# Appendix A

# Data Analysis with Excel

Computers are used for data analysis in any modern physics laboratory, and the Physics 103 lab is no exception. We have built our data analysis system around the program Excel, which is widely used on and off campus. We've added some Workshop Physics (WP) tools to make graphing data easier, and to let you do regression calculations with uncertainties, but otherwise we are using the standard, off-the-shelf software.[1]

If you are already familiar with Excel, great! If not, we'll give brief instructions here. Like any software, it can be confusing at first, so don't hesitate to ask your instructors and your fellow students for help. Play around with the program a bit to get comfortable with it.

## A.1 Starting Things Up

- If the computer isn't already on, turn it on and wait for it to boot up.
- If the Physics 103 window isn't already open, double click on the Physics 103 icon to open it
- Double click on Excel with WPtools (look for the X logo) to get the program running.

## A.2 Entering Data: a Simple Example[1]

When Excel is started up, you need to open a spreadsheet to work in. If you are asked if you want to reopen WPtools, click No. Then go to File→New and click on OK to open a new Workbook. (If you wanted to open a pre-existing spreadsheet, you would use the File→Open menu command; if you want to save the new spreadsheet, use the File→Save menu command. Since we'll be working with fairly small datasets, neither of these is really necessary for your lab work.)

Suppose you wish to record some $(x, y)$ data pairs in two columns of a spreadsheet. Go to Excel, and start entering the data in the upper left most cell (called A1). To do this, move the cursor to this cell and click on it with the left mouse button. Enter the first $x$ data value

[1]Some online documentation for WPtools is at
http://physics.dickinson.edu/~wp_web/wp_resources/Documentation.html

[1]Note: Menu commands are described as follows: File→Open means move the cursor to the word File on the line near the top of the screen, press *and hold* the **left** mouse button, drag the cursor down to the word Open, and release the button.

here, pressing Return when you are done. The cell below it (called A2) will automatically be selected next; enter the second $x$ data value here. Work down the first data column in this way. If you need to go back and correct any of the numbers, simply move the cursor with the mouse, click on the relevant cell, and re-enter the number.

Once you've entered the first column of numbers, move the cursor to the top cell of the second column (cell B1) and click the left button to select it. Enter the first $y$ data value. Press Return/Enter, and enter subsequent $y$ data in the rest of the cells.

## A.3 Calculations in Excel

Now that your two columns of data are in the computer, select them. Do this by moving the cursor to the top left cell, pressing *and holding* the **left** mouse button, dragging the cursor to the lowest filled cell in the second column, and then releasing the mouse button. The block of numbers you entered will now be "selected", indicated by a blue-grey color.

Go to the WPtools pull-down menu and select Linear Fit. Immediately Excel will display a plot of your data, along with the values and uncertainties of the best-fit line. Print out a copy of the results if desired.[2]

Sometimes you will want to transform your raw data in some way before plotting it. For example, you may have entered two columns of data as above, but you want to convert the $y$ values from inches to meters. This is where a spreadsheet program becomes really handy. Select a blank cell somewhere on the sheet (cell C1 would be a good place). Instead of entering a number, enter the formula '=0.0254*B1, and press Return. Excel will display the expected numerical value in cell C1, and it will also remember the formula. This is useful for two reasons, first, if you change the value in B1, the number in C1 will be automatically updated. Second, you can copy the formula in C1 to other cells, transforming the rest of column B using the same formula. To do this, first select cell C1. The cell becomes outlined, and note that there is a little square in the lower right corner of the outline. Move the cursor to this square, push and hold the **left** mouse button, drag the cursor down several cells, and release the mouse button. Voila! Excel will use the same formula to multiply all the cells in column B by 0.0254.

Excel can do much more complicated arithmetic. For example, you could use the formula =sqrt(A1) * B1 to take the square root of the values of cells in column A, multiply them by the values in column B, and put the result in some other column.

