This page is deliberately left blank.

Measurement Analysis 2: Probabilistic Uncertainty, Least Squares Fitting, and Graphical Analysis

Goals of this activity

After completing this activity you should be able to

- Interpret and calculate the mean, standard deviation, and standard error of the mean (SEM) of a set of repeated measurements.
- Use a spreadsheet to perform a Least Squares Fit statistical reduction of data to determine the statistically best values (and uncertainties) for the slope and y-intercept of a linear relationship.
- Know the standards for clear, professional, easily-readable graphs (complete with error bars).

1 Probabilistic (or Statistical) Uncertainty — mean, standard deviation and SEM

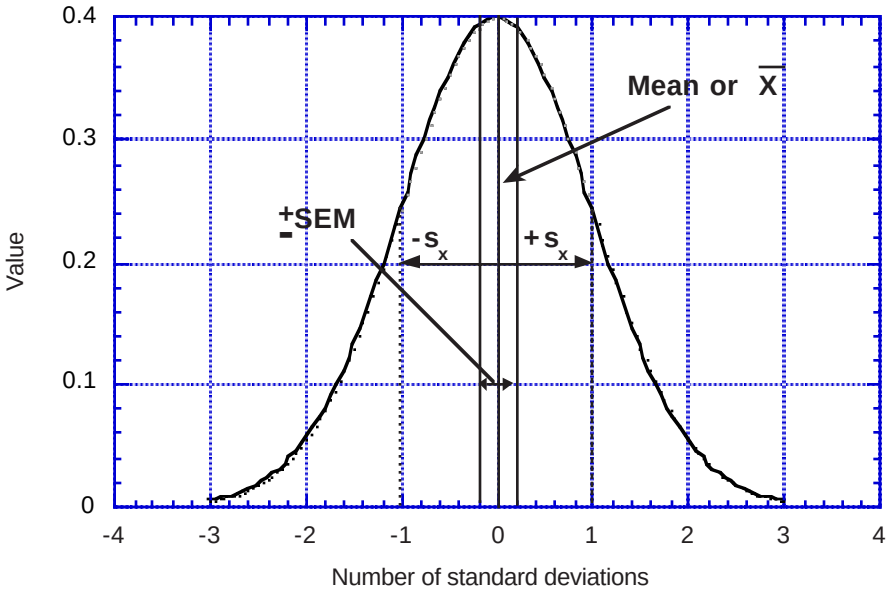

Figure 1: A Gaussian Distribution

When we make a repeated measurement and experience random uncertainties due to resolution limitations, we can treat these uncertainties probabilistically — we assume that the distribution of uncertainty will follow the Gaussian or Normal distribution as shown in Figure 1. This procedure gives a better uncertainty calculation than the propagation methods discussed earlier: those methods always assumed the *worst or maximum uncertainty* in any situation, while statistical treatments give the *most likely uncertainties*.

We begin our discussion with the idea of a statistical mean value. The statistical *mean value* is exactly equivalent to the quantity we knew in high school as the simple average for a set of repeated measurements. For N repetitions of a measurement X_i, the statistical mean is written as

$$\bar{X} = \frac{1}{N}\left[X_1 + X_2 + X_3 + \cdots + X_N\right]$$

or written more compactly,

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{1}$$

For example, suppose that we measure the speed of sound five times in the laboratory and collect the following data: 341 m/s, 344 m/s, 338 m/s, 340 m/s and 343 m/s. The mean value would be:

$$\begin{aligned} \bar{v} &= \frac{1}{N}\sum_{i=1}^{N} v_i = \frac{1}{5}(v_1 + v_2 + v_3 + v_4 + v_5) \\ &= \frac{1}{5}(341 + 344 + 338 + 340 + 343)\ \text{m/s} = \frac{1}{5} \times 1706 = 341.2 \approx 341\text{m/s}. \end{aligned} \tag{2}$$

