Group Projects: Data Analysis and Statistics

Statistics allow an ecologist to do three important things: 1) quantitatively describe and summarize characteristics of data, 2) draw conclusions about a habitat, community, or population from samples of them and 3) objectively assess differences and relationships between sets of data.

A. Basic Data Analysis

1) Each group should have one computer. Enter your data into an Excel spreadsheet as follows:

Treatment	Replicate number	Starting #leaves (Day 0)	#leaves day ____	# leaves day ___	Etc...	Notes and observations

2) Construct a graph of the mean #/leaves/cup over time. Ask your instructor if you need help using the graphing functions in Excel.
3) Calculate r from beginning to end for each of your cups. Refer to the handout for the pilot study if you do not remember how to calculate r.
4) When all groups get to this point, your instructor will show you how to calculate the mean, standard deviation, and standard error for each of your treatments using Excel. If you are unfamiliar with these terms, read section B. below.
5) Graph the mean and standard error for each of your treatments.
6) Use appropriate inferential statistics to draw conclusions about differences among treatments. See you instructor for help with these.
7) Talk to the members of your group about how best to present the data in a presentation.

B. Descriptive statistics: (summarized from Brower et. al, 1990).
We will use descriptive statistics to assess the central tendency and variability in the experiments you conducted with *Lemna*.

A measure of the central tendency or average of a group is the **mean** (the mean is also referred to as the average). The mean of a sample is calculated as:

$$\overline{X} = \sum X \big/ n$$

where $\overline{X}$ (pronounced "X bar") is the conventional symbol for the mean, $\sum X$ is the summation of all values of *X* in the sample, and *n* is the number of data points in the sample.

Calculating a mean or other average gives only a partial description of a data set. For example, the following two samples have the same mean (11): 1,6,11,16,21 and 10,11,11,11,12. To help describe these samples, we also need a measure of how variable the data are.

One measure that is very useful in describing variability is the deviation of the data from their mean. To calculate the **standard deviation**, we first need to calculate the *sum of squared deviations from the mean*, referred to simply as *sum of squares* (abbreviated SS).

$$SS = \sum (X - \overline{X})^2$$, which is mathematically equivalent to $SS = \sum X^2 - [(\sum X)^2]/n$

The *sample variance* is

$$s^2 = SS/DF,$$

where DF is a quantity called *degrees of freedom*. For sample variance, DF = *n* – 1. The sample *standard deviation* (abbreviated *s*, or SD) is:

$s = \sqrt{s^2}$

Standard deviation is generally reported as a measure of variability in preference to the variance because it has the same units as the original data.

Another useful value to know is how much our estimate of the mean would vary if we were to repeatedly sample a treatment (i.e. had an infinite number of cups to work with that we repeatedly took a subset to count and calculate the mean). The *standard error* (denoted SE) is simply:

$SE = s/\sqrt{n}$

C. Inferential Statistics

We will use inferential statistics to draw conclusions about differences in your treatments. These statistical tests allow us to measure the probability of obtaining differences between groups simply due to chance events, not true biological effects. These conclusions are based on the (infamous) *p-value* that you may have heard of.

We commonly use the value $p < 0.05$ as a cutoff for differences that are consistent with biological effects. In other words, if the chance of getting our observed difference between groups is less than 5%, we view this result as consistent with an effect of our treatment.

On the other hand, if an inferential statistical test gave a $p > 0.05$, this means we would expect to see the observed amount of difference between groups more than five percent of the time simply due to sampling error and other random effects. In other words, there is little support for treatment effects with this result.

Because the design of each project varies, you will need to consult with you lab instructor about the appropriate type of test to use. These may include t-tests, ANOVA, and chi-square analyses.

Cited references:

Brower, J.E., J.H. Zar, and C. V. von Ende. 1990. *Field and laboratory methods for general ecology*. Wm. C. Brown, Dubuque, IA.