DATA PRESENTATION AND INTRODUCTION TO INFERENTIAL STATISTICS
BI 311
Homework 04

Introduction:

You have begun the exploration of quantitative ecology with a simple **comparative study** or "natural" experiment. You and your group members have designed and are now implementing a field study. The exercises in this homework will help you determine how to begin treating your data.

Most ecological experiments involve making measurements of two or more groups of subjects and then graphically or statistically comparing the data. Making statistical comparisons between groups of subjects requires describing the type of variable the investigator will measure, defining a random sampling scheme, determining the number of samples needed to adequately measure the variables of interest, and selecting an appropriate graphical/statistical analysis. We have already completed homework that introduced concepts of defining variable types, implementing sampling schemes and deciding on an appropriate sample size. The following exercises will help you display your data in figures and tables, and to decide on an appropriate statistical analysis.

Skills:

1. Learn how to use WORD to make data tables appropriate for scientific publications.
2. Learn how to use EXCEL to make bar graphs for presenting categorical data.
3. Learn how to use a decision tree to select an appropriate statistical test.

Materials:

- A Handbook of Biological Investigation, Ambrose et al. 2007
- Access to WORD and EXCEL

EXERCISE 1-CONSTRUCTING TABLES USING WORD

Background:

Scientific communications present key findings in illustrations that take one of two forms: tables and figures. As the name implies, tables consist of tabular data presented in rows and columns. Tables usually present numerical data that would be too tedious to write out in the text body of a paper or that cannot be summarized in a figure. All tables should have a title at the top that starts "Table 1." Each scientific journal or conference will have its own specifications for the construction of tables and figures. As an example, we will use the style guide for authors submitting articles to the scientific journal EMERGING INFECTIOUS DISEASES.

Procedure:

1. Read chapter 12 of Ambrose et al. (2007) for samples of both tables and figures and general tips illustrating data.
2. Visit and briefly read the formatting instructions for tables and figures for the journal EMERGING INFECTIOUS DISEASES available at the following links: http://wwwnc.cdc.gov/eid/pages/tables.htm
3. Watch my brief video on creating a table in WORD available at the following link: http://www.screencast.com/t/XokgI7Ye
4. Open the example data available in the EXCEL file named "Homework 4 Data" available on MOODLE.
5. On the first sheet (named "Table Data") of the EXCEL file are mosquito data for 3 counties in Montana. Use the table tools in WORD and the formatting requirements of the journal EMERGING INFECTIOUS DISEASES to recreate the table you see in the image below. Be sure to add a table title and abbreviation footnotes. For abbreviation definitions: LTI = mean light trap index = number of sorted *Cx. tarsalis* divided by the number of traps, and MIR = minimum infection rate = number of positive samples divided by the number of *Cx. tarsalis* tested and then multiplied by 1,000.

Year	Sheridan County		Blaine County		Lake County	
	LTI	MIR	LTI	MIR	LTI	MIR
2005	210.4	1.4	86.0	0.0	39.8	0.0
2006	131.0	1.3	245.0	3.6	27.3	0.0
2007	334.0	6.0	185.0	7.6	44.7	0.0
2008	62.6	3.2	113.3	1.1	50.1	0.0
2009	197.5	1.0	152.7	0.7	25.8	0.0
2010	380.0	1.1	75.1	0.0	37.0	0.0
2011	1,690.0	0.0	204.6	0.0	46.9	0.0
2012	21.6	5.8	80.8	6.2	21.6	0.0

6. Use the following checklist to ensure your table is complete:
    a. Is the table less than 17 cm wide?
    b. Did you use 8 point Arial font?
    c. Do you only have horizontal rules (lines) under the column headings?
    d. Did you eliminate vertical lines?
    e. Is the title brief but complete?
    f. Did you only capitalize the first word in the title?
    g. Did you use a footnote to explain any abbreviations that appear in the title or elsewhere?
    h. Does every column have a column heading?
    i. Did you include units of measure in the column headings?
    j. Did you use a footnote to explain any symbols?
    k. Did you organizing your footnotes using the following indicators in order *, †, ‡, §, ¶, #?
7. Save your WORD file to submit as part of your homework (see below).

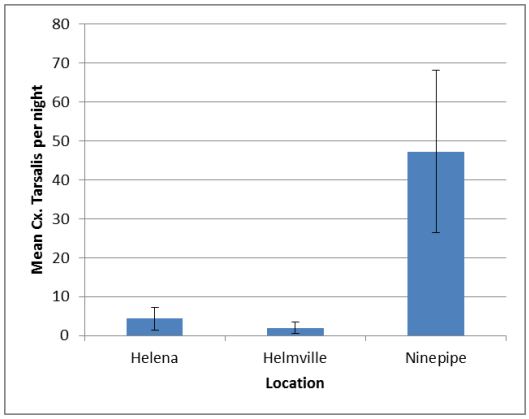
EXERCISE 2- USE EXCEL TO FORMAT FIGURES FOR PUBLICATION

Background:

Figures include maps, diagrams, and graphs that illustrate data trends. Because we humans are visually oriented, we intuitively understand data presented in figures much better than data presented in tables. Thus, always present data in figures when you have an option between a figure and table. Figures have a title (sometimes called a legend) at the bottom of the figure starting with "Figure 1."