Often we wish to know by how much measurements deviate from the mean value. The quantity that relays this information is known as the *standard deviation* and is indicated by the lowercase letter s with an appropriate subscript, as in s_X. (Note that there are several different standard deviations in statistics; here we intend the sample standard deviation.) To determine the standard deviation, first the mean value $\bar{X}$ must be calculated, then s_X^2 is calculated by taking the sum of the squares of the deviations of each point from the mean and dividing that sum by $N - 1$:

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{X} - X_i)^2 \tag{3}$$

Then the standard deviation s_X is

$$s_X = \sqrt{s_X^2} \tag{4}$$

Note that Equation 1 for the mean contains a $\frac{1}{N}$ term while Equation 3 for standard deviation squared contains $\frac{1}{N-1}$. The mean and standard deviation for ANY Gaussian curve entirely define the curve and are CONSTANTS.

Recall the speed of sound data we collected in the laboratory. Having already found the mean value of 341 m/s for this data, we now wish to calculate the standard deviation of this sample.

Using Equation 3, we first find s_v^2:

$$\begin{aligned} s_v^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(\bar{v}-v_i)^2 \\ &= \frac{1}{5-1}[(\bar{v}-v_1)^2+(\bar{v}-v_2)^2+(\bar{v}-v_3)^2+(\bar{v}-v_4)^2+(\bar{v}-v_5)^2] \\ &= \frac{1}{4}[(341-341)^2+(341-344)^2+(341-338)^2+(341-340)^2+(341-343)^2]\ \mathrm{m^2/s^2} \\ &= \frac{1}{4}[0+9+9+1+4]\ \mathrm{m^2/s^2} = \frac{1}{4}\times 23\mathrm{m^2/s^2} \quad (5) \\ &\approx 5.8\ \mathrm{m^2/s^2} \quad (6) \end{aligned}$$

Then we use Equation 4 to find the standard deviation:

$$s_v = \sqrt{s_v^2} = \sqrt{5.8\ \mathrm{m^2/s^2}} \approx 2.4\ \mathrm{m/s}$$

After we know the mean value $\bar{X}$ and the sample standard deviation s_X of a set of measurements, we can determine the *Standard Error of the Mean* ($\sigma_{\bar{X}}$ or sometimes SEM) calculated from our limited set of data as:

$$\mathrm{SEM} = \sigma_{\bar{X}} = \frac{s_X}{\sqrt{N}} \quad (7)$$

With $\bar{X}$ and $\sigma_{\bar{X}}$ known, we can write our measurement in the usual form: $(\bar{X} \pm \sigma_{\bar{X}})$. Note that the SEM is very sensitive to reduction by taking more data: the more data, the less uncertainty in the measure.

To find the standard error of the mean $\sigma_{\bar{v}}$ of our speed of sound data, we use Equation 7:

$$\mathrm{SEM}\ = \sigma_{\bar{v}} = \frac{s_v}{\sqrt{N}} = \frac{2.4\ \mathrm{m/s}}{\sqrt{5}} = 1.07\ \mathrm{m/s} \approx 1.1\ \mathrm{m/s}$$

Finally, we would conclude that our measured speed of sound is $(\bar{v} \pm \sigma_{\bar{v}}) = (341 \pm 1)$ m/s. Notice that the number of decimal places in the measured value and its uncertainty agree (in this case, both have zero decimal places).

2 Least Squares Fitting

Often you will determine a quantity by examining a linear plot of N collected data points (x_1, y_1), (x_2, y_2), ... (x_i, y_i), ... (x_N, y_N). We will call these (x_i, y_i) where i may have any value from 1 to N.

Because of measurement uncertainties, the plot of these points will not exactly define a straight line, and a number of different lines can be drawn through the data points. The

problem becomes one of 'goodness of fit' — which of the many possible different lines we can possibly draw is the 'best' fitting one?

