Biology 171L – Environment and Ecology Lab  
Lab 2: Descriptive Statistics, Presenting Data and Graphing Relationships

Introduction

Long lists of data are often not very useful for identifying general trends in the data or the significance of a particular treatment in affecting the outcome of an experiment. There are, however, statistical procedures that facilitate the summarization, presentation, and analysis of the data. These procedures may allow a scientist to look through the noise in the data to see major trends. The use of these procedures is particularly beneficial in biological science, where the variability in the data may be very great.

In this laboratory exercise you will learn how to calculate averages and standard deviations to summarize your koa haole seed pod data. In addition, you will learn how to graph data to summarize data trends and illustrate relationships.

Some Definitions

Statistics is the scientific study of numerical data based upon variation in nature. In statistics, we begin by making observations, or measurements. A single observation is a datum. A collection of observations is called data. The actual attribute being measured is called a variable. For example, a number of variables could be measured on a rat: weight, body length, tail length, sex, fur color, eye color, the number of whiskers on its snout, the number of toes on its right forefoot, aggressiveness, and health.

A statistical population represents the totality of individual observations for a variable about which we would like to make some inferences or generalizations. In this context, you should distinguish between a biological population and a statistical population. Thus our population could be the rat weights of the world or the tail lengths of the world. We could define our population more specifically, e.g., the rat weights of Honolulu, or the weights of rats fed a specific high protein diet. Note that we rarely collect data on the entire population.

Since it is impractical, not to mention costly, to collect data from an entire population, we usually collect a small subset, or sample, of observations from the population that hopefully represents the population. In general, a larger sample is more representative of the population than a smaller sample.

Descriptive Statistics

We can use statistics to summarize and describe the data. Thus we can give meaning to a long list of numbers. There are two basic types of descriptive statistics: statistics of location and statistics of dispersion. Statistics of location tell us about the central tendency of the data, i.e., where the center of the data lies. A familiar statistic of location is the average, or arithmetic mean. The average is calculated by summing up the values of the individual observations and dividing by the number of observations. Other statistics of location include the median and mode.

Statistics of dispersion tell us about the spread of the data, indicating their variability. One simple statistic of dispersion is the range, which is the difference between the minimum value and the maximum value. Note that the range is highly influenced by sample size. A statistic of dispersion that is not biased by sample size is the standard deviation, which can be thought of as the average deviation of the individual observations from the mean. A high value for the standard deviation would indicate high variability in the data. Calculation of the standard deviation will be demonstrated in class.

Graphing Data

Besides calculating averages and standard deviations to summarize a data set, it is often useful to graph the values. In this way, a picture of the data may be made. Sometimes graphs allow us to see trends in the data that would not otherwise be apparent.

One type of graph is the frequency histogram. In this type of graph, the horizontal axis represents the possible values for a particular variable. The vertical axis represents the frequency, or number of times, a particular value was observed. This frequency is plotted as a vertical bar. For example, if we observed five seed pods bearing 12 seeds per pod, then the bar height for this valued would be five units high.

Graphs may be drafted to illustrate relationships between variables. For example,
we may want to illustrate the relationship between weight and body length in rats. In this case, we could use the horizontal axis to represent the values for rat body length and the vertical axis to represent values for rat weight. The body length and weight of each rat could then be plotted. A line may then be drawn through the points to indicate the trend observed.

Presenting Tables and Figures

A table is a list of values arranged in columns and rows for presentation. Be aware that a useful table is one that does not require further explanation. In other words, a table should be able to stand alone for anyone to interpret without having to refer to more descriptive text elsewhere. Thus every table should have an identifying number, descriptive title, column headings, and row headings. The units for a particular variable should also be indicated (usually in the column or row headings). Any other information pertinent to the table may be placed in a legend below the table. An example of a table (Table I) is provided below.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>The Effects of High Fat Versus Low Fat Diets on Rat Weights and Lengths</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AVERAGE</td>
</tr>
<tr>
<td>DIET TYPE</td>
<td>WEIGHT (g)</td>
</tr>
<tr>
<td>high fat</td>
<td>545.5</td>
</tr>
<tr>
<td>low fat</td>
<td>346.7</td>
</tr>
</tbody>
</table>

The high fat diet consisted of 50% fat, while the low fat diet consisted of 5% fat. Each rat was fed ad libidum. Length was measured from the tip of the snout to the base of the tail. The values presented averages for samples of 100 rats.

Figures may be graphs, photographs, or drawings. As with tables, figures should be able to stand alone without having to refer to text elsewhere. Figures require assigned figure numbers, descriptive titles, and legends if necessary. The axes of graphs should be adequately labeled with units indicated. Examples of graphs (Figs. 1 & 2) are presented below.

The following rules regarding the presentation of data in tables or figures should be followed:

1. Each figure or table must be identified by a figure or table reference number.

2. Each figure or table must have a clear descriptive title. If the figure or table refers to biological specimens, then the scientific names of these specimens are generally referred to in the title.

3. In tables, column/row headings should clearly identify the variable and its units of measure.

4. In figures such as graphs, the axes should be clearly labeled and the units of measure indicated.

5. In figures, such as diagrams or maps, the pertinent features should be clearly labeled.
6. Do not cram figures and graphs into a small space on a sheet of paper; try to fill up a whole sheet of paper.

