## Presentation of data and graphics

## Introduction

In most biology studies, large amounts of numerical data are collected. Several methods have been developed to organize and present these data. Large amounts of data are generally presented either as tables or in some type of graph. Either of these methods allows clear and concise communication with the reader. Understanding these methods of data presentation will help you present data in writing assignments and also will aid in your understanding of models presented in class. This lab will introduce you to tables and several types of graphs. You will be expected to understand the conventions presented in this lab and to follow these conventions in all papers and assignments for this class.

When it is necessary to provide the reader with several numbers, a table is a convenient way to organize your information. Tabular presentation allows the reader to easily find the data values. One disadvantage of tables is that they do not lend themselves to composing a mental picture of the trends occurring.

Graphic presentations are used to show trends or patterns and assist visualization. There are several different types of graphs which are commonly used in animal biology. There are some characteristics which are common to most of these types of graphs. In general the independent variable will be plotted on the horizontal axis ( $x$, or abscissa) and the dependent variable is plotted on the vertical axis ( $y$, or ordinate). It is important to always clearly label the axes of every graph and indicate what units are being plotted. If more than one type of symbol is used, a key is necessary to show what each symbol indicates.

Whether data is presented in a table format or in a graph it is necessary to indicate to the reader what data is being presented. For a table the title is at the top of the table. For graphs the legend is located below the graph. Think of this as the "caption" for the figure it gives most of the information, the key just indicates the symbols (look at page 7 for a labeled example). In either case the title or legend should be as informative as possible. It should be possible to read the title or legend and understand 1) what is being presented, 2) its statistical significance and 3) its importance without reference to the text of the paper. Each table should be numbered consecutively in the order of presentation and the title should begin with "Table $x$ " Graphs are also numbered consecutively and are called figures. The legend for each figure begins with "Figure $x$ ".

## Procedures

This skills exercise consists of examples of several types of graphs and tables and a number of exercises for you to complete. You will need to complete all of the assignments and turn this exercise at the beginning of the next class. This is an individual activity, not a group activity.

We expect you to prepare figures and tables in these formats throughout this class and your biology degree so be sure to refer to this exercise when composing tables and graphs for your research paper and laboratory analyses.

## Tabulating Data

Below is an example of standard table format. There are no vertical lines! There must be a complete horizontal line separating the title from the table, a horizontal line separating the column headings from the data, and a horizontal line at the end of the table. A partial horizontal line may be used to group several columns under the heading. If data are missing, a dash (or N/A) will be recorded in the cell. Units of measure may be indicated either in the table of sometimes in the title.

Table 1. Population estimates for collared lizards at two sites in Oklahoma over two years.

|  | Wichita Mountains |  | Glass Mountains |  |
| :--- | :--- | :--- | :--- | :--- |
|  | 1992 | 1993 | 1992 | 1993 |
| Sex |  |  |  |  |
|  | 25 | 37 | 41 | 45 |
| Males | 21 | 38 | 39 | 40 |
| Females | 46 | 75 | 80 | 85 |
| Total |  |  |  |  |

## Exercise: Table of Blood Groups

Different races and cultural groups of people in the world have differing percentages of the blood types. Caucasians (Scottish) have 53\% type O and $32 \%$ type A. They have $12 \%$ type B and $3 \%$ type AB. Chinese have $10 \%$ type AB and $25 \%$ type A. Type B is most common at $34 \%$ and the rest is type O. Navajo Indians only show two major blood types; $75 \%$ are type O and the rest are type A. Australian Aborigines also only exhibit these two blood types; however type A is most common at $52 \%$. The Baka people have all four blood types, with $31 \%$ type O, $30 \%$ type A, $29 \%$ type B and the rest type AB.

Tabulate this information. Be sure to include a title above the Table.


Figure 1. Head width of male collared lizards (Crotaphytus collaris) from Oklahoma as a function of body size (snout-vent length).

