Data Analysis

INTRODUCTION TO DATA ANALYSIS

Data analysis requires interpretation of results and comparison to relevant theories (or theoretical values). Experimental data is transformed into useful or more desired values. In order to understand the process, it can be helpful to map out how experimental data is related to desired outcomes/results.

Usually there are some mathematical manipulations required in order to obtain the desired results (it consists mostly of algebra, logarithms, and exponents). There is a small **math review** at the back of your textbook. Please be sure that you understand the math relevant to the course; if not, please work with a TA or the instructor.

Using a spreadsheet and/or graphing program like Microsoft® Excel can make data analysis and graphing easier and more accurate. This is especially helpful when performing the same calculation many times or when determining best-fit lines to data. TAs and/or the instructor can help you.

You should be able to work with fundamental units and the measurement of mass, length, volume, temperature and time, as well as uncertainty, systematic and random error, and significant figures.

UNCERTAINTIES IN DATA AND RESULTS

Accuracy and Precision

Science is interested in repeatable phenomena and in properties that can be measured both qualitatively and quantitatively. All measurements are uncertain and subject to error, so the result of a single measurement cannot be trusted.

Errors can be mistakes (when proper procedure was violated or goofed), and data must be thrown out. Random errors, on the other hand, arise from variations in how you read instruments and handle samples, as well as how the equipment is functioning, from one time to the next. Random errors lead to different results each time you measure the same quantity (sometimes too high; sometimes too low). The closer your measurements agree, the more *precise* your data. Precision is a measure of how closely a group of measurements agree among themselves.

Even if the results of repeated measurements are precise, they are not necessarily *accurate*. Accuracy is a measure of how closely the reported value agrees with the "true" value of the property. Your best guess at the true value is the average value you measured, which may or may not be very accurate. If the 50-mL volumetric flask used to make solutions actually held only 45 mL, the solutions would all be more concentrated than thought. The results would be inaccurate, even if the precision was deceptively excellent. This is an example of a systematic error, an error built into the equipment or procedure. Systematic errors affect the result the same way each time the measurement is made.

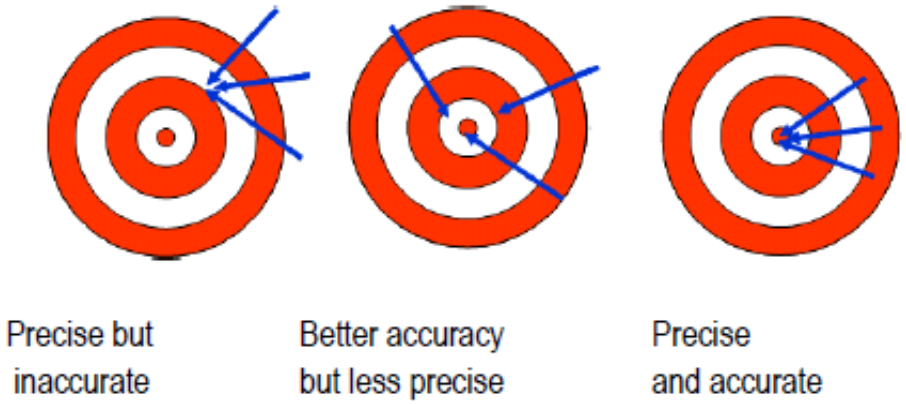

Figure 1

Summary

Precision does not guarantee accuracy, and vice versa. Experiments involve the following.

1) Random errors, which lead to different values when the same property is measured repeatedly. These errors reduce precision.

2) Systematic errors, which can be uncovered by comparing the measured value with a known result or by calibrating the equipment. These errors reduce accuracy.

SIGNIFICANT FIGURES

Counted and defined quantities are exact and therefore have an infinite number of significant figures. The numerical readings that instruments give are always limited in precision. As you use each instrument, get a sense of the precision it can provide and thus the precision associated with a result. Do not record results that exceed the limit of precision of the instrument. The precision associated with a measurement is often called the uncertainty and is reported as a ± value after the data point or result.

