Graphing Data

Purpose

Graphing data collected in any experiment serves many purposes. Graphs are used in science not only to communicate information, but also to organize data and assist in looking for patterns and relationships among variables. Recognizing the relationships between phenomena in nature is one of the primary purposes of science. Once a relationship is recognized and evaluated, it becomes possible to make predictions.

Graphs quickly show any data trends or correlation between the independent and dependent variables. Given some type of correlation, graphical representation aids analysis and development of an equation relating the two variables. In addition, graphs make it easier to spot erroneous data points.

This document will explain how to properly graph a data set and develop an equation relating the variables involved. While all graphs may be drawn by hand, use of a computer or graphing calculator speeds the process and makes for easier analysis. Most graphing programs calculate the best fit for a data set by using some version of a least-squares approximation. Correctly graphing by hand requires manually calculating the best fit, a topic not covered here.

Graphing Data

The following procedure shows how to construct a graph for a given set of data. To illustrate the procedure, a complete data set and associated graph appear on the final page. The examples given in each step refer to this illustration.

Setting up the axes

1. Locating variables: When plotting data on coordinate axes, the dependent variable goes on the vertical (y) axis and the independent variable goes on the horizontal (x) axis. Remember, the independent variable *is a function of* the dependent variable as y *is a function of* x, $y = f(x)$.

   Ex: The dependent variable is distance and the independent variable is time. The graph shows distance as a function of time.

2. Labeling axes: Clearly label each axis with the variable name and its associated units in parentheses. Write the title of the graph at the top in the format *dependent variable* vs. *independent variable*.

Ex: The vertical label is "distance (m)" and the horizontal label is "time (s)." The title of the graph is "Distance vs. Time."

3 Setting scales: Choose convenient scales to allow *equal spaces to represent equal changes* on each axis. To determine an acceptable value for the axis interval, find the difference between the smallest and largest values for the variable. Obtain a reasonable number of intervals by dividing the difference by five. Using five as the divisor is somewhat arbitrary, but results in an appropriate number of intervals. Too many intervals crowd a graph and too few make it difficult to plot points. After dividing the number by five, round the resulting dividend to the nearest convenient counting number. Any number easily counted in multiples works well (i.e., multiples of 2, 5, or 10).

Develop a scale for each axis using the rounded dividend as the interval. Begin with a number smaller than the smallest number to be graphed (preferably 0, but if this results in a large amount of unused space, it is not necessary) and end with a number that allows the largest number to be graphed.

Ex: The largest time value is 35 s and the smallest is 0 s, yielding a difference of 35 s. Dividing this difference by five gives 7 s. Since counting by 7 s is somewhat awkward, 5 s intervals were chosen.

Similarly, the largest distance value is 95 m and the smallest is 25 m, yielding a difference of 70 m. Again, dividing by five gives an interval of 14 m. A 10 m interval is used to simplify counting.

Plotting the data

1. Marking data points: Once the axes have been constructed and labeled properly, mark each data point with a large, distinct marker. Simple dots may be confused with stray marks or be covered up by the line of best fit. Use an X or * or some other geometric shape that will prevent these shortcomings. When marking multiple data sets on the same axes, use a different marker for each set.

Ex: Large, dark triangles are used to clearly show the location of data points.

2. Handling repeated trials: Remember, the dependent variable *is a function of* the independent variable, meaning *one and only one* y-value may exist for each x-value. Therefore, average all repeated trials and plot the average value as a single data point. The number of data points appearing on any graph will be the same as the number of distinct values of the independent variable, regardless of the number of trials.

Ex: The data points on the graph correspond to the average values shown in the data table. Note there are 9 data points and 9 distinct values for time.

Analyzing the graph

1. Correlation: The first step in analyzing the graph is to decide if there is any correlation among the data points. A graph showing no correlation would have data points scattered about in a random manner, filling the entire graph much like stars in the night sky. This results when the two variables are unrelated (i.e., the dependent variable *does not* depend on the independent variable) or when the experiment is flawed.

Slight correlation would result in some regions of the graph containing more data points than others. These graphs will show a general trend to the data but will not yield an equation relating the two variables. A graph with most points along the diagonal from the origin (i.e., sloping up) indicates a direct relationship -- as one variable increases, so does the other. A graph with most points along the opposite diagonal (i.e., sloping down) indicates an inverse relationship -- as one variable increases, the other decreases.

Strong correlation is indicated by data points falling between narrow boundaries. The data points may form a straight line or some other type of curve. When this occurs, an equation may be developed that expresses a relationship between the variables.

Ex: The data points lie between narrow boundaries along a straight line. The upward slope indicates a direct relationship. Since the data show strong correlation, an equation may be found relating distance and time.