The line with the best fit can be fairly easily determined if we make three basic assumptions about the nature of our measurement uncertainties:

1. The absolute uncertainties are nearly the same for all data points (we can use single uncertainty values σ_x and σ_y).
2. The uncertainties are principally in the dependent variable – the measured y_i, with effectively trivial uncertainties ($\sigma_x \sim 0$) in the independent variable x_i.
3. The uncertainties are random in nature (not systematic or due to human error).

Given these three assumptions, we can statistically determine the equation for the best line by calculating the sum of the squares of the distances between the theoretical line and our data points, and then by minimizing this sum of squares. This is done by taking partial derivatives of that theoretical quantity, and so we will not derive the formulas here (although the derivation will be shown in lecture to interested parties). This whole process of minimizing the squares of the distances (the uncertainties) is widely known as the *Least Squares Fit* algorithm. The algorithm is given below:

$$m = \frac{N(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\Delta} \tag{8}$$

$$b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{\Delta} \tag{9}$$

where $$\Delta = N(\sum x_i^2) - (\sum x_i)^2$$

You should be aware that $(\sum x_i)^2 \neq \sum x_i^2$.

2.1 Uncertainties in Least Squares Fitting

When calculating the uncertainties in a least squares fit, we must first calculate σ_y, which gives the standard deviation of the y values from the straight line. This value is given by:

$$\sigma_y^2 = \frac{1}{N-2} \sum (y_i - b - m x_i)^2. \tag{10}$$

Knowing σ_y^2, the values for the uncertainties in m and b can be found from:

$$\sigma_{\bar{m}}^2 = \frac{N\sigma_y^2}{\Delta} \tag{11}$$

$$\sigma_{\bar{b}}^2 = \frac{\sigma_y^2 \sum x_i^2}{\Delta} \tag{12}$$

Note that $\Delta = N(\sum x_i^2) - (\sum x_i)^2$ is a very useful quantity to evaluate early and have at hand when calculating least square fits. Also, you will need to take $\sqrt{\sigma_{\bar{m}}^2}$ and $\sqrt{\sigma_{\bar{b}}^2}$ to find the final uncertainties and write $(m \pm \sigma_{\bar{m}})$ and $(b \pm \sigma_{\bar{b}})$.

Example

1. Data collected by a Physics 152L student to describe the motion of a skier who moved from a level section of a ski run onto a smooth slope at $t = 0$ seconds is compiled in Table 1.

Time t (s)	Uncertainty in time δt (s)	Velocity v (m/s)	uncertainty in v δv (m/s)
0.50	0.01	7.10	0.20
0.75	0.01	7.90	0.20
1.00	0.01	8.50	0.40
1.25	0.01	9.50	0.30
1.50	0.01	10.00	0.40
1.75	0.01	10.50	0.40
2.00	0.01	11.20	0.30

Table 1: Velocity and time data describing a skier's descent

If we assume that the velocity data can be fitted by the linear equation:

$$v(t) = v_i + at, \tag{13}$$

we can determine the initial velocity v_i and the acceleration a of the skier on the slope.

(a) Perform a least squares fit upon the data, and determine v_i, a, and their associated uncertainties $\sigma_{\bar{v}_i}$ and $\sigma_{\bar{a}}$.

(b) Graph these data following the standards for graphical presentation of laboratory data.

(c) What was the skier's initial velocity as she started into the slope?

(d) What was her acceleration on the slope?

The Solution

(a) First we need to calculate a series of statistics from our data. We will use Table 2 to assist in the least squares fitting. (Note that we cannot calculate the final column until *after* we determine the values for b and m from our least squares fit).

i	x_i	y_i	x_i^2	$x_i y_i$	$(y_i - b - mx_i)^2$
	time t (s)	velocity v (m/s)	t^2 (s^2)	vt (m)	(m/s)^2
1	0.50	7.10	0.2500	3.5500	0.0115
2	0.75	7.90	0.5625	5.9250	0.0002
3	1.00	8.50	1.0000	8.5000	0.0041
4	1.25	9.50	1.5625	11.8750	0.0661
5	1.50	10.00	2.2500	15.0000	0.0062
6	1.75	10.50	3.0625	18.3750	0.0100
7	2.00	11.20	4.0000	22.4000	0.0062
N	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum(y_i - b - mx_i)^2$
7	8.75	64.70	12.6875	85.6250	0.1043

Table 2: A least squares fit table

Always carry extra (non-significant) decimal places during LSQ fit calculations, then round to the correct number of significant digits when you finally interpret the results. Otherwise, each time you make a calculation you will introduce a round-off error. In this example, we will carry at least two non-significant digits throughout, and will dispose of these at the end.

