SUMMARIZING, PRESENTING AND ANALYZING DATA

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 data from the previous lab activity. 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 (see Appendix I). 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 the televised presentation (see also downloadable handout on calculating means and standard deviations).

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 also 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 I
The Effects of High Fat Versus Low Fat Diets on Rat Weights and Lengths

DIET TYPE	AVERAGE WEIGHT (g)	AVERAGE LENGTH (cm)
high fat	545.5	55.3
low fat	346.7	53.2

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.

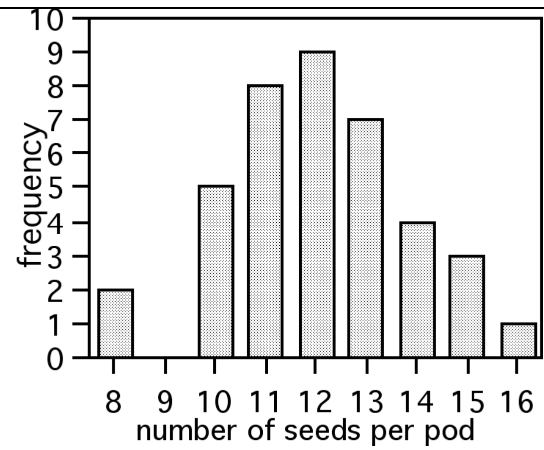

Figure 1. Frequency histogram of number of seeds per pod for koa haole, *Leucaena leucocephala*, seed pods.

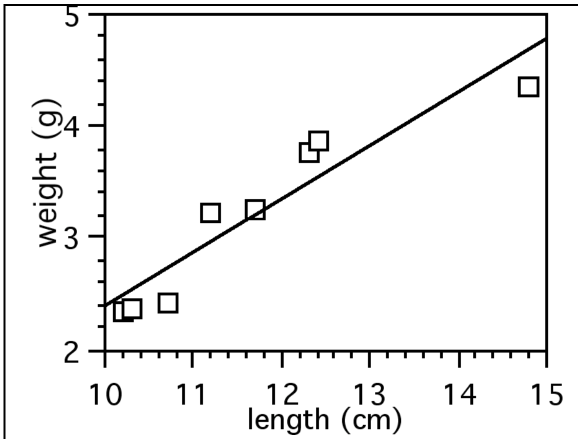

Figure 2. Weight-length relationships of koa haole, *Leucaena leucocephala*, seed pods.

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 clear 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

****You will need to download the data from the course website or course WebCT site when the data are available.****

I. AVERAGES AND STANDARD DEVIATIONS: CLASS DATA

Using the class data, calculate the averages and standard deviations for each of your variables measured (foot length, height, weight, pulse rate before exercise, and pulse rate after exercise) separated according to gender. Present these values in a table. Be sure to follow the rules for presenting data in tables.

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

II. FREQUENCY HISTOGRAM FOR PULSE RATE

Using the pooled class data for the pulse rate, plot two frequency histograms: one for pulse rate before exercise and one for pulse rate after exercise. Group the data into intervals of 5 beats per minute (e.g., 60-64, 65-69,70-74, 75-79, 80-85, etc.) for the plot.

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

III. GRAPHING RELATIONSHIPS: FOOT LENGTH VERSUS HEIGHT AND BODY WEIGHT VERSUS HEIGHT.

Prepare two graphs using the class data as described below (Sections III.A. & III.B.).

In preparing these graphs, it should be obvious to the reader which points refer to males and which points refer to females (you could use solid circles to represent female data and open circles to represent male data).

Be sure the graph axes are properly labeled (with units of measure presented) and the graphs have descriptive titles.

A. Foot Length Versus Height

Draw a graph that presents foot length as a function of height.

B. Body Weight Versus Height

Draw a graph that presents body weight as a function of height.

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 (draw separate lines for male and female data respectively).

IV. DATA INTERPRETATION AND ANALYSIS

A. Comparing Males and Females

After reviewing the descriptive statistics for the foot lengths, heights, body weights, and pulse rates, describe how males and females are likely to be different from each other with respect to these variables. Provide possible explanations for these differences.

B. Comparing Pulse Rate Before and After Exercise

Describe your observations regarding the affect of exercise on pulse rate (in terms of the averages and frequency histogram). Provide possible explanations for these differences.

C. Quantitative Relationships Between Variables

After reviewing the graphs, describe how foot length relates to height and how body weight relates to height. Were these results expected? Is gender an important factor influencing these relationships? Explain.

VOCABULARY

statistics
datum/data
variable
statistical population
sample
descriptive statistics
statistics of location
statistics of dispersion
average (arithmetic mean)
range
standard deviation
frequency histogram
table
figure