_A: The two worm populations have different densities."

The Mann-Whitney test puts each data set in numerical order and then assigns a "rank" each data point. We will be using a statistical package to calculate the "critical value at a 95% confidence limit. If the calculated value of U1 or U_2 is greater than or equal to the tabled value, then H_0 is rejected and H_A is declared true. For our example, the two sample sizes are 5 and 7, so the critical value for testing at the 0.05 significance level is 30. As neither U_1 nor U_2 is as large as 30, the null hypothesis, H_0, is not rejected, and it is concluded that the two samples of worms came from populations having the same densities.

Goodness of Fit   *(Chi Square – see separate sheet)*
Ecological data often take the form of a distribution of frequencies of occurrences, in which case we may wish to ask whether the observed distribution is significantly different from some hypothesized distribution. For example, let us assume that we have determined that the bottom of a section of stream is 50% sand, 30% gravel, and 20% silt. We have further observed that for a certain species of fish, 8 individuals were in the sand areas, 18 were in the gravel portions, and 4 were in the vicinity of silt. A typical null hypothesis would be that the fish have no preference among the three substrates-that individuals of this species distribute themselves in the stream without respect to substrate type. The alternate hypothesis is that the fish are not distributed independent of substrate, but that they do show substrate preference.

If the null hypothesis were true, and the distribution of fish is independent of substrate type, then 50% of the total number of fish (15 of the total sample of 30 fish) would have been expected to be in sand areas, 30% (9) in the gravel, and 20% (6) in the silt.

Note that the chi-square calculation uses <u>frequencies</u> only; it never uses percentages or proportions.