Then using the values from Table 2:

$\Delta = N(\sum x_i^2) - (\sum x_i)^2 = (7)(12.6875) - (8.75)^2 = 12.2500$ seconds^2

$m = \frac{N(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\Delta} = \frac{(7)(85.6250) - (8.75)(64.70)}{12.2500} = 2.7143 \sim 2.71$ m/s^2

$b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{\Delta} = \frac{(12.6875)(64.70) - (8.75)(85.6250)}{12.2500} = 5.8500 \sim 5.85$ m/s

Then to calculate the uncertainties:

$\sum(y_i - b - mx_i)^2 = 0.1043$ (m/s)^2

$\sigma_y^2 = \frac{1}{N-2}\sum(y_i - b - mx_i)^2 = \frac{0.1043}{7-2} = 0.0209$ (m/s)^2

$\sigma_{\bar{m}} = \sqrt{\frac{N\sigma_y^2}{\Delta}} = \sqrt{\frac{(7)(0.0209)}{12.2500}} = 0.1092 \sim 0.11$ m/s^2

$\sigma_{\bar{b}} = \sqrt{\frac{\sigma_y^2 \sum x_i^2}{\Delta}} = \sqrt{\frac{(0.0209)(12.6875)}{12.2500}} = 0.1470 \sim 0.15$ m/s

Therefore $m = (2.71 \pm 0.11)$ m/s^2 and $b = (5.85 \pm 0.15)$ m/s .

(b) The complete plot of this data was previously seen in this activity as Figure 2. Note that, as required in Physics 152L, the slope and y-intercept have been labeled and their meanings have been noted.

(c) The skier's initial velocity as she started into the slope was $v_i = 5.85$ m/s, $\sigma_{\bar{v}_i} = 0.15$ m/s.

(d) Her acceleration on the slope was $a = 2.71$ m/s^2, and $\sigma_{\bar{a}} = 0.11$ m/s^2. Therefore $(a \pm \sigma_{\bar{a}}) = (2.71 \pm 0.11)$ m/s^2 and $(v_i \pm \sigma_{\bar{v}_i}) = (5.85 \pm 0.15)$ m/s .

