AP BIO **Packet on Graphing and Analysis**

Excerpts from *AP Biology Quantitative Skills: A Guide for Teachers*

Descriptive statistics and graphical displays are… useful for summarizing larger data sets and for presenting patterns in data. Tables, while useful for collecting and compiling data to answer a question, are often not the best way to make sense and communicate the results of an investigation. For illustration, consider a nonbiological example: exam grades. Suppose that a class of 100 AP Biology students sat for an exam that was worth 100 points. The teacher posts the anonymous results in a table of 100 unsorted numbers. With their own score and the table of results, a student could compare his or her score to others, but such a comparison would likely not be obvious. However, if instead the teacher posted a mean (average) score and a standard deviation (how variable the scores were) and presented the student scores in a graphical display known as a histogram, students could quickly compare their scores to others in the class. Likewise, the teacher would have a good summary for how the class did as a whole. The same is true for AP Biology investigations. It is all about communication. Descriptive statistics and graphical displays share the same goal of presenting and summarizing data—one with numbers and one with visual displays. Most often the two are combined when presenting the results of an investigation.

The AP Biology laboratory manual is designed to encourage students to ask their own questions by designing and carrying out investigations. This process of inquiry requires data analysis and communication of results. The data collected to answer questions generated by students will generally fall into three categories: (1) normal or parametric data, (2) nonparametric data, and (3) frequency or count data. Normal or parametric data are measurement data that fit a normal curve or distribution. Generally, these data are in decimal form. Examples include plant height, body temperature, and response rate. Nonparametric data do not fit a normal distribution, may include large outliers, or may be count data that can be ordered. A scale such as big, medium, small (qualitative) may be assigned to nonparametric data. Frequency or count data are generated by counting how many of an item fit into a category. For example, the results of a genetic cross fit this type of data as do data that are collected as percentages.

The typical questions asked in an AP Biology lab investigation can likewise be divided into two groups: those questions that compare phenomena, events, or populations (Is A different from B?), and those questions that look for associations between variables (How are A and B correlated?).

Understanding the five graph types used in this guide will be sufficient for getting started with the investigations in the new lab manual and the AP Biology course. **Bar graphs** are graphs used to visually compare two samples of categorical or count data. Bar graphs are also used to visually compare the calculated means with error bars of normal data…. **Scatterplots** are graphs used to explore associations between two variables visually. **Box-and-whisker** plots allow graphical comparison of two samples of nonparametric data (data that do not fit a normal distribution). **Histograms**, or frequency diagrams, are used to display the distribution of data, providing a representation of the central tendencies and the spread of the data.

Bar Graphs

Many questions and investigations in biology call for a comparison of populations. For example, Are the spines on fish in one lake without predators shorter than the spines on fish in another lake with predators? or Are the leaves of ivy grown in the sun different from the leaves of ivy grown in the shade? If the variables are measured variables, then the best graph to represent the data is probably a bar graph of the means of the two samples with standard error indicated (Figure 1). In Figure 1, the sample standard error bar (also known as the sample error of the sample mean) is a notation at the top of each shaded bar that shows the sample standard error (SE, in this case, ±1). Most of the time, bar graphs should include standard error rather than standard deviation (discussed in Chapter 2). The standard error bars provide more information about how different the two means may be from each other. Sample standard error bars are not particularly easy to plot on a graph, however. In Excel, for example, the user needs to choose the "custom error bar" option. An Internet search will yield links to video instructions on "how to plot error bars in Excel."