For example, the electronic pan balance reads to 0.01 g. If you feel that the mass reading is reproducible to the limit of the balance's accuracy, the uncertainty in a single mass reading would be ±0.01 g. Similarly, with the electronic analytical balance, the uncertainty must be at least ±0.0001 because that instrument reads to only the nearest tenth of a milligram.

Record and report only significant figures (digits that mean something). One way to figure out how many significant figures an instrument provides is to make the same reading several times. If you weigh the same object several times and find that your measurements are all within ±0.01 g of each other, you develop confidence in ±0.01 as the precision you get from the balance. If your measurements are within ±0.05 instead, you will know that for you, today, on that balance, you are not getting the expected precision.

Not all instruments and equipment read to ±1 of the most precise digit. For example, the level of the meniscus in a 100-mL graduated cylinder can probably only be read to ±0.5 mL. You need to record this value in your notebook, and all values measured using this cylinder can only be reported to ##.0 or ##.5 mL.

SIGNIFICANT FIGURES IN CALCULATIONS

You must determine how many significant figures to include when *performing a calculation*. Mathematically, there are rules to use. These are used in the absence of any knowledge of experimental precision (*see error propagation*).

If you add or subtract two measurements, the significant figures in the answer are limited by the measurement with the fewest figures past the decimal point. The number of significant figures in the measurements is not important; the last significant figure in the least precise measurement determines the last significant figure in the answer.

If two people weigh 158 and 98 lb, the sum and difference in their weights is 256 lb and 60. lb, respectively (to the ones place). If two objects weigh 12.07 and 7.4 g, the sum and difference of their weights is 19.5 g and 4.7 g, respectively (to the tenths place).

If you multiply or divide several measurements, the answer contains as many significant figures as there are in the measurement with the fewest significant figures.

If you multiply 1.987 cm × 43 cm, the resulting 85.441 cm^2 must be rounded to 85 cm^2. The number 43 implies that it might really be 42 or 44. Suppose it was actually 42. The result of 42 cm × 1.987 cm is 83.454 cm^2, which is still close to 85. The last three digits have no significance and it would be incorrect to include them. Similarly with division, 1.987 / 43 gives a calculator display of 0.0462093, which must be rounded to 0.046.

MEAN AND STANDARD DEVIATION

To overcome the effects of random error, an experiment is usually repeated. If a measurement has been repeated n times and to give the values $x_1, x_2, x_3, \ldots x_n$, all of the individual measurements aren't reported; instead, the average or mean value, $<x>$, is calculated. Add all the values and divide the sum by the number of values.

$$< x > \equiv \frac{(x_1+x_2+\ldots+x_n)}{n} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

The sign $\equiv$ means that this equality holds *by definition*. In its first use in Eq. 1, the average of x is *defined* to be the sum of the x values divided by the number of values. The second equality introduces the capital sigma notation for the sum.

The average alone says nothing about *precision*. An agreed-upon way to express how closely the data points cluster around the mean is to report standard deviation (fluctuation in data). For example, three people weighing 100, 150, and 200 lb have the average weight of 150, just as do three people weighing 149, 150, and 151 lb. But the groups of people are very different.

The customary way of communicating the precision is to give the standard deviation, σ, defined as follows.

$$\sigma \equiv \sqrt{\frac{(x_1-<x>)^2+(x_2-<x>)^2+...+(x_n-<x>)^2}{n-1}}$$

$$\sigma \equiv \sqrt{\frac{\sum_{i=1}^{n}(x_i- < x >)^2}{n-1}} \quad (2)$$

Note: $(x_i - <x>)$ is the difference between the i^{th} measurement and the average value. Squaring these terms makes the results positive and weights the larger deviations more than the smaller ones. To get a sense of the precision, the percent uncertainty (or percent error) is often reported: $(\sigma \ / \ < x >) \times 100\%$.

The standard deviation is reported as the "± value" after the average and has only one significant figure. The error determines the decimal place of uncertainty, and the reported value is also truncated to the same decimal place as the uncertainty.

About 68% of individual measurements should fall within 1σ of the mean (∼34% above and ∼34% below); about 95% should fall within 2σ. The mean and standard deviation clearly communicate a best guess at the actual value and how well the individual measurements randomly cluster around this value. "Random" means that they cluster around the mean with a frequency resembling the bell-shaped curve.