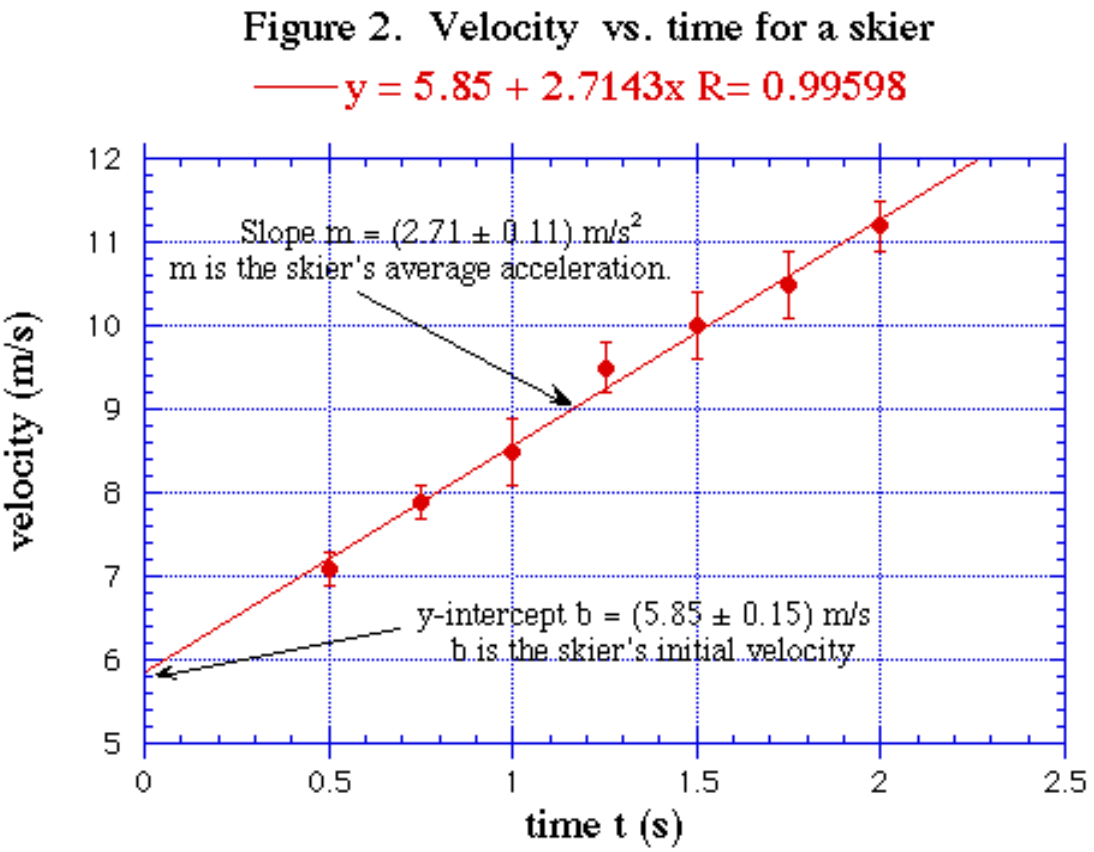

Figure 2: Example of a graph meeting all of the Physics 152L standards described in the next section. The uncertainties in the slope m and the y-intercept b are from the least squares fit performed above. The equation for the line and the value of the Linear Correlation Coefficient (R) displayed under the main title was generated by *KaleidaGraph* is optional; it serves as a quick check of whether the least squares calculations are correct.

All of the enclosed graphs and statistical fits were computer-generated using hardware and software available for your use in PHYS 14 (next to your laboratory room). There are samples and explanation files for PHYS 152L graphs and fits on those machines as well as student monitors paid to assist you with your work during weekly open hours. Please explore and make use of these facilities. *You must use computer software to generate plots and least squares fits as a requirement for completing PHYS 152L, and this activity (MA2) must be completed using computer software for the calculations and the plots.* These skills are extremely useful to science and engineering practitioners and students. We recommend that you use *Kaleidagraph* for your plots; it is much easier to use than *Excel* for plotting. *Kaleidagraph* may be accessed in the PUCC Macintosh laboratories, and working demonstration

copies of it may be downloaded from `http://www.synergy.com` for both the Macintosh and PC platforms.

3 Graphing standards

All laboratory graphs must follow the guidelines listed below.

1. **Titles and labels.** All graphs should be clearly labeled with both a figure number and/or an explanatory title directly above the graph. The title text should briefly explain the graph independently of other text elsewhere.

2. **Axes.** Both the horizontal and vertical axes of graphs should be clearly labeled with variable names and units, and numerical values at major tickmarks should be given. The conventional way of deciding what quantity goes with which axis is to plot the dependent or measured value is on the upright axis and the independent or controlling variable on the horizontal axis. For example, a plot of velocity vs. time (i.e., $v(t)$) would show velocity on the vertical axis and time on the horizontal axis.

3. **Error bars.** When plotting points with known uncertainties, error bars should be included. If your plotting software does not allow error bars, pencil them in on your graph.

4. **Slope and y-intercept.** When plotting linear relationships, the slope and y-intercept of the line are of interest. These values and their units should also be clearly displayed on the graph. Sometimes the x-intercept is also of interest; if this is the case, it should also be labeled and its value indicated. In Physics 152L the physical significance of the the slope and y-intercept should be stated, as shown in Figure 2.