You might also want to take differences between the successive items in your data list. If you type into cell C2 the formula =B2-B1, and then use the little square to fill this formula into the cells B3, B4, *etc.*, then you will obtain the differences in column C.

If you do a transformation like this, and then you want to do a plot or a curve fit, the columns of data you want to plot may not be adjacent to each other. No problem. Say you want to plot cells A1-A10 on the horizontal axis and cells C1-C10 on the vertical axis. First

[2] You must first "grab" the plot by left-clicking on an open area inside it. If the plot legends are obscuring the graph, drag them aside with the mouse. Your can add labels to your plot using the Edit labels option on the WPtools menu bar.

select A1-A10. (Go to cell A1, hold down the left button, drag the cursor to A10, then release the mouse button). Then *hold down the* Ctrl *key* and select C1-C10. Now both cell groups A1-A10 and C1-C10 will be selected, but not B1-B10. Run the WPTools→LinearFit routine, and you will get the plot you want.

## A.4 Accumulating Values via Excel Tricks

There will be times in the Physics 103 lab when you want to accumulate sums of a series of values. For example, you might have measured a series of time intervals,

$\Delta t_1$ = interval between event 1 and event 2,
$\Delta t_2$ = interval between event 2 and event 3,
$\Delta t_3$ = interval between event 3 and event 4,
*etc.*

You may wish to convert these into a continuous time scale. In other words, you may want to declare that $t = 0$ at the time of event 1, and then find

time of event 2 = $\Delta t_1$,
time of event 3 = $\Delta t_1 + \Delta t_2$,
time of event 4 = $\Delta t_1 + \Delta t_2 + \Delta t_3$,
*etc.*

This is easy to do. Say that $\Delta t_1$, $\Delta t_2$, *etc.*, are in cells A1, A2, *etc.*, and you want to put the accumulated times in column B. First put a 0 in cell B1 (since $t = 0$ for the first event). Then go to cell B2 and enter the formula =SUM($A$1:A1). The SUM function simply adds up the cells in the range specified.

The usefulness of the $ notation becomes apparent when you want to calculate the rest of the times. Select B2, move the cursor to the square in the lower righthand corner of the cell border, press and hold the left mouse button, drag the cursor down several cells, and release the button. The cells in column B are now filled with SUM functions, but in a special way: The $A$1 in the SUM function call remains the same in all the cells (because of the $), but the second part of the function call changes from A1 to A2 to A3, ... In other words, cell B3 now reads =SUM($A$1:A2), cell B4 reads =SUM($A$1:A3), and so on. These are exactly the formulae we want for the event time calculations, so column B is now filled with calculated values of $t$.

## A.5 Further Notes about Workshop Physics Routines

- Use the WPtools→Polynomial Fit menu command, and set Order=2 to fit lines of the form $y = c_0 + c_1 x + c_2 x^2$.
- If you enter non-numerical text in the cell above each column of data, it will be used to label the horizontal and vertical axes on the plot.

- The fitting routines always use the first selected column for the horizontal ($x$) points, and the second selected column for the vertical ($y$) points.
- Empty rows are usually ignored (but partially-empty rows may corrupt the fit).
- To delete a plot, select it (move cursor to it and click once), then press the Delete key. To delete a column of the sheet, select the entire column (by clicking on the letter at the top) and use Edit→Delete.
- If data are modified after running a fit, the associated plot will be automatically updated, but the fit parameters will not be re-calculated. Usually it is best to delete both the old plot and fit parameters after updating data.

# Appendix B

# Estimation of Errors

While the subject of error analysis can become quite elaborate, we first emphasize a basic but quite useful strategy, discussed in secs. B.1-2. Then, we distinguish between random (or statistical uncertainties and systematic uncertainties in sec. B.3. Random uncertainties follow the famous bell curve, as sketched in secs. B.4-5. The important distinction between the uncertainty on a single measurement, and the uncertainty on the average of many repeated measurements is reviewed in secs. B.7-7. The subject of propagation of errors on measured quantities to the error on a function of those quantities is discussed in sec. B.8.