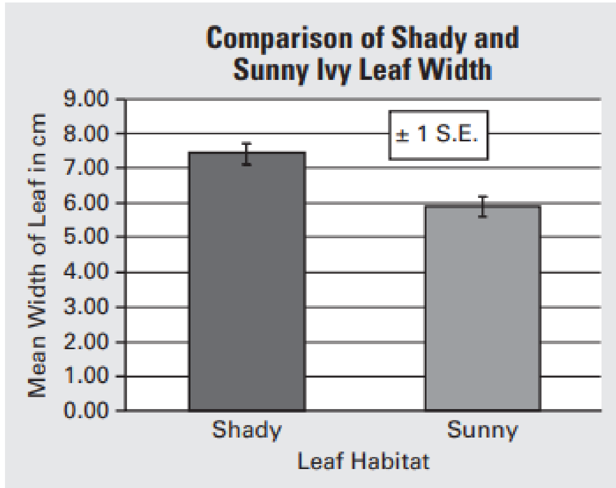

Figure 1. A Bar Graph

Figure 2 shows a variation of a bar graph that only plotted a point for each mean, but there are still standard error bars around each mean. Note that the light intensity (the independent variable, or the variable manipulated) is on the x-axis, and the rate of photosynthesis (the dependent variable) is on the y-axis. A standard bar graph could have been used, but the bars would have cluttered up the display. For this reason, only the means were plotted. The points imply a function between the two variables. Had a line been drawn between the points, this would have been a line graph. However, since data were not taken at all light intensities, a line is not appropriate. With more advanced mathematical techniques, one could draw a line of best fit to describe the relationship.

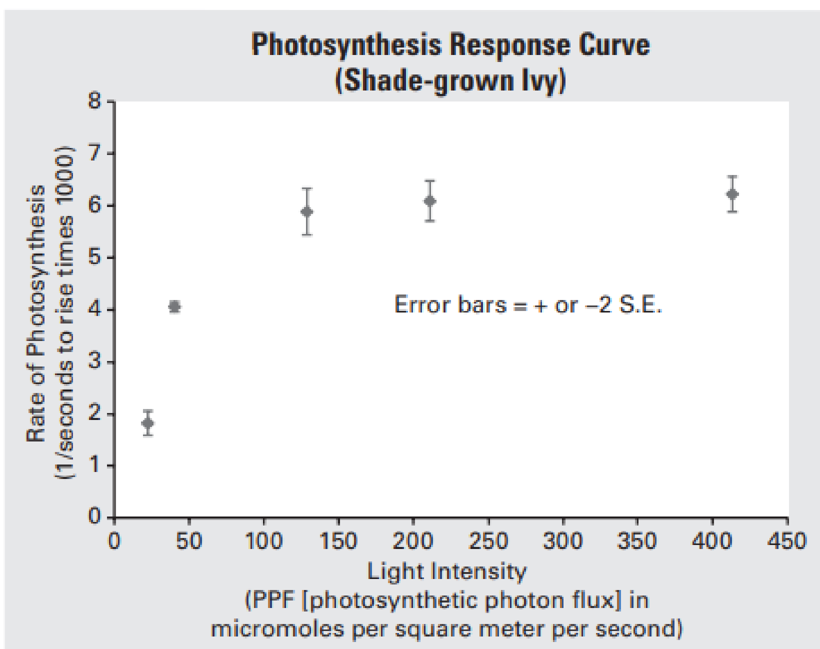

Figure 2. Variation of a Bar Graph with Only the Means Plotted

Scatterplots

When comparing one measured variable against another—looking for trends or associations—it is appropriate to plot the individual data points on an x-y plot, creating a scatterplot. If the relationship is thought to be linear, a linear regression line can be calculated and plotted to help filter out the pattern that is not always apparent in a sea of dots (Figure 3). In this example, the value of r (square root of R2) can be used to help determine if there is a statistical correlation between the x and y variables to infer the possibility of causal mechanisms. Such correlations point to further questions where variables are manipulated to test hypotheses about how the variables are correlated. Students can also use scatterplots to plot a manipulated independent x-variable against the dependent y-variable. Students should become familiar with the shapes they'll find in such scatterplots and the biological implications of these shapes. For example, a bell-shaped curve is associated with random samples and normal distributions. A concave upward curve is associated with exponentially increasing functions (for example, in the early stages of bacterial growth). A sine wave–like curve is associated with a biological rhythm.