The line drawn through the data points in this graph is not drawn by guesswork. This line is the result of a mathematical process known as regression, which minimizes the errors between the line and all of the data points. Unless you have used the regression process to mathematically estimate the line that best fits the data it is inappropriate to draw a line on your graph. It is also inappropriate to "connect the dots." The $\mathrm{R}^{2}$ value indicates how closely the relationship between the X and Y variables is represented by the line. $\mathrm{An} \mathrm{R}^{2}$ of 1.0 is perfect correlation.

A major use of graphs and statistics is to understand the relationship between two variables. Often we try to predict one of the variables, the response variable, from a second variable, the explanatory variable. The explanatory variable is part of our experimental design, the response variable is what we measure and hope to predict. Usually when you are asked to "plot variable A as a function of variable B", variable A will be the response (dependent) variable and will be placed on the Y axis. One possible exception is when the explanatory (independent) variable is some value which actually varies vertically (height, altitude, depth). In this case the axes may be changed to provide a clearer picture of the pattern of the data. Notice the legend for the figure is located below the figure.

## Correlation/Causation

Science is aimed at understanding how the world works. Identifying the causes of events is a key part of the process of science. A first step in that process is the discovery of patterns or associations between events. When two events are associated it raises the possibility that one is the cause of the other. A correlation between two variables is a powerful clue that there may be a causal relationship between those two variables. For example, taking streptomycin is associated with reduced bacterial infection. However, two variables can be associated, or correlated without one being the cause of the other. The two variables might both be due to a common cause. For example, data shows that the number of violent crimes tend to increase when ice cream sales increase. Clearly one of these is not the cause of the other, rather both of these variables increase with hot temperatures. Hot weather encourages both ice cream consumption and bad behavior.

In science, you often hear these ideas expressed as "correlation does not require causation". Please keep this in mind as you work through this packet, a correlation between two variables is a clue, not definitive evidence that one variable causes the other. Establishing cause requires a controlled experiment, not just a correlation.

Exercises: Plot the following data in the space provided and answer the questions.
An agronomist is interested in studying the effect of fertilizer on wheat yield. He adds different amounts of a new type of fertilizer to several plots of wheat and measures the amount of wheat produced. His data is below. Plot this data in the space provided. Be sure to write a thorough figure legend.

| Plot \# | Fertilizer (kg/ha) | Yield (kg/ha) |  |
| :--- | :---: | :---: | :---: |
| 1 | 10 |  | 201 |
| 2 | 25 | 260 |  |
| 3 | 50 | 302 |  |
| 4 | 75 | 330 |  |
| 5 | 100 | 350 |  |
| 6 | 125 | 363 |  |
| 7 | 150 | 392 |  |
| 8 | 175 | 405 |  |
| 9 | 200 | 417 |  |



What is the relationship between the amount of fertilizer and wheat yield?

Does this graph prove a causative relationship between fertilizer application and wheat yield? In your response, be sure you consider the difference between correlation and causation.

## Bar graph

A plain bar graph is used when the data values expressed have no relationship to each other except that they were the result of the same experiment or set of measurements. It is used to display the frequency distribution of a categorical variable. (That is, there is no mathematical or numeric relationship between the values of the independent variable.) The bars in a bar graph should be separate.

In an extension of the experiment concerning the effects of fertilizer on wheat yield, the agronomist tests the effects of his new fertilizer on several varieties of wheat. He applies the fertilizer at the rate of $100 \mathrm{~kg} / \mathrm{ha}$ to all varieties of wheat. He then measures the yield (in hg/ha) and calculates the increase in yield relative to plots of the same variety without fertilizer. The data below show the percentage difference in yield of the fertilizer plots relative to the unfertilized plots. Express these data as a bar graph.

| Wheat Variety | Percentage change in yield |
| :--- | :---: |
| 1 | $13 \%$ |
| 2 | $31 \%$ |
| 3 | $18 \%$ |
| 4 | $27 \%$ |
| 5 | $03 \%$ |
| 6 | $21 \%$ |



## Stacked Bar Diagrams (also called Mosaic Plots)

In a stacked bar diagram each bar is divided into portions. It is important to note that the bars are stacked on top of each other. Stacked bar diagrams are very useful to express differences in composition. Often the data for stacked bar diagrams are expressed as proportions or percentages so that the total height of all bars is the same. The portions of each bar should be stacked in the same order and a key should be included to indicate what each portion of the bar represents. These graphs allow easy comparison among groups.


