Appendix A
Statistics

Data Handling:

Introduction

This document describes the various procedures that may be applied in the processing of data obtained in the modern physics lab. There are four sections, Error Analysis, Significant figures, graphical analysis and curve fitting.

1 Error Analysis

In the laboratory there are generally two types of error which are usually responsible for the experimental errors in the measured quantity – systematic and random errors.

Systematic errors are due to known causes and can, in theory, be removed. These types of errors usually manifest themselves as measured values in which are consistently too high or too low. Systematic errors may be further divided into four types.

- Instrumental – An incorrectly calibrated instrument. E.g. a scale that has not been zeroed.
- Observational – Operator error. E.g. parallax when reading a meter.
- Enviromental – E.g. variation of gravity with altitude.
- Theoretical – Simplification of model or approximation in equations describing the system.

Random errors are positive and negative fluctuations that result in about 50% of the recorded measurements being too high and about 50% being too low. Sources of random errors are not always apparent. Possible causes are:

- Observational – E.g. error in judgment when reading the smallest division of a scale measuring device.
- Environmental – E.g. Unpredictable variations in supply voltage to , or temperature variations in experimental equipment.

Figure 1. illustrates the effects of these two type of errors. Figure 1(a) shows typical results from the measurement of some quantity in the presence of only random errors. In this case, the values are distributed about the true value. Figure 1(b) shows the same measurement but in the presence of systematic errors as well as random errors. In this case, the values are spread about some displaced value rather than the true value.

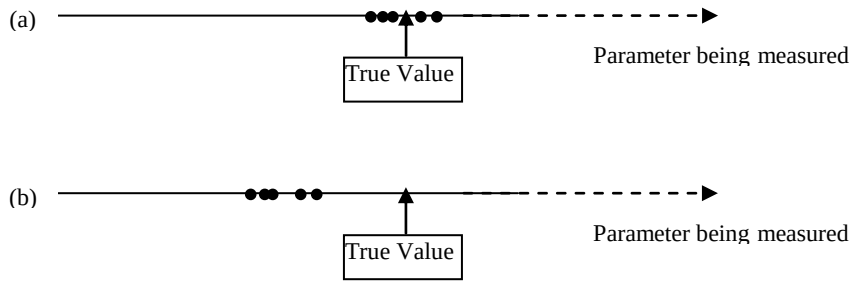


Figure 1: Set of measurements (a) with only random errors present (b) with both random and systematic errors present.

When a physical quantity, such as a length measured with a ruler, is measured several times, then a distribution of readings is obtained because of random errors. For such a set of data the mean or average value is defined by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

where x_i is the ith recorded value and n is the total number of measurements. These n values will be distributed about $\bar{x}$ as shown in Figure 1(a). A smaller spread of values about the mean indicates a higher precision.

Simply defining the mean value of a set of data points is not sufficient. We need to estimate the *precision* or *uncertainty* in the value. There are different ways of doing this, we shall only consider the *standard deviation*.

The standard deviation is defined as

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

If s is small, then the spread in the measured values will be small and the precision in the measurements will be high. The error of uncertainty in $\bar{x}$ is the *standard deviation of the mean*,s_m, defined as

$$s_m \equiv \frac{s}{\sqrt{n}} \quad (3)$$

Where s is the standard deviation and n is the total number of measurements. From this we can say that the average value is

$$\bar{x} \pm s_m \quad (4)$$

1.1 The Gaussian distribution

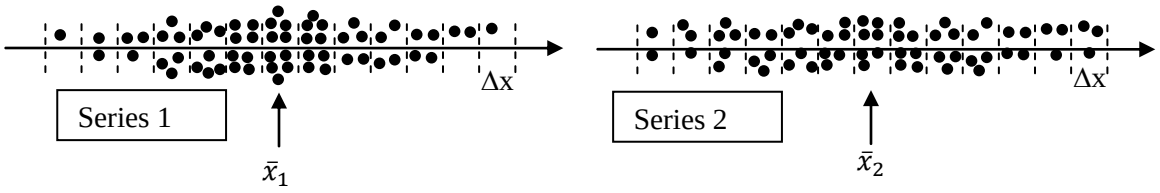
Figure 2 shows the results from two different sets of position measurements. Both sets of data contain n measurements. The x axis has been divided into equal increments of width Δx and each dot represents a measured value. The vertical position of the dots is simply to make the dots more visible and is of no physical significance.