5. **Size and clarity.** In order to express all of this information clearly and legibly, you should choose axis limits so that the region of interest occupies most of your graph. Labeling should be done with reasonably large size numbers and letters.

4 Expressing Linear Relationships

Often you will determine a quantity by examining a plot of collected (x, y) data points. If the quantity plotted on the vertical axis depends linearly on the quantity plotted along the horizontal axis, ideally (ignoring measurement uncertainties) these data points will fall on a straight line whose equation can be written in the form:

$$y = mx + b \tag{14}$$

where b is known as the y-intercept and m is the slope. The slope can be further defined as

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} \tag{15}$$

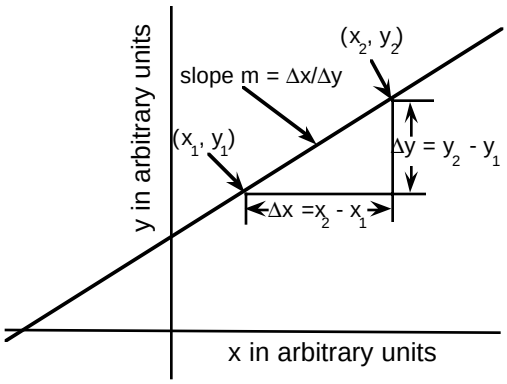

Figure 3: Illustration of the slope of a straight line.

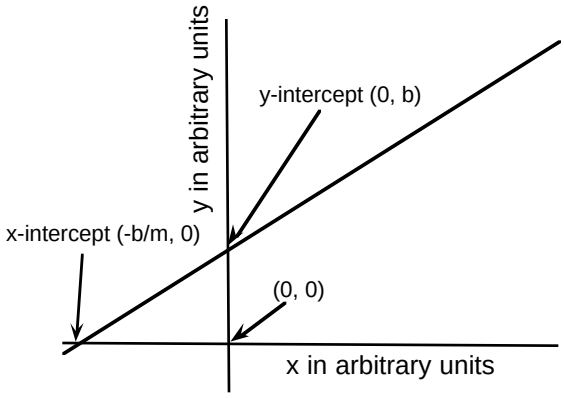

Figure 4: x- and y-intercepts of a straight line.

where (x_1, y_1) and (x_2, y_2) are two points on the line. This is illustrated in Figure 3.

The y-intercept is the position where a linear graph crosses the y-axis and corresponds to the value for which the value of x is zero. Sometimes we need to know the x-intercept rather than the y-intercept; if this is the case we can apply the fact that $y = 0$ at the x-intercept to our equation for a straight line as follows:

$$y = 0 = mx_{int} + b,$$

$$mx_{int} = -b$$

Thus,

$$x_{int} = -\frac{b}{m} \tag{16}$$

and we can readily determine the x-intercept knowing the y-intercept and the slope. The x- and y-intercepts of a straight line are shown in Figure 4.

5 References

1. Bevington, P. R., *Data reduction and uncertainty analysis for the physical sciences* (McGraw-Hill, New York, 1969).

2. Taylor, J. R., *An introduction to uncertainty analysis: the study of uncertainties in physical measurements* (University Science Books, Mill Valley, CA, 1982).

3. Young, H. D., *Statistical treatment of experimental data* (McGraw-Hill, New York, 1962).

Measurement Analysis Problem Set MA2

Name _____ Lab day/time _____

Division _____ GTA _____

This assignment is due at the start of E3 for your division. Some portions of this assignment must be done by hand, and other parts must be done with a spreadsheet plus graphing software for full credit.

1. The mass of an object was measured eight times on a balance readable to ±1 g, and the following values were recorded: 416 g, 423 g, 399 g, 410 g, 410 g, 417 g, 402 g, and 431 g. Find $\bar{m}$, s_m, and $\sigma_{\bar{m}}$ by hand. Carry extra digits through calculations and rationalize your numbers at the end, showing all work.