The three weights 100, 150, and 200 lb give $<x>$ = 150 lb and $\sigma = 50$, which is reported as 150 ± 50 lb. The percent error is about 30%.

$$\sigma = \sqrt{\tfrac{1}{2}[(100-150)^2 + (150-150)^2 + (200-150)^2]} = \sqrt{\tfrac{1}{2}[5000]} = 50 \quad (3)$$

The three weights 149, 150, and 151 lb give $<x>$ = 150 lb and $\sigma = 1$, which is reported as 150. ± 1 lb. The percent error is about 1%.

$$\sigma = \sqrt{\tfrac{1}{2}[(149-150)^2 + (150-150)^2 + (151-150)^2} = \sqrt{\tfrac{1}{2}[2]} = 1 \quad (4)$$

RELATIVE ERROR AND PERCENT ERROR

When an average and standard deviation are used, relative error is $\frac{\sigma}{< x >}$; percent error (% error) is $\frac{\sigma}{<x>} \times 100\%$.

When individual data points are used, replace standard deviation with uncertainty and the average with the measured data.

ERROR PROPAGATION

Many times, results are used to calculate the value of some other quantity. How do the uncertainties in values propagate into uncertainties in a calculated result? Calculus is required to find a general answer that works for any calculation. The answers for calculations involving the addition,

subtraction, multiplication, and division of two numbers A ± a and B ± b are given below.

$$\text{Addition: } (A \pm a) + (B \pm b) = (A + B) \pm (a + b) \tag{5}$$

$$\text{Subtraction: } (A \pm a) - (B \pm b) = (A - B) \pm (a + b) \tag{6}$$

$$\text{Multiplication: } (A \pm a)(B \pm b) = (AB) \pm (AB)(\tfrac{a}{A} + \tfrac{b}{B}) \tag{7}$$

$$\text{Division: } \tfrac{(A \pm a)}{(B \pm b)} = (\tfrac{A}{B}) \pm (\tfrac{A}{B})(\tfrac{a}{A} + \tfrac{b}{B}) \tag{8}$$

For addition and subtraction, the uncertainty in the result is the sum of all the individual uncertainties. For multiplication and division, the uncertainty is the product of the result times the sum of the relative uncertainties of each quantity involved, a/A, b/B, c/C, etc. If you multiply a quantity $(B \pm b)$ by a *constant* known exactly or to much greater precision, the result is $A(B \pm b) = AB \pm Ab$.

For example, the difference in buret readings of 49.06 and 12.74 mL, both with an uncertainty of 0.5 mL, is:

$$\text{Volume} = V_{\text{final}} - V_{\text{initial}} = 49.06 \text{ mL} - 12.74 \text{ mL} = 36.32 \text{ mL}.$$

The uncertainty is simply the sum of the uncertainties of the terms being subtracted:

$$\text{Uncertainty in Volume} = 0.5 \text{ mL} + 0.5 \text{ mL} = 1 \text{ mL}$$

The volume is reported as 36 ± 1 mL. The size of the error, 1 mL, makes the 10ths decimal place in 36 mL meaningless.

As another example, an experimental value of the gas constant, R, can be determined. If the temperature was (296 ± 2) K, the volume was (1000.0 ± 0.3) mL, the pressure (748.6 ± 0.5) torr, and the amount of gas was (0.040 ± 0.001) moles, the answer given by a calculator is the following.

$$R = \tfrac{pV}{nT} = \tfrac{(748.6 \text{ torr})(1000.0 \text{ mL})}{(0.040 \text{ mol})(296 \text{ K})} = 63,226 \text{ torr} \cdot \text{mL/K} \cdot \text{mol} \tag{9}$$

The *fraction* of this answer that is uncertain is, from the last of the general equations above, the sum of the fractional uncertainties of the four terms involved.

$$\text{Fractional uncertainty in } R = \tfrac{0.5}{748.6} + \tfrac{0.3}{1000.0} + \tfrac{0.001}{0.040} + \tfrac{2}{296} = 0.033 \tag{10}$$

Uncertainty in R = (0.033)(63,226 torr · mL/K · mol) = 2100 torr · mL / K · mol.