In Figure 2(a) the values in series 1 are more closely clustered and so represent a more precise set of measurements. In Figure 2(b) the number of measured values N(x) for each increment of Δx is shown. The series 1 graph has a more sharper peak which again indicates that the data of series 1 is more precise than series 2. If the number of measurements become very large, then the measured values are distributed evenly about the mean value as shown in Figure 2(c). For very large *n*, the standard deviation is denoted by σ and each curve in figure 2(c) represents the frequency with which some value *x* is obtained as the result of a single measurement. Ideally, the analytical expression for these curves is given by

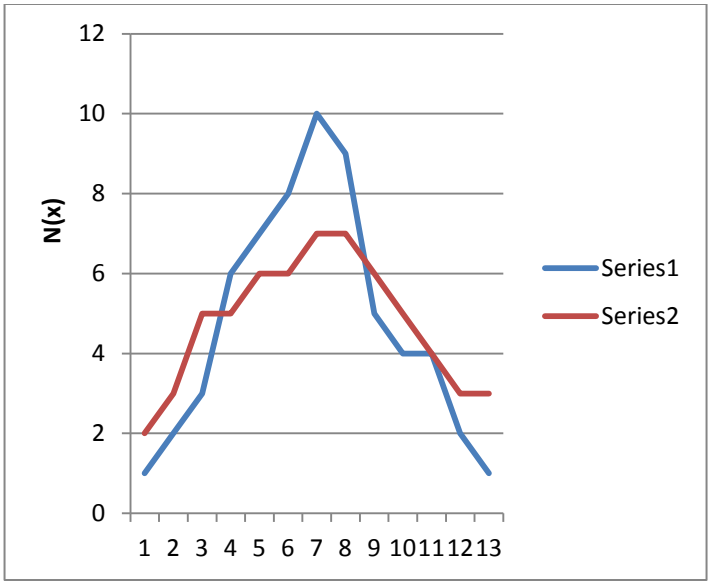
$$N(x) = \frac{n}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right] \quad (5)$$

Where *n* is the very large number of measurements, $\bar{x}$ is the mean value and σ is the standard deviation. This equation defines the *Gaussian (or Normal) distribution*. If the measurements are carried out to great precision σ will be small and the distribution will be a sharp peak about $\bar{x}$.

(a)



(b)



(c)

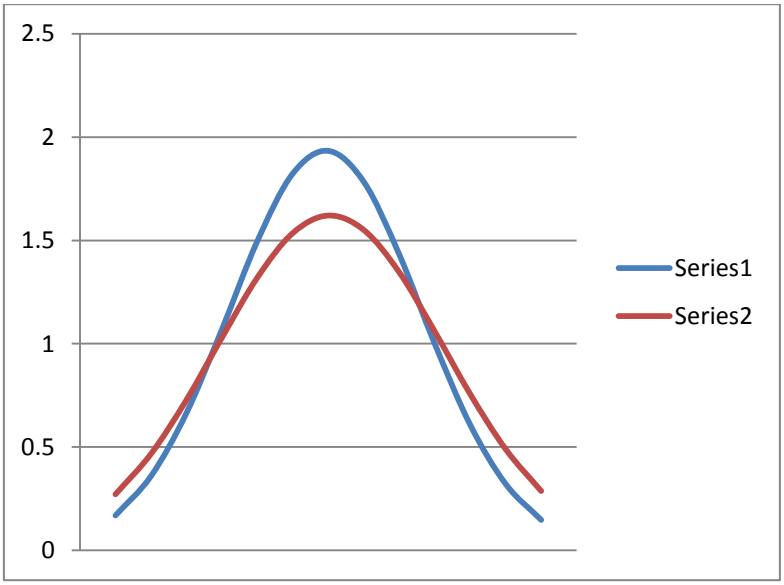


Figure 2: Two Sets of measurements for the same physical quantity. (a) Each dot represents a measurement. (b) N(x) is the number of measurements in the increment Δx. (c) Distribution for very large n. In this limit, the distribution approaches the Gaussian or normal distribution and x_1 and x_2 approach the same value.