2. Determining the equation: Graphs with strong correlation may be used to find an equation relating the variables. A linear graph (straight line) is the only graph that yields a rapid and accurate equation. If the data fall along a straight line, the equation may be found immediately as described below. If the data fall along another curve, the curve must be straightened as described in step 3.

When plotting data on a computer or graphing calculator, the program will provide the slope (m) and y-intercept (b) of the line that best fits the data. Keep in mind most programs will attempt to find the best linear fit even if the data fall on a curve. When this happens, modify the graph as described in step 3 until the data fall on a straight line. If you are graphing the data by hand, drawing the line that best approximates the data will provide a crude m and b. A least-squares fit algorithm -- described in most good algebra texts -- will do much better.

Given m and b, the equation may be found by using the slope/intercept form of a line:

$$y = mx + b$$

Instead of using y and x, use the dependent and independent variables, respectively. Also, remember to include the units for m and b -- most programs forget to include them! (The units of b will be the same as the dependent variable. The units of m will be the units of the dependent variable divided by the units for the independent variable.)

Ex: The data lie along a straight line, so no modification is necessary. The computer reports the following values for m and b:

$$m = 2.0 \qquad b = 25$$

The units for b must be the same as distance; namely, meters. The units for m must be meters divided by seconds. Using the slope/intercept equation and substituting d (distance) for y and t (time) for x gives:

$$y = mx + b$$
$$d = 2.0\ (m/s)\ t + 25\ m$$

3. Modifying data: When the data fall on a curve other than a straight line, the graph must be modified before an equation may be found. To do this, modify the data on the x-axis (the independent variable) and regraph the data. Instead of

plotting y vs. x, try y vs. x^2, x^3, _x, ln(x), e^x, sin(x), or any other function f(x) you like. Many computer programs will do this directly. (Graphing by hand requires construction of a new graph each time.) Continue trying new functions until the graph yields data that lie along a straight line.

Sometimes it may be difficult to tell which graph gives the best results as more than one may be close to linear. When this happens, refer to the correlation coefficient (COR). Computer programs report the COR as a number between zero and one. The closer to unity, the better the data approximate a straight line.

Once the data have been modified and regraphed to give a linear relationship, the procedure in step 2 may be used to find the equation. The only difference is that x is replaced by f(x) in the slope/intercept formula. The function f(x) is the modification that gives the best fit.

Ex: Although not necessary in the problem illustrated, sometimes data modification will be necessary. Suppose a graph of y vs. x^2 gives a straight line when plotting a new distance vs. time data set. The computer reports $m = 2.3$ and $b = 0.01$. The equation would be as follows:

$$y = mx^2 + b$$

$$d = 2.3\ (m/s^2)\ t^2 + 0.01\ (m)$$

Final considerations

1. Erroneous data points: Sometimes the vast majority of the data points will lie on a nice line and a few will lie elsewhere. When this happens, a careful examination of the experiment may be in order. These points may exist due to experimental error or mathematical miscalculation. Repeating the experiment for these points (or points with similar values of the independent variable) will confirm this. If, however, repetition yields the same result, then the points may indicate a subtlety not considered.
2. Small values of *b*: The y-intercept (*b*) tells where the graph crosses the vertical axis. If the value of *b* is small compared to typical values of the dependent variable, this term may be dropped from the equation.

3. No correlation: Some experiments do yield no correlation between the variables -- both variables change independently of each other. When this happens, there is a flaw in the design of the experiment. One or more interfering variables has not been controlled properly.

   One important relationship that often *appears* to give no correlation is that of a constant function. When the dependent variable changes sporadically but by small amounts (i.e., change is small compared to value), the equation for the data can be written:

   $$y = constant$$

   This indicates that the dependent variable *does not* depend on the independent variable -- a horizontal line. This is considered a null result but still a successful experiment.

4. Recognizing slopes and intercepts: The slope (*m*) and intercept (*b*) correspond to physical quantities. Sometimes these may be interpreted

easily; other times it is more difficult. Careful consideration of the units often provides valuable clues. Uncovering these physical quantities provides valuable insight and leads to generalized equations.

Ex: The equation found relating distance and time was:

$$d = 2.0\ (m/s)\ t + 25\ m$$

The slope, measured in meters per second, corresponds to a speed. The intercept, measured in meters, corresponds to a distance. With this in mind, a generalized equation may be written:

$$d = vt + d_o$$

Sample Data Table and Associated Graph

time (s)	distance 1 (m)	distance 2 (m)	distance 3 (m)	avg. dist. (m)
0	25	25	25	25
4.3	34	34	33	33.7
8.2	42	43	42	42.3
12.8	51	51	49	50.3
17.1	58	58	59	58.3
21.6	68	66	67	67
26.4	77	79	78	78
30.7	87	87	88	87.3
35	96	95	94	95