Laboratory Procedure:

1. Watch my brief video on making an EXCEL bar graph available at the following link: http://www.screencast.com/users/Hokit/folders/Science%20Communication/media/0f7e3287-febc-42a9-970e-c0465d000656
2. Re-open the "Homework 4" EXCEL file and move to the second sheet titled "Figure data". Recreate the figure you see below (with corrections) using the chart tools in EXCEL.



3. Use the following figure checklist to ensure your figure is complete:
    a. Did you include both the figure (on a separate worksheet) and the original data?
    b. Is the figure at least 8.5 cm wide?
    c. Did you use Calibri or Arial font in at least 12 pt?
    d. Did you italicize the species name?
    e. Did you include a figure legend (title) at the bottom of the figure?
    f. Did you omit unnecessary boxes, borders or horizontal lines in the figure? E.g. did you eliminate the horizontal lines in the above figure?
    g. Did you rotate the y-axis label to be parallel to the y-axis?
    h. Did you add error bars?
4. Copy your figure and paste it in the WORD file you used to create your table in exercise 1 above.

EXERCISE 3- USING A DECISION TREE TO SELECT AN APPROPRIATE STATISTICAL TEST

Background:

You have already been introduced to **descriptive statistics** including **measures of central tendency** (mean and median), **measures of variability** (range, variance, and standard deviation), and **measures of confidence** (standard error and confidence interval). Although useful for detecting trends and patterns, descriptive statistics are not always useful for making conclusions about the significance of observed trends and patterns. **Inferential statistics** allow you to make comparisons between data sets and/or relationships between variables and to make conclusions about the statistical significance of the differences.

Whenever data sets are statistically compared, a **null model** (AKA null hypothesis) is stated to the effect that there is no significant difference (or association in correlation studies) between the data sets. This is because, statistically, it is possible to disprove but impossible to prove a hypothesis. There is always, no matter how small, some chance that the observed results were due to chance. **P-values** are a statistical estimate of the probability that the observed results are due simply to chance and not due to "real" differences/associations between the data sets.

If data sets are statistically different (or associated) then we reject the null and accept the alternative that there is a significant difference/association between the data sets. However, making such a conclusion comes with the risk that we may be making an error. There are two types of errors in hypothesis testing. **Type I errors** occur when we reject the null hypothesis in favor of the alternative when, indeed, the null is true. This is an error that scientists do not want to make. When we test for differences/associations between data sets, we want to be able to conclude that any detected differences/associations are real and not due to chance. By convention, we calculate inferential statistics such that when we reject the null, there is less than a 5 percent chance that we are wrong. **Type II errors** occur when we accept a null that is actually false. Type II errors can occur when sample sizes are too small or the measured variable has too much variation (dispersion).

Typically when we use inferential statistics we have measured a variable we call the **dependent variable** that may or may not be influenced by **factors** (or **independent variables**). We often make comparisons between data sets that differ with respect to their independent variables and we wish to know if this difference influences the dependent variable. For example, we could test whether upstream versus downstream location on a river (independent factor = location) influences one or more measurement variables (e.g. dependent variable = stream flow rate). Note that, in this case, the independent variable is a nominal variable (upstream versus downstream) while the dependent variable is a continuous, ratio variable (e.g. 100.5 CMS). This is not always the case. Some independent variables may be continuous, some dependent variables discrete, and any combination of independent and dependent variable is possible. Also, a nominal (sometimes called categorical) variable can have two or more levels (or groups). For example, the location variable has two levels, upstream and downstream. This is important because the number and type of independent and dependent variables will partially determine which statistical analysis is appropriate.

Not only are the type and number of dependent and independent variables important in determining which statistical test to use, the **frequency distribution** of the dependent variable is critical. Parametric data analyses are powerful statistical tests that are often used by biologists. However, **parametric tests** have several assumptions, the most common of which is that the data (continuous dependent variables) are normally distributed. These tests are considered "robust" to small deviations from normality and often data can be transformed to meet parametric assumptions. If not, **nonparametric tests** can be used, although they are generally considered less powerful. The **power** of a test lies in its ability to detect differences between populations without committing a type I error.

So how do you decide which type of inferential statistic to use? Using a 'decision tree' can help decide which test is most appropriate.

Laboratory Procedure:

Read Chapter 7 of Ambrose et al. (2007) to further examine the statistical concepts described above. Then move on to Chapter 8 and page 50 to use a decision tree to decide which statistical test is appropriate for the following list of studies.

Study A- You wish to test for a difference between mean snout-vent-length (SVL) of male versus female scrub lizards. Because lizards cannot be both male and female, you know that your data are independent of each other. However, your descriptive statistics tell you that your data is not normally distributed. Which test should you choose?

Study B- You wish to test for a correlation (association) between scrub lizard size and age. Which test should you use?

Study C- You wish to test for differences between mean SVL of hatchling, juvenile, and adult scrub lizards. Preliminary tests reveal that the data is normally distributed. Which test should you use?

Study D- You wish to test for differences between the distributions in the presence/absence of scrub lizards versus habitat patches with sandy/non-sandy soils. Which test will you use?

HOMEWORK SUBMISSION

Each person should create a WORD file named *Homework 4*, answer the following questions, and submit their file to me via email.

1. Exercise 1: Include your completed table.
2. Exercise 2: Paste a copy of your completed figure.
3. Exercise 3: List which type of statistical analysis you would use for the hypothetical studies A-D.