Conversely, if the set of measurements is of low accuracy, σ will be large and the distribution will be broad about $\bar{x}$.

To obtain the probability P(x) of obtaining some value of *x* as a result of a single measurement, we divide (5) by n and define P(x) to be N(x)/n to give

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2}\right] \tag{6}$$

1.2 Estimation of random Errors

Although there are precise mathematical methods for calculating random errors, these can be time consuming. A sufficient estimate of measurement errors may be obtained more readily from a subjective standpoint. For example, a length measurement is only likely to be as accurate as the smallest division on the ruler. Any measurement using a ruler whose smallest increment is 0.1cm would have an uncertainty in the measurement of ± 0.1cm.

Further error can be introduced in what is being measured. Figure 4 shows an example where common sense and judgment are needed to provide a suitable estimate. In Figure 4 the distance d_1 represents a clearly define measurement, limited only by the precision of the measuring instrument. However, there is the further complication of deciding where the center of the bodies lay in the measurement for d_2. The error in this measurement would clearly be larger than the former.

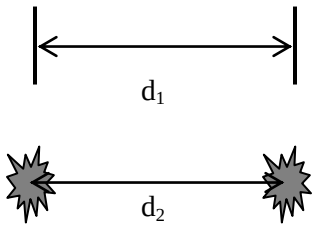


Figure 4

1.3 Propagation of Errors

The propagation of errors is a method by which the error in some final value, which depends on two or more other values with known estimated errors, may be calculated from these known errors. We consider first the method for addition and subtraction of errors, then extend the process to multiplication and division of errors.

Let x, y and z be three measured values with estimated errors δx, δy and δz respectively. We would express these results as

$$x \pm \delta x \quad y \pm \delta y \quad z \pm \delta z \tag{7}$$

Now let w be some known function of the measured values, i.e w = w(x,y,z). We wish to calculate w and its associated error δw. From statistical theory

$$\delta w = \sqrt{\left(\frac{\partial w}{\partial x}\delta x\right)^2 + \left(\frac{\partial w}{\partial y}\delta y\right)^2 + \left(\frac{\partial w}{\partial z}\delta z\right)^2} \tag{8}$$

Equation (8) is the basic formula for error propagation.

1.3.1 Addition and Subtraction of measurements

Suppose that

$$w = ax + by + cz \tag{9}$$

where a, b, and c are known positive or negative constants and x, y and z are measured values with associated errors δx , δy and δz respectively. From (8) $\partial w/\partial x = a$ etc and so

$$\partial w = \sqrt{(a\delta x)^2 + (b\delta y)^2 + (c\delta z)^2} \tag{10}$$

If one of the estimated errors is significantly larger than the others then, to a first approximation, we may ignore the others; for example if $c\delta z$ is much larger, equation (10) approximates to

$$\partial w = \sqrt{(c\delta z)^2} = c\delta z \tag{11}$$

This will often suffice when estimating errors while the experiment is being performed. The final lab write up will require a full treatment.

Example

Suppose that three measured lengths and the associate errors are.