The experimental R is 63,000 ± 2,000 torr · mL / K · mol or $6.3 \times 10^4 \pm 0.2 \times 10^4$ torr · mL / K · mol.

The density-by-geometry error propagation would be calculated in experiment 1.

$$\rho = \frac{m}{V} = \frac{m}{(\frac{\pi d^2}{4})l} = \frac{4}{\pi}\frac{m}{\pi(d)^2(l)} \tag{11}$$

$$\rho_{\text{error}} = \pm\frac{4}{\pi}\frac{m}{(d)^2(l)}(\frac{\mu}{m} + \frac{\delta}{d} + \frac{\delta}{d} + \frac{\lambda}{l}) \tag{12}$$

Please note that **there is no such thing as post-lab error**. Errors in calculations need to be fixed before you turn in your results and your lab report. Post-lab calculation error is not a possible source of error to be offered in the discussion sections of your reports.

MAKING AND INTERPRETING GRAPHS

Graphs, which show how one variable depends on another, are of great value when determining how different properties affect each other. Graphs often highlight *relationships among various properties of interest* and show information that might not be noticed from a table. Here are a series of helpful guidelines:

Step 1

Each pair of experimental values represents a *data point* for your graph. Decide which values belong on the y-axis and which on the x-axis. ***Usually*** the independent variable is plotted on the x-axis and the dependent on the y-axis. For example, pressures can be measured by setting the temperature of a fixed volume of gas. Temperature is the *independent variable* (x-axis), and pressure is the dependent variable (y-axis).

Step 2

Prepare a data table which pairs up each value of the independent variable with the corresponding value for the dependent variable over the entire data range. For example:

temp, °C	0	10	21	29	40	51	62
pressure, torr	697	723	751	770	800	826	855

Step 3

Decide on the clearest layout for the graph and determine the ranges for both variables. Here, the ranges extend from 0 to 62°C (just under 65°C) and from just over 690 to just under 860 torr (a range of ∼170 torr, *not* of 855 torr). Values do not have to start at 0 but can start just below the smallest value of the variable in the data. Also, decide what increments to use on each axis. The scales in the two directions do not have to match. Data points should spread out over the whole graph, which should, in turn, cover at least half the page.