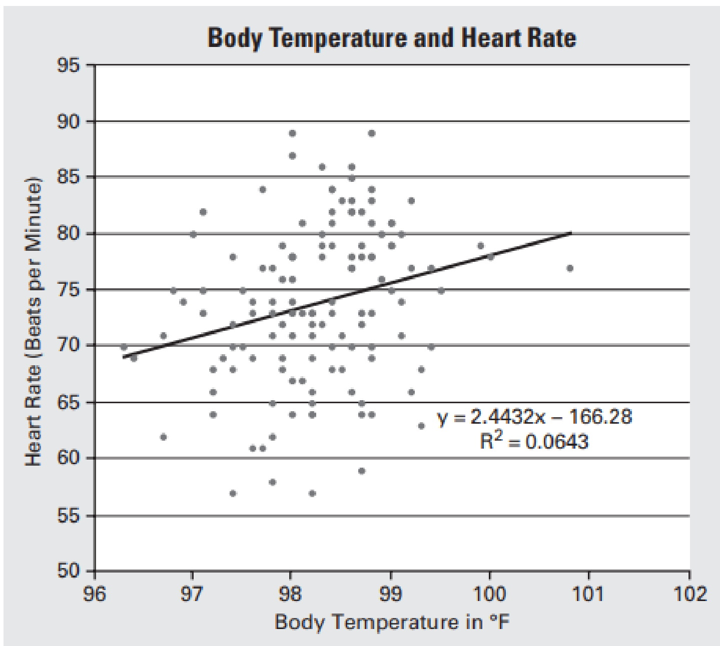

Figure 3. A Scatterplot with a Linear Regression Line

Box-and-Whisker Plots

The appropriate descriptive statistics are medians and quartiles, and the appropriate graph is a box-and-whisker plot (whisker plot). In the graph, the ticks at the tops and bottoms of the vertical lines show the highest and lowest values in the dataset, respectively. The top of each box shows the upper quartile, the bottom of each box shows the lower quartile, and the horizontal line represents the median. The graph allows the investigator to determine at a glance, in this case, that the ash leaves appear to decay the fastest and the beech leaves take longer to decay. Excel and most other spreadsheet programs do not plot box-and-whisker plots automatically. If the data to be graphed are best plotted with box-and-whisker plots, use Google to find a video or instructions for making a box-and-whisker plot in Excel.

Bag	% Decay		
Number	Ash	Sycamore	Beech
1	51	40	34
2	63	33	15
3	44		26
4		52	21
5	48	48	
6	32	35	11
7	70	44	19
8	48	63	32
9	57	40	

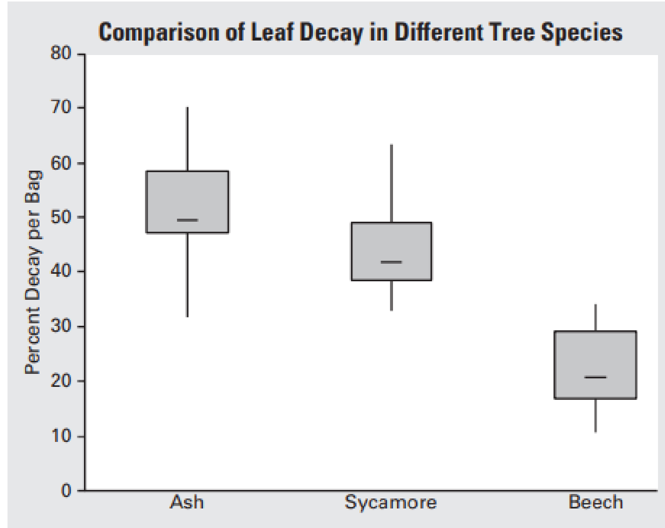

Figure 4. Nonparametric Data and Their Representation in a Box-and-Whisker Plot [*Source: Redrawn from "Merlin_examplels.xls," available as part of a download at: http://www.heckmondwikegrammar.net/index.php?highlight=introduction&p=10310*]

Histograms

When an investigation involves measurement data, one of the first steps is to construct a histogram to represent the data's distribution to see if it approximates a normal distribution (which will be defined shortly). Creating this kind of graph requires setting up bins—uniform range intervals that cover the entire range of the data. Then the number of measurements that fit in each bin (range of units, shown in Figure 5) are counted and graphed on a frequency diagram, or histogram. If enough measurements are made, the data can show an approximate normal distribution, or bell-shaped distribution, on a histogram.