7. A general rule of thumb is that each figure/table should be able to stand alone without forcing the reader to read additional text to understand it.

Procedures and Assignments

I. AVERAGES AND STANDARD DEVIATIONS

Using your koa haole seed pod data from the previous lab activity (pod length in cm, pod mass in gm, and number of seeds per pod), calculate the averages and standard deviations for each of your variables (thus you will have three means and three standard deviations to calculate). These values should be presented in a table along with the pooled class data (see Part II below).

Don't forget to provide a title and proper row and column headings for this table.

II. TABLE OF AVERAGES AND STANDARD DEVIATIONS

You will be given a copy of the averages and standard deviations for the data pooled over the entire class. Include these class values, along with your own values into a table (Table I) that allows the reader to compare your individual values with the pooled values.

Don't forget to provide a title and proper row and column headings for this table.

III. FREQUENCY HISTOGRAM FOR NUMBER OF SEEDS PER POD USING CLASS DATA

Using the pooled class data for the number of seeds per pod, plot a frequency histogram (Figure 1) for this variable only.

Be sure the axes are properly labeled and the graph has a descriptive title.

IV. GRAPHING RELATIONSHIPS: WEIGHT VERSUS LENGTH AND NUMBER OF SEEDS PER POD VERSUS LENGTH

Using Excel prepare two graphs (using your data only) as described below. Be sure the axes are properly labeled and the graphs have descriptive titles.

A. Weight Versus Length

Draw a graph (Figure 2) that presents seed pod weight as a function of seed pod length.

B. Seeds/Pod Versus Length

Draw a graph (Figure 3) that presents the number of seeds per pod as a function of seed pod length.

C. Linear Trends in the Data

Using a ruler, draw the "best fit" straight line that illustrates the trend of the data in each graph described in Parts A & B above.

V. DATA INTERPRETATION AND ANALYSIS

A. Comparing Individual Data to Class Data

Were your values much different from the combined class data? Explain what effect might be responsible for any differences observed.

B. Description of Number of Seeds Pre Pod

After examining the frequency histogram for the number of seeds per pod data, comment of the central tendency and variation of the data. Does your impression from this graph match the actual values calculated?

C. Quantitative Relationships Between Variables

After reviewing the graphs, describe how pod weight and length are related to each other and how the number of seeds per pod relates to pod length. Were these results expected? Explain.

VOCABULARY

statistics
datum/data
variable
statistical population
sample
descriptive statistics
statistics of location
statistics of dispersion
average (arithmetic mean)
range
Summary of Assignment to be Turned In

Your laboratory summary (submitted at the next lab meeting) should consist of the following elements:

1. Descriptive title for assignment.
2. Brief introduction describing the purpose of this lab activity and its objectives.
3. Table (Table I; created using Excel) of averages and standard deviations for your individual data and the class pooled data.
4. Frequency histogram (Figure 1) for number of seeds per pod (class pooled data only).
5. Weight-versus-length graph (Figure 2) for your koa haole seed pod data.
6. Seeds/Pod-versus-length graph (Figure 3) for your koa haole seed pod data.
7. Interpretation and analyses of the data and relationships observed (see Section V.)

The assignment must be typed. Written text must utilize correct spelling, complete sentences, and correct grammar.
CALCULATING MEANS AND STANDARD DEVIATIONS

MEANS

The average value or mean is calculated by summing up the values of the observations (the Xi's) and dividing this total by the number of observations (n):

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

An example of how to make these calculations is illustrated below:

<table>
<thead>
<tr>
<th>i</th>
<th>X_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

\[
\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 2 + 3 + 5 + 8 + 10 = 28
\]

n = 5

\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{28}{5} = 5.6
\]

STANDARD DEVIATIONS

The standard deviation is conventionally calculated by finding the difference between each value and the mean value, squaring each of these differences, summing up these squared differences, dividing this number by the number of observations minus one, and taking the square root of this result:

\[
\text{S.D.} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}
\]

However, this formula leads to substantial rounding error because the subtraction step takes place before squaring the differences between the values and the mean. A more robust formula is the following:

\[
\text{S.D.} = \sqrt{\left[ \frac{\sum_{i=1}^{n} X_i^2}{n} - \left( \frac{\sum_{i=1}^{n} X_i}{n} \right)^2 \right]} / (n-1)
\]
Using the symbols (butterfly, whale & octopus) to represent the quantities defined above, the formula presented on the previous page may be expressed as:

\[
S.D. = \sqrt{\frac{\sum (X_i)^2 - (\sum X_i)^2}{(n - 1)}}
\]

Arrange the data as follows:

<table>
<thead>
<tr>
<th>(X_i)</th>
<th>((X_i)^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>28</td>
<td>202</td>
</tr>
</tbody>
</table>

= 202 \quad = 28 \quad = 5

\[
S.D. = \sqrt{\frac{202 - \frac{28^2}{5}}{(5 - 1)}}
\]
\[
\sqrt{\frac{(202 - \frac{784}{5})}{(4)}} \\
= \sqrt{\frac{(202 - 156.8)}{(4)}} \\
= \sqrt{\frac{45.2}{4}} \\
= \sqrt{11.3} = 3.36
\]