$L_1 \pm \delta L_1 = 5.94 \pm 0.3$ cm, $L_2 \pm \delta L_2 = 1.64 \pm 0.2$ cm, $L_3 \pm \delta L_3 = 4.73 \pm 0.1$ cm

and let the quantity to be calculated, L, be defined as $L = L_1 + 2L_2 + 3L_3$
then from equation (10).

$$\delta L = \sqrt{(a\delta L_1)^2 + (b\delta L_1)^2 + (c\delta L_3)^2} = \sqrt{(1 \times 0.3)^2 + (2 \times 0.2)^2 + (3 \times 0.1)^2} = \pm 0.58 \text{ cm}$$

and so $L = 23.41 \pm 0.6$cm

1.3.2 Multiplication and Division

Suppose that

$$w = kx^a y^b z^c \tag{12}$$

where k, a, b and c are constants, positive or negative. From equation (8)

$$\delta w = [\, (kax^{a-1}y^b z^c)^2(\delta x)^2 + (kx^a by^{b-1}z^c)^2(\delta y)^2 + (kx^a y^b cz^{c-1})^2(\delta z)^2]^{1/2} \tag{13}$$

Dividing both sides by w and writing $w = kx^a y^b z^c$ on the left hand side of the equation yields

$$\delta w/w = 1/(\, kx^a y^b z^c\,)\, [\, (kax^{a-1}y^b z^c)^2(\delta x)^2 + (kx^a by^{b-1}z^c)^2(\delta y)^2 + (kx^a y^b cz^{c-1})^2(\delta z)^2]^{1/2} \tag{14}$$

Moving the $1/(kx^a y^b z^c)$ under the root and then simplifying

$$\frac{\delta w}{w} = \sqrt{\left(\frac{a\delta x}{x}\right)^2 + \left(\frac{b\delta y}{y}\right)^2 + \left(\frac{c\delta z}{z}\right)^2} \tag{15}$$

2 Significant Figures

The significant figures in a number are the figures that are obtained directly from measurements and exclude any zeros included for the purpose of locating the decimal point. A measurement and its experimental error should have their last significant digits in the same position relative to the decimal point – e.g. 3.141 ± 0.003, 314 ± 2 or $(3.14 \pm 0.01)x10^{-1}$.

The correct number of significant figures to which a result should be quoted should ideally be found from error analysis. However, this does take time and usually when performing calculations in the laboratory it is sufficient to retain enough significant figures such that round off errors is not a problem, but not retaining so many as to make the calculation unnecessarily long winded, e.g.,

$$0.91 \text{ x } 1.25 = 1.1 \text{ (wrong)} \tag{16}$$

Here, the numbers 0.99 and 1.11 are defined to around 1% whereas the result is defined only to 10%. So the accuracy of the final result has been reduced by a factor of 10 as a consequence of discarding too many digits.

$$0.91 \text{ x } 1.25 = 1.1375 \text{ (wrong)} \tag{17}$$

Here the extra digits have no meaning, and give an incorrect indication of the accuracy of the experiment.

$$0.91 \text{ x } 1.25 = 1.138 \text{ (okay)} \tag{18}$$
$$0.91x1.25 = 1.14 \text{ (Best)}$$

3 Graphical Analysis

When analyzing experimental data it is often necessary to present data in a graphical form. This section provides guidelines for presentation and analysis.

Speed (m/s)	Time (s)
0.45 ± 0.06	1
0.81 ± 0.06	2
0.91 ± 0.06	3
1.01 ± 0.06	4
1.36 ± 0.06	5
1.56 ± 0.06	6
1.65 ± 0.06	7
1.85 ± 0.06	8
2.17 ± 0.06	9

Table 1: Set of measurements of speed of a particle at different times.

- Select the range of the axes to use as much of the graph as possible. If data occupies only a small portion of the graph, it will hinder interpretation of the graph.
- Give the graph a concise title.
- Label the axes and include units.
- Select a scale for each axis and start each axis at zero (if possible).
- Use error bars to indicate errors in measurement.
- Draw a smooth curve through the data points. If errors are random, approximately 30% of them will not lie within their error range of the best fit curve.

When graphing your data, you will be using a computer. Nevertheless, we will review the process for analyzing data in the absence of a computer as many points remain relevant. In the next section we will look at more precise ways of analyzing data which ideally, but not necessarily, requires a computer.

Consider the set of data provided in Table 1. A graph is shown in Figure 5. From this graph it is clear that the speed varies linearly with time. The general equation for a straight line is $y = mx + c$, where m is the slope of the line and c is the y intercept. The data in table 1 follows the kinematics equation $v = at + v_o$.