## B.1 67% Confidence

Whenever we make a measurement of some value $v$, we would also like to be able to say that with 2/3 probability the value lies in the interval $[v-\sigma, v+\sigma]$. We will call $\sigma$ the uncertainty or error on the measurement. That is, if we repeated the measurement a very large number of times, in about two thirds of those measurements the value $v$ would be in the interval stated.

## B.2 A Simple Approach

Repeat any measurement three times, obtaining a set of values $\{v_i\},\ i = 1, 2, 3$. Report the average (mean),

$$\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_i \qquad (\text{for } N = 3), \tag{B.1}$$

as the best estimate of the true value of $v$, and the uncertainty $\sigma$ as

$$\sigma = \frac{v_{\max} - v_{\min}}{2}. \tag{B.2}$$

If you take more than three measurements, you can still implement this procedure with the aid of a histogram. Divide the range of observed values of $v$ into 5-10 equal intervals (called bins). Located the bin that contains each measurement, and draw a box one unit high above that bin. Stack the boxes on top of one another if more than one measurement falls in a bin. To estimate the error, determine the interval in $v$ that contains the central 2/3 of the measurements, *i.e.*, the central 2/3 of the boxes you just drew, and report the error as 1/2 the length of this interval.

## B.3 Random and Systematic Uncertainties

The uncertainty in a measurement of a physical quantity can be due to intrinsic random uncertainty (colloquially: error) as well as to systematic uncertainty.

Random uncertainties lead to difference in the values obtained on repetition of measurements. Systematic uncertainties cause the measurement to differ from its ideal value by the same amount for all repetitions of the measurement.

Random uncertainties can arise from vibrations of the components of a set-up driven by random thermal fluctuations, random noise in the electronics, and/or many other small but uncontrolled effects including quantum fluctuations.

In principle, the effect of random uncertainties can be made as small as desired by repetition of the measurements, such that the dominant uncertainty is due to systematic effects (which can only be reduced by designing a better measurement apparatus).

## B.4 The Bell Curve

In many cases when a measurement is repeated a large number of times the distribution of values follows the bell curve, or Gaussian distribution:

$$P(v) = \frac{e^{-(v-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}, \tag{B.3}$$

where $P(v)dv$ is the probability that a measurement is made in the interval $[v, v+dv]$, $\mu$ is true value of the variable $v$, and $\sigma$ is the standard deviation or uncertainty in a single measurement of $v$. See Figure B.1.

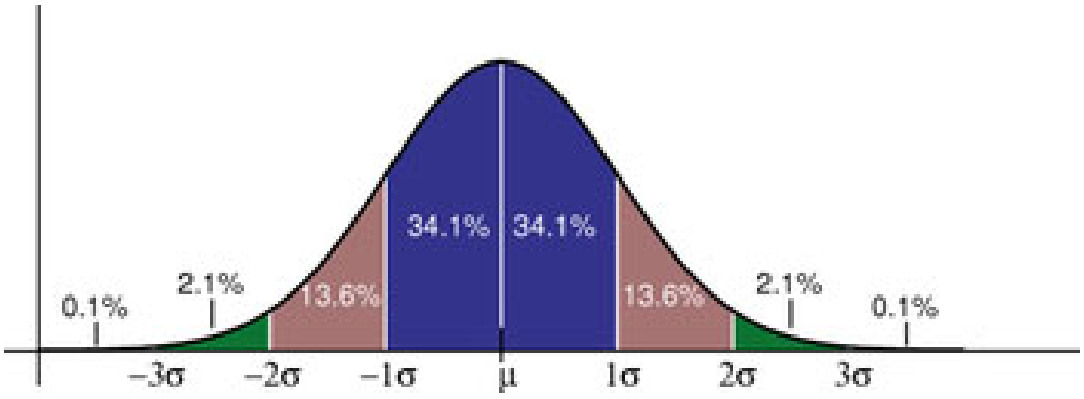


Figure B.1: The probability distribution measurements of a quantity with true value $\mu$ and Gaussian uncertainty $\sigma$ of a singe measurement. About 68% of the measurements would fall in the interval between $\mu - \sigma$ and $\mu + \sigma$, and 95% would fall in the interval $\mu \pm 2\sigma$.