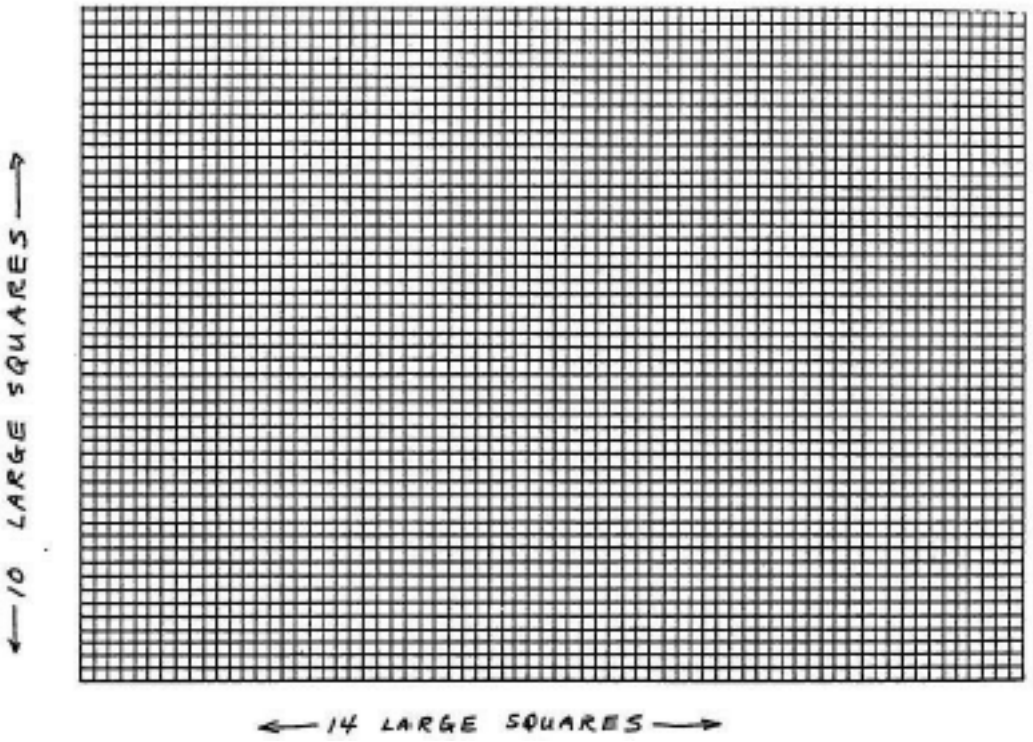

Figure 2

On the graph, there are 14 large squares with five lines in each. The range in x is from 0 to $\sim$70°C, so two large squares for each 10°, or one small square to each °C, is about right. The vertical range is $\sim$170 torr, so 200 torr in the 10 large squares will fit. Assign 20 torr to each large square or 4 torr to each small one. You may want to turn the paper on its side in order to make the data fit better. Spread the range of each variable out over a reasonable distance on its axis so your graph doesn't look like either a horizontal or a vertical line (worthless).

Step 4

Draw the straight lines for the horizontal x-axis and the vertical y-axis. Make tick-marks crossing the axes at equal spacing to show intermediate values of x and of y, ranging from the smallest to the largest values of these variables. If either axis does not go all the way to 0 at the lower left-hand corner of the graph, break the axis just before that corner with a symbol to show that there is a break in the data. [See Figure 3.]

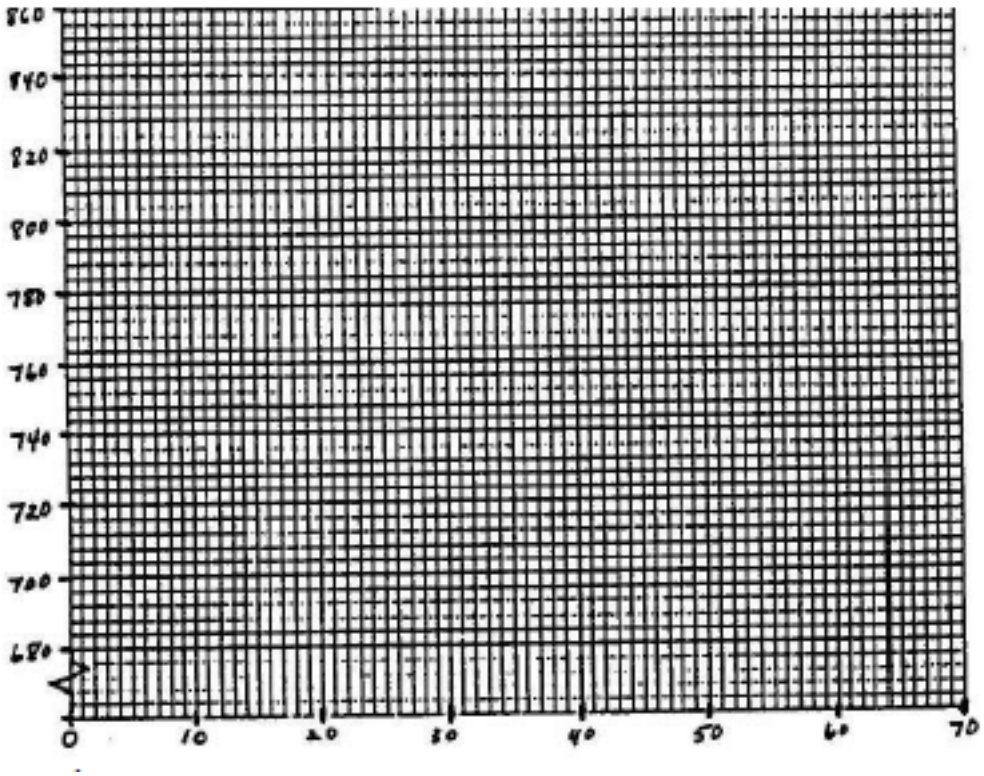

Figure 3

Step 5

Mark a point on the graph for each data pair. [See Figure 4.] If you can estimate the error in each variable, draw a small cross, centered on the point. The height of the cross represents the uncertainty in y; the length represents the uncertainty in x.