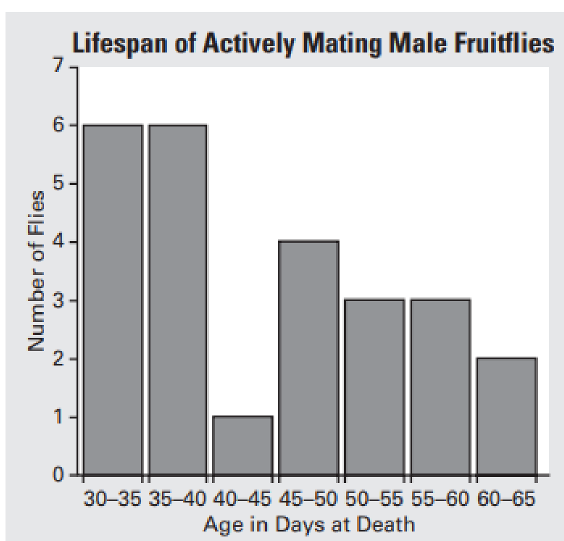

Figure 5. Histogram Showing Nonparametric Data [*Source: Hanley, James A., and Stanley H. Shapiro. "Sexual Activity and the Lifespan of Male Fruitflies: A Dataset That Gets Attention."* Journal of Statistics Education *2, no. 1 (1994).*]

Line Graphs

When the data in a graph represent more or less continuous process, a line graph may be permitted. In the following example, we would be studying a chemical reaction in which substance A is being used up. You can relate the two kinds of information (time & quantity of A) to one another on a line graph.

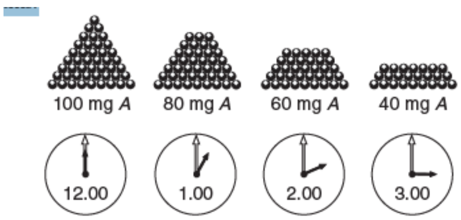

Figure A.1

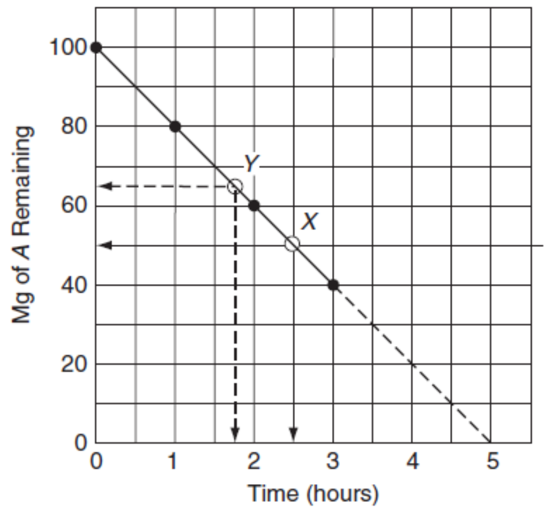

Figure A.3

It should be clear by looking at our graph that only the measurements we actually made are those indicated by the dots. However, because the information on both scales of the graph is assumed to be continuous, we can use the graph to find out how much A would have been found if we made our measurements at some other time, say 2.5 hours. We merely locate the line that corresponds to 2.5 hours on the time scale and follow it up until it crosses our graph at point X and read 50 mg of A remaining. Similarly, we can estimate how long it would take for 65 mg of A to remain, see point Y. This is called **interpolation**, which can only be done within the limits of the graph we actually measured. Notice, also, that part of the graph is drawn with a broken line. In making a line graph we are only allowed to connect the points of our actual measurement. The broken line is called an **extrapolation**. It is an estimate based on our data, but goes beyond our actual experience.

Sometime a line graph represents a rate. Figure A.3 represents the rate of the decomposition of substance A. To discover the rate, we find the slope of the graph at that point. Rate can be calculated at one specific point, between two specific points, or overall for the entire process (using the first and last points). Rate/slope is calculated using the *rise over run* calculation.

$$Slope = \frac{\text{rise}}{\text{run}} = \frac{\Delta Y}{\Delta X} = \frac{y_2 - y_1}{x_2 - x_1}$$

Graphing

A graph must contain five major parts:

a. Title
b. The independent variable
c. The dependent variable
d. The scales for each variable (including written label of the scale and the unit)
e. A legend

- The **TITLE**: depicts what the graph is about. By reading the title, the reader should get an idea about the graph. It should be a concise statement placed above the graph. Every graph needs a descriptive title. Poor examples of titles: ***Lab 1A, Pea Growth, Pizza is Awesome!*** Examples of a descriptive title: ***The effect of light intensity on pea plant growth. Relationship between study time and score on the AP Biology Exam***