Figure 2. Diet composition of four species of aquatic turtles (Trachemys scripta, Apalone mutica, Apalone spinifera, and Pseudemys concinna) in a riverine habitat. Composition of diet is expressed as a percentage of the total diet by volume.

Is reproduction hazardous? If not, then it is difficult to explain why adults in many organisms hold back on the number of offspring they appear capable of raising. Opplinger et al. (1996) investigated the impact of reproductive effort on the susceptibility to malaria in wild great tits (Parus major) breeding in nest boxes. They divided 65 nesting females into two treatment groups. In one group of 30 females, each bird had two eggs stolen from her nest, causing the females to lay an additional egg. The extra effort required might increase stress on these females. The remaining 35 females were left alone, establishing the control group. A blood sample was taken from each female 14 days after her eggs hatched to test for infection by avian malaria, a common parasite. The resulting data is shown in the contingency table below. Depict this data as a stacked bar plot (mosaic plot), remember to use the relative frequency of the events or percentages, not raw numbers.

Table 2. Contingency table showing the incidence of malaria in female great tits in relation to experimental treatment.

|  | Experimental Treatment Group |  |  |
| :--- | :---: | :---: | :---: |
|  | Control Group | Egg-removal group | Row Total |
| Malaria | 7 | 15 | 22 |
| No malaria | 28 | 15 | 43 |
| Column Total | 35 | 30 | 65 |



Is reproduction hazardous? Interpret your graph to answer the question.

## Arithmetic and Semilogarithmic plots

A logarithmic scale is a non-linear scale used when there is a large range of quantities being compared. Semilogarithmic means only one axis is logarithmic. Some types of ecological phenomena are more easily graphed and interpreted on a logarithmic or semilogarithmic plot. Functions can also be plotted with both axes on a logarithm scale.

Bacteria reproduce by binary fission and have very high rates of population growth. Suppose you start on day zero with one bacterium of a species which doubles its population size every 24 hours. Starting at day zero, graph the population size of these bacteria for 15 days on both arithmetic and semilogarithmic plots to compare their appearance.


If you draw a table or figure you must write a detailed legend!

## Histograms

A histogram is a graphical representation of a frequency distribution. It is especially useful for presenting a large number of data values. Although a histogram superficially resembles a bar graph they are actually very different. In a histogram the horizontal axis does not have a numerical scale. The bars in a histogram represent ranges of values on this axis. The height of each bar represents the number of values that fall within this range. The bars in a histogram touch. There is a certain amount of art in creating a histogram that clearly displays your data. Although there is no absolute right way to go about creating a histogram, below are some general guidelines.

You have captured a large number of small fish (Notropis baileyi) from a river and wish to show in a graph the sizes of fish in this population. First determine the range of data values. The next step is to decide how many bars you wish to display. Divide the range of the data values by the number of bars. This number will be the width of each bar. Usually you will round this number to some even value (e.g. 5 instead of 4.875). Pick a starting point just below the smallest data value. Add the bar width to this value. This will be the range for the first bar in your histogram. Then count the number of observations (fish in this case) that fall in this range. Proceed with this process until you have reached the bar that includes the largest data value. Note: You must arrange the intervals for your bars so that no data value can fall on the boundary between two bars.


Figure 3. Length-frequency histogram of Notropis baileyi caught in a Texas river.