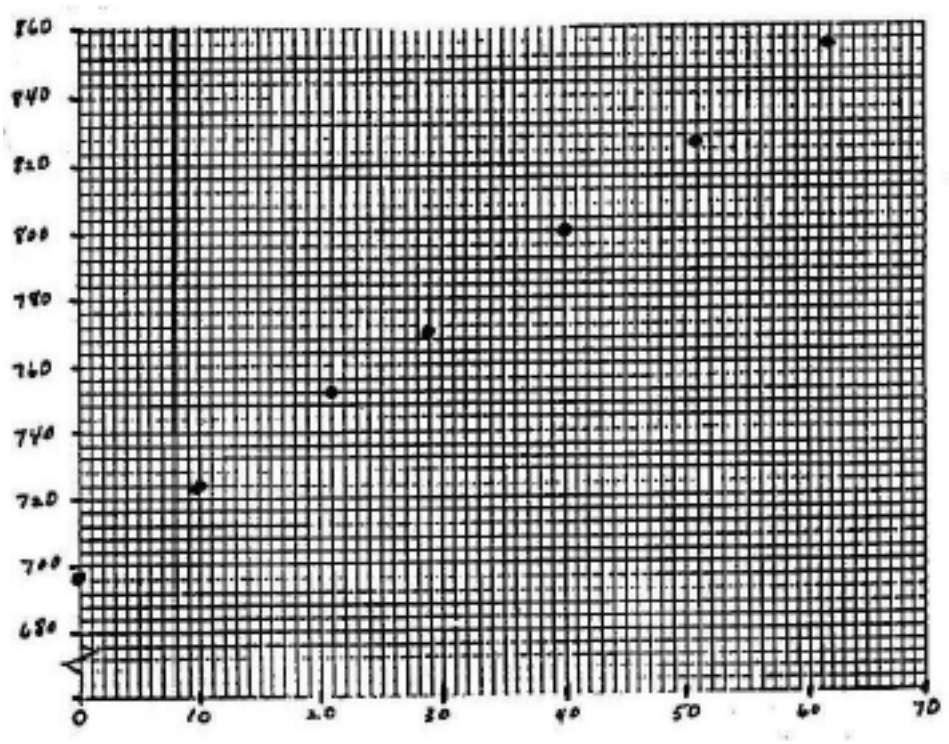

Figure 4

Step 6

Draw the best straight line, the "best-fit line" if the relationship between variables is more or less linear. The majority of your data points should lie close to the line. There are theories of how best to draw this line and computer programs to calculate the line for you (as will some calculators).

The equation for a straight line is: $y = mx + b$, where m is the slope and b is the y-intercept.

To find m and b from a straight line: find two well-separated points on your line where it precisely crosses the intersection of two of the graph paper lines (x- and y-values easily read from the graph). Two such points are shown on the graph. The first such point is labeled x_1, y_1, and the second is labeled x_2, y_2.

The slope is the change in y over the change in x (the difference between y_2 and y_1 divided by the difference between x_2 and x_1). This is often described as the change in y over the change in x or the "rise over run". Once you know m, find b by solving the equation $b = y - mx$ for an easily-read (x, y).

The resulting equations are shown below.

$$m = \frac{y_2 - y_1}{x_2 - x_1} \tag{13}$$

$$b = y_1 - (\frac{y_2 - y_1}{x_2 - x_1})x_1 \tag{14}$$

Step 7

Label each axis neatly (property and units). Also put a title on your graph, such as "Pressure Versus Temperature for Air Sealed in a One-Liter Tank."