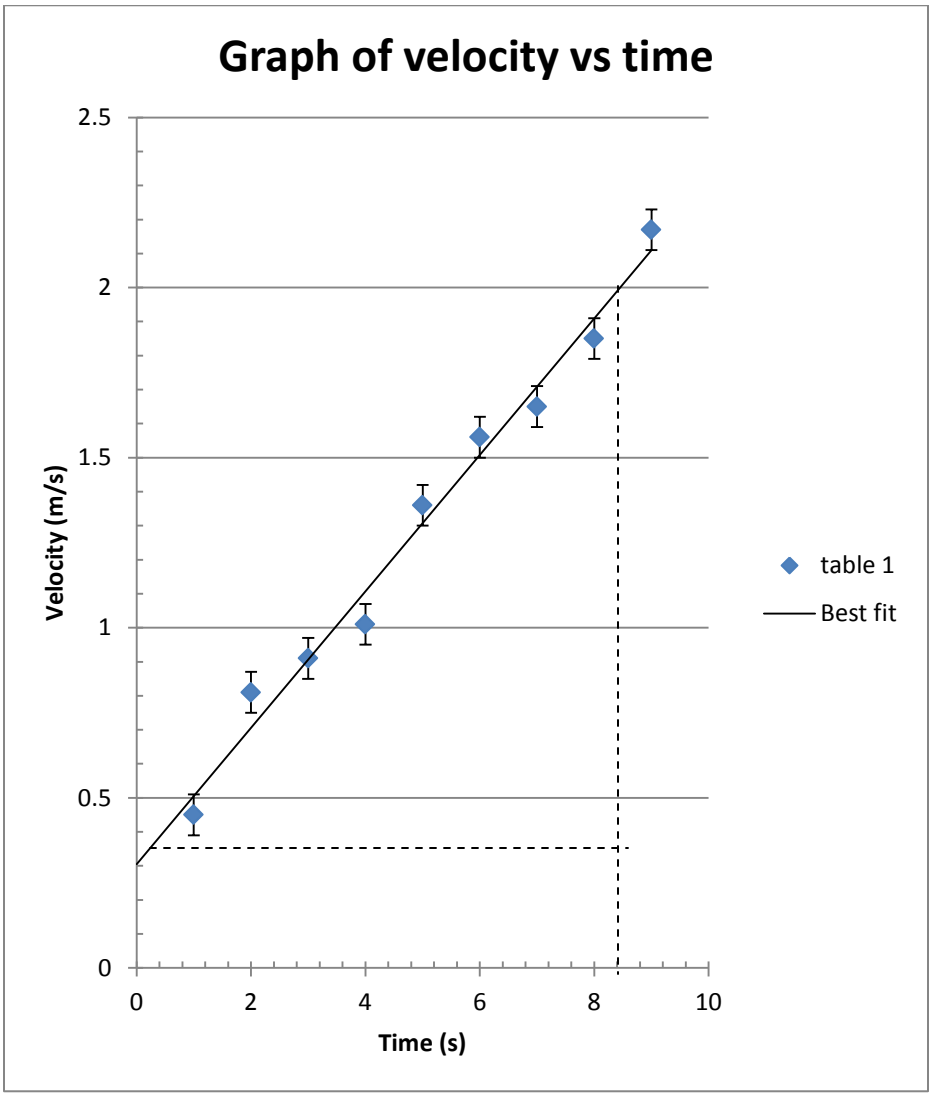


Figure 5

From inspection of the graph, v_o = 0.30 m/s. To determine the slope we select two well separated points on the line (which are not data points) and then

$$\text{a = slope} = \frac{\Delta v}{\Delta t} == \frac{(2-0.3)m/s}{(8.4-0)\ s} = 0.20\ m/s^2 \qquad (19)$$

The equation of the line for the data set is then v = 0.20t + 0.3(m/s)

As an example of a non-linear relationship, we consider the variation of the position of the particle with time. The graph of the data shown in Figure 6 is taken from Table 6. The distance measurement is given an uncertainty of 3%. The error bars for each data point changes accordingly.