   (a) $\bar{m} = ($) g

   (b) $s_m = ($) g

   (c) $\sigma_{\bar{m}} = ($) g

   (d) Write the final measurement $(\bar{m} \pm \sigma_{\bar{m}}) = ($ $\pm$) g

   (e) What is the percentage uncertainty of the measurement? $\frac{\sigma_{\bar{m}}}{\bar{m}} \times 100\% =$ %

2. The following data were collected by a Physics 152 student to describe the motion of a motorcycle racer who applied the brakes at $t = 0$ seconds:

Time t (s)	Uncertainty in time δt (s)	Velocity v (m/s)	uncertainty in v δv (m/s)
0.20	0.01	14.0	0.4
0.40	0.01	12.6	0.4
0.60	0.01	11.0	0.3
0.80	0.01	9.7	0.3

If we assume that the velocity data can be fitted by the linear equation:

$$v(t) = v_i + at \tag{17}$$

we can determine the initial velocity v_i and the acceleration a of the motorcyclist.

(a) How does Equation 17 relate to the standard equation for a line? Identify the y-intercept and the slope of Equation 17.

(b) Create an appropriate table and perform a least squares fit upon the data, and determine the slope and intercept of the graph and their associated uncertainties. Perform all calculations for this first LSQ fit by hand – write your answers on a separate sheet of paper, showing all calculations. Carry at least two extra decimal places through all statistical calculations, and round appropriately when expressing the final results. Use the correct units.

(c) Using a software plotting package (preferably *Kaleidagraph*; *Excel* is fine but harder to use for graphing with labels, etc.), graph these data following the standards for graphical presentation of laboratory data. Label the slope and y-intercept and say what information they provide about the motion of the motorcyclist. Attach your plot to this page.

(d) What was the driver's initial velocity when the brakes were first applied?

(e) What was the driver's acceleration (deceleration)?

3. The linear relationship between the restoring force applied by a spring and its distortion is known as Hooke's Law and is written as:

$$F_{restoring} = -kx \tag{18}$$

where $F_{restoring}$ is the restoring (resisting) force applied by the spring, k is a positive quantity known as the spring constant and x is the amount the spring is stretched from its natural equilibrium position x_e. If the total length of the stretched spring is x_s, $x = x_s - x_e$. The negative sign means that the force is opposite in direction to the displacement—the spring tries to return to its original size by pulling against the direction of the stretch. Typical values for k for lab springs are on the order of one newton per meter (1 N/m).

In the laboratory, it is fairly easy to apply a force to the spring and the displacement or stretch of the spring. This applied force $F_{applied}$ is equal in size and opposite in direction to the restoring force $F_{restoring}$ such that:

$$F_{applied} = -F_{restoring} = -(-kx) = kx \tag{19}$$

Therefore, you will plot $F_{applied}$ vs. x, where $x = x_s - x_e$ is the stretch of the spring.

The following data were collected to describe the stretch of a spring in a Physics 152 lab:

Force F applied (N)	Amount of stretch $x = x_s - x_e$ (m)
0.203	0.250
0.409	0.455
0.596	0.676
0.810	0.882

(a) Use a spreadsheet to prepare an appropriate table for performing a least squares fit on the data and determine the best values for the slope and y-intercept along with their uncertainties. Print out your spreadsheet, showing the intermediate sums and calculations on the sheet so the essential steps are obvious. *Note: this same spreadsheet can be re-used to solve the LSQ fits in E3 lab and E6 lab.*

(b) Plot *Applied force F vs. Stretch* $x = x_s - x_e$ from this data according to laboratory graphing standards. Label the slope and y-intercept and interpret them. Note that you have no uncertainty bars for this example.

(c) What is the value of the spring constant and its uncertainty?

(d) What should the value for the y-intercept be according to Equation 18? Does the data agree with theory within the uncertainties given by the LSQ fit?