Male fireflies of the species Photinus ignitus attract females with pulses of light. Flashes of longer duration seem to attract the most females. During mating, the male transfers a spermatophore to the female. Besides containing sperm, the spermatophore is rich in protein that is distributed by the female to her fertilized eggs. The data below are measurements of spermatophore mass (in ng) of 50 males (Cratsley and Lewis 2003). Create a histogram depicting the frequency distribution of the mass measurements.

| 2.5 | 2.8 | 3.0 | 3.3 | 4.5 | 5.5 | 8.5 | 1.2 | 3.5 | 3.3 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 3.1 | 1.5 | 1.4 | 4.7 | 5.2 | 6.1 | 7.4 | 3.8 | 3.9 | 4.2 |
| 2.4 | 3.5 | 3.8 | 3.9 | 4.8 | 5.8 | 7.2 | 7.7 | 4.1 | 3.9 |
| 3.5 | 3.3 | 2.1 | 1.8 | 3.9 | 4.2 | 6.4 | 4.1 | 3.8 | 3.7 |
| 2.3 | 2.0 | 1.6 | 3.3 | 4.5 | 4.9 | 8.2 | 3.9 | 4.0 | 4.1 |



Describe the shape of this histogram by referring to the figures on the next page.


Figure 4. Examples of some histogram shapes.

Table 2. Description of what type of graphical method to use with different types and numbers of variables.

| Single Variable | Type of data | Graphical Method |
| :--- | :--- | :--- |
|  | Categorical <br> Numerical (continuous) | Bar graph <br> Histogram, cumulative frequency <br> distribution |
| Comparison of <br> more than one <br> variable |  |  |
|  | Two categorical variables <br> One numerical and one <br> categorical variable <br> Two numerical variables | Grouped bar graph, mosaic plot <br> Grouped histograms, line plot, <br> cumulative frequency distribution <br> Scatter plot, line plot, map |

## Box and Whisker plots

Box plots and box and whisker plots are a useful way of visually summarizing the descriptive statistics of a set of data. They depict the data's distribution by plotting minimum, maximum, median, and first and third quartiles (percentiles dividing the data into fourths) of the data. They are very useful for visually indicating where the bulk of the measurements lie.
E.g.


Figure 5. Box-plot of the response of land snail species richness to burned vs unburned management type of all sampled sites showing median, first and third quartiles, and maximum and minimum values (Nekola 2002).

Draw a box and whisker plot for the following data:
Table 3. Running speed ( $\mathrm{cm} / \mathrm{s}$ ) of male spiders before and after self-amputation of a pedipalp.

| Descriptive <br> statistic | Before <br> amputation | After <br> amputation |
| :--- | :--- | :--- |
| Mean | 2.9 | 3.5 |
| Maximum | 3.7 | 5.5 |
| Minimum | 1.5 | 2.2 |
| $1^{\text {st }}$ quartile | 2.3 | 3.2 |
| $3^{\text {rd }}$ quartile | 3.1 | 4.9 |



## Heatmap

A heatmap (chloropleth map) is a graphical depiction of data where the individual values are represented as colors usually shaded in proportion to the data being represented. They are often used to depict genetic as well as geographical population data. Note: I made up this heatmap example data to make it simple - all the others in this packet are from real studies.


Figure 6. Heatmap showing change in mRNA concentration (picomolar) of five genes in response to incubation of sea star (Echinodermata) embryos in varying concentrations of NaCl . Average seawater has $35 \mathrm{~g} / \mathrm{L} \mathrm{NaCl}$.

What color indicates an increase in mRNA concentration?

Which genes are downregulated (have lower expression and therefore decrease in mRNA concentration) in the presence of higher than normal NaCl ?

Use the map of the southern part of Texas to make a useful geographic heat map depicting the current data (2016) on land snail species richness of each county. Decide on a useful and appropriate breakdown of the categories and be sure you include a Key to the colors being used for each category. I put a chunk of a Tx county map below so you can identify the counties.

Table 4. Land snail species richness of Texas counties from current museum records.

| County | Species <br> Richness |
| :--- | :--- |
| Brooks | 6 |
| Cameron | 55 |
| Hidalgo | 28 |
| Jim Hogg | 5 |
| Kenedy | 3 |
| Starr | 13 |
| Webb | 10 |
| Willacy | 29 |
| Zapata | 4 |



Why do you think Kenedy, Zapata, and Jim Hogg Counties appear species-poor?


Figure 7. Outline map of Texas counties.
If you draw a table or figure you must write a detailed legend!