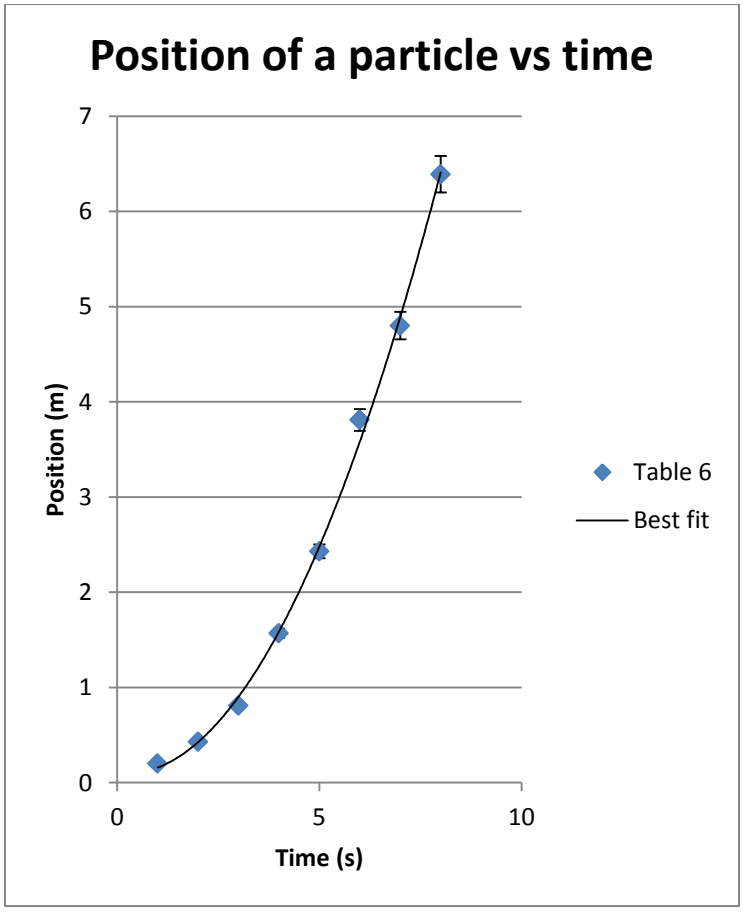


Figure 6

Distance (m)	Time (s)
0.20± 3%	1
0.43± 3%	2
0.81± 3%	3
1.57± 3%	4
2.43± 3%	5
3.81± 3%	6
4.80± 3%	7
6.39± 3%	8

Table 6

Clearly a linear treatment of the data would be inappropriate. From the form of the curve it would not be unreasonable to assume a power or polynomial relationship i the form of $y = at^n$.

In this case, from kinematics we would expect a relationship of the form $d = \frac{1}{2} at^2$, where *a* is the particles acceleration. In this case, a graph of d vs t^2 would yield a straight line. You can do this yourself to see that $d = mt^2 + d_o$.

3.1 Semi-log Data Plotting

Often the relationship between measured variable is not linear. For example, Lambert's Law relates the intensity of Light, I, transmitted through a sample of thickness x

$$I = I_o e^{-\mu x} \tag{20}$$

where I_o is the incident intensity and μ is the absorption coefficient and depends on the sample and the wavelength of the radiation.

Figure 7 shows a set of measurements of I for different values of x. From the smooth curve, it is unclear whether the data obeys Lambert's Law. To find the relationship between I and x, it is necessary to make a semi-log plot. A semi-log plot has a logarithmic y axis and a regular x axis. A semi-log plot of the data is shown in figure 8. Missing from both graphs are the error bars, the uncertainty would not be one fixed value for all measurements but would be determined for each data point individually using the methods of propagation of errors for equation (20) as outline in the section 1.