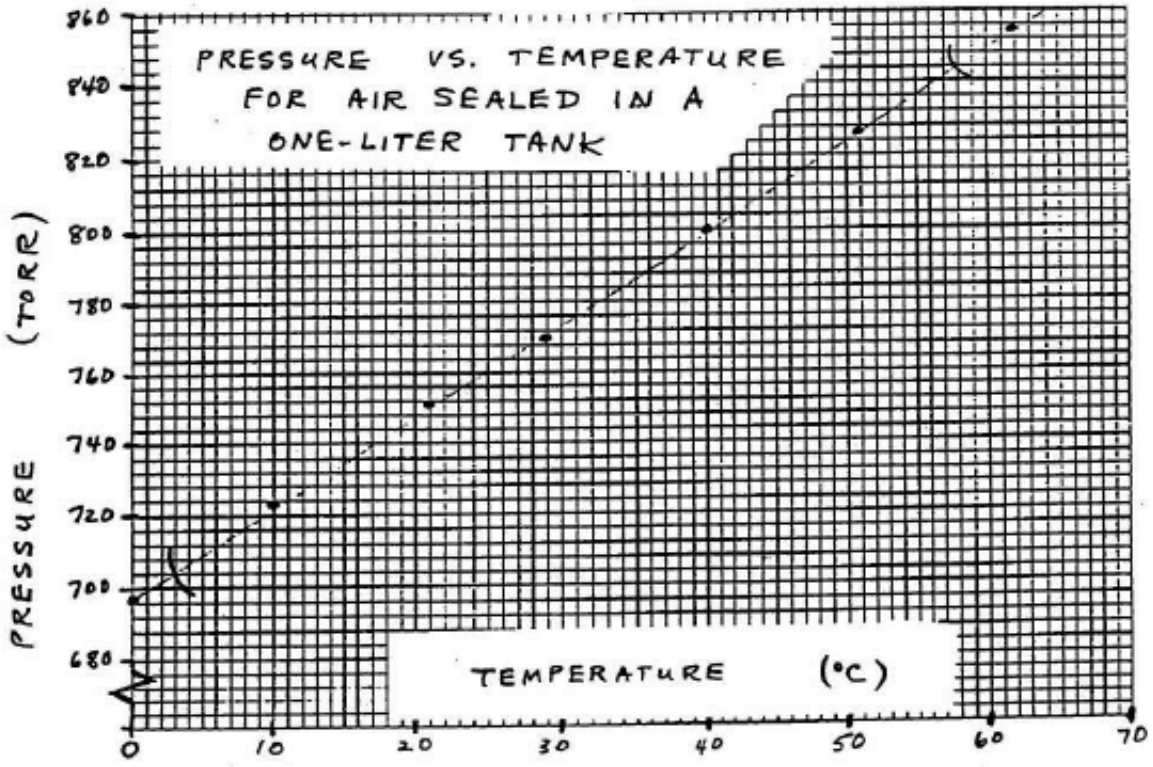

Figure 5

PLOTS OF NONLINEAR RELATIONSHIPS

Sometimes the experimental theory suggests that your plot not be so simple as just p against T. Instead, you may need to plot some function of p against some function of T. For example, if the theory related p to T (given in Kelvins, not Celsius) through the following equation.

$$\log p = \frac{a}{T} + b \tag{15}$$

To generate a plot with a linear relationship, compare this equation to that of a straight line below.

$$y = mx + b \tag{16}$$

The equivalent of y is $\log p$, and the equivalent of x is $1/T$. A plot of $\log p$ against $1/T$ would have a slope of a and a y-intercept of b.

Do this graph the same way you did the one above, except the table made in Step 2 contains t in Celsius, T in Kelvins, and $1/T$. It also contains log p in addition to p:

T, °C	0	10	21	29	40	51	62
T, Kelvin	273	283	294	302	313	324	335
1/T, K^{-1}	.00366	.00353	.00340	.00331	.00319	.00309	.00299
P, torr	697	723	751	770	800	826	855
log p	2.843	2.859	2.876	2.886	2.903	2.917	2.932

Plot log p on the y-axis (values ranging from 2.84 to 2.94) against $1/T$ on the x-axis (values ranging from .0037 to .0029). The data for the y-axis come from the fifth row and the corresponding data for the x-axis come from the third row. The slope would be negative because higher values of log p are associated with lower values of $1/T$.