- The **INDEPENDENT VARIABLE**: is the variable that can be controlled by the experimenter. It usually includes time (dates, minutes, hours, etc.), depth (feet, meters), and temperature (Celsius). This variable is placed on the X axis (horizontal axis).
- The **DEPENDENT VARIABLE**: is the variable that is directly affected by the independent variable. It is the result of what happens because of the independent variable. Example: How many oxygen bubbles are produced by a plant located five meters below the surface of the water? The oxygen bubbles are dependent on the depth of the water. This variable is placed on the Y-axis or vertical axis.
- The **SCALES** for each Variable: In constructing a graph one needs to know where to plot the points representing the data. In order to do this a scale must be employed to include all the data points. This must also take up a conservative amount of space. It is not suggested to have a run on scale making the graph too hard to manage. The scales should start with 0 and climb based on intervals such as: multiples of 2, 5, 10, 20, 25, 50, or 100. The scale of numbers will be dictated by your data values. Every axis must have a label and the unit. No exceptions!
- The **LEGEND**: is a short descriptive narrative concerning the graph's data. It should be short and concise and placed under the graph.

Other factors that may be included:

- The **MEAN** for a group of variables: To determine the mean for a group of variables, divide the sum of the variables by the total number of variables to get an average.
- The **MEDIAN** for a group of variables: To determine median or "middle" for an even number of values, put the values in ascending order and take the average of the two middle values. e.g. 2, 3, 4, 5, 9, 10 Add 4+5 (2 middle values) and divide by 2 to get 4.5
- The **MODE** for a group of variables: The mode for a group of values is the number that occurs most frequently. e.g. 2, 5, 8, 2, 6, 11. The number 2 is the mode because it occurred most often (twice).

The graph below includes all of the necessary parts. Index marks would not be necessary when using graph paper. Using a software program, like Excel, would allow the inclusion of the index marks.

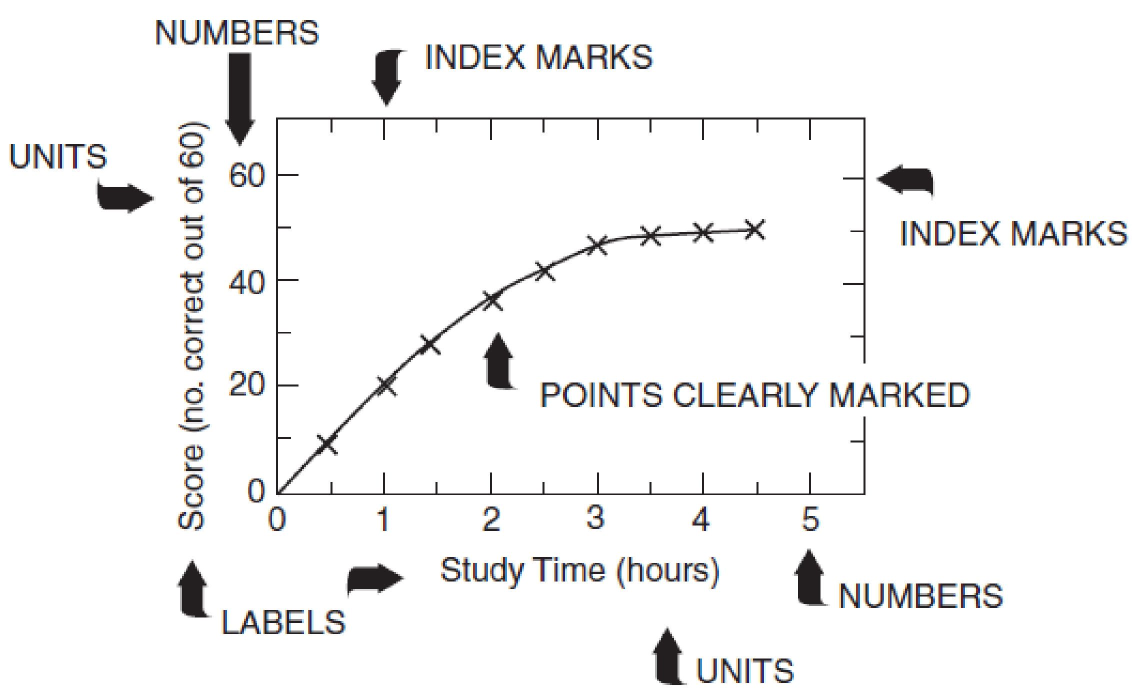

Figure A.7: Relation Between Study Time and Score on a Biology Exam in 2011