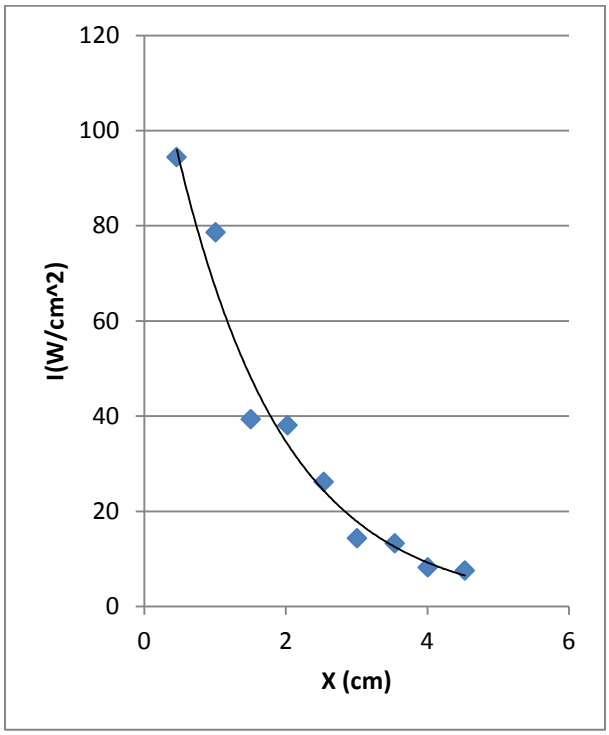


Figure 7

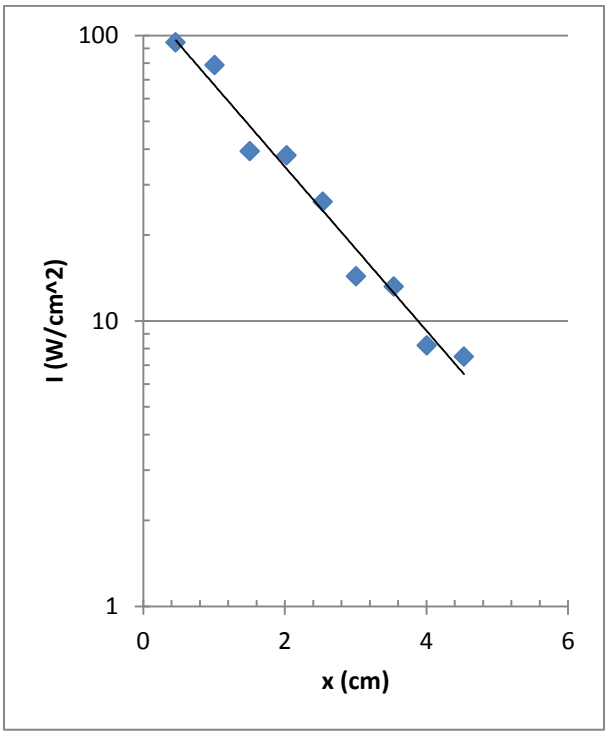


Figure 8

Note on the semi-log

- The y axis has no zero
- When reading values off the y axis you read the logarithm of the value and not the value, e.g. log 9 and not 9

The smooth curve drawn through the data in Figure 8 is a straight line with a negative slope and the intensity at the point where the line intersects the vertical axis ix I_o. Taking the logarithm of (20)

$$\begin{aligned}\log I &= \log(I_o e^{-\mu x}) \\ &= \log e^{-\mu x} + \log I_o \\ &= -\mu x \log e + \log I_o \\ &= -0.434\ \mu x + \log I_o \end{aligned} \qquad (21)$$

Compare equation (21) to the general equation of a line $y = mx + b$, we see that $y = \log I$, $m = -0.434\ \mu$ and $b = \log I_o$. So a plot of log I versus x will yield a straight line with a gradient of -0.434μ and a y-intercept of $\log I_o$. We can calculate the slope as follows:

$$\text{slope} = \frac{\Delta(\log I)}{\Delta x} = \frac{\log 10 - \log 100}{(3.80 - 0.40)cm} = -0.294\ \text{cm}^{-1}$$