The Table lists the confidence that a single measurement from a Gaussian distribution falls within various intervals about the mean. If the 100 students in Ph103 each make 100

Table B.1: The probability (or confidence) that a measurement of a Gaussian-distributed quantity falls in a specified interval about the mean.

| Interval | Confidence |
|---|---|
| $\pm\sigma$ | 68% |
| $\pm 2\sigma$ | 95% |
| $\pm 3\sigma$ | 99.7% |
| $\pm 4\sigma$ | 99.994% |

measurements during these lab sessions, then 10,000 measurements will be taken in all. The Table tells us that if those measurements have purely Gaussian 'errors', then we expect one of those measurements to be more than $4\sigma$ from the mean.

## B.5 Estimating Uncertainties When Large Numbers of Measurements Are Made

One can make better estimates of uncertainties if the measurements are repeated a larger number of times. If $N$ measurements are made of some quantity resulting in values $v_i,\ i = 1, ...N$ then the mean is, of course,

$$\bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_i, \tag{B.4}$$

and the standard deviation of the measurements is

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(v_i - \bar{v})^2}. \tag{B.5}$$

Calculus experts will recognize that the operation $(1/N)\sum_{i=1}^{N}$ becomes $\int P(v)\,dv$ in the limit of large $N$. Then, using the Gaussian probability distribution (B.3) one verifies that

$$\bar{v} = \langle v \rangle = \int_{-\infty}^{\infty} vP(v)\,dv, \qquad \text{and} \qquad \sigma^2 = \left\langle (v-\bar{v})^2 \right\rangle = \int_{-\infty}^{\infty} (v-\bar{v})^2 P(v)\,dv. \tag{B.6}$$

## B.6 The Uncertainty on Mean of a Uniformly Distributed Quantity

Not all measurable quantities follow the Gaussian distribution. A simple example is a quantity with a uniform distribution, say with values $v$ equally probable over the interval $[a, b]$. It is clear that the average measurement would be $(a + b)/2$, but what is the uncertainty of the measurement? If we adopt the simple prescription advocated in secs. B.2 we would

report the uncertainty as $(b-a)/3$ since 2/3 of the time the measurement would fall in an interval $2(b-a)/3$ long. If instead we use the calculus prescription for $\sigma$ given in eq. (B.6) we find that

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{b-a}{3.46}, \tag{B.7}$$

which result is often used by experts.

## B.7 The Uncertainty in the Mean

Thus far we have considered only the uncertainty or spread in measured values of some quantity $v$. A related but different question is: what is the uncertainty on our best estimate of $v$ (which is just the mean value of our measurements, $\bar{v} = (1/N)\sum v_i$)?

The uncertainty on the mean $\bar{v}$ is surely less that the uncertainty, $\sigma$, on each measurement $v_i$. Indeed, the uncertainty on the mean is given by

$$\sigma_{\bar{v}} = \frac{\sigma}{\sqrt{N}}, \tag{B.8}$$

where $\sigma$ is our estimate of the measurement error obtained from one of the methods sketched previously.

Appendix C illustrates eq. (B.8) using measurements of $g$ from past Ph103 labs.

## B.8 The Uncertainty on a Function of Several Variables (Propagation of Error)

In many cases we are interested in estimating the uncertainty on a quantity $f$ that is a function of measured quantities $a,\ b,\ ...\ c$. If we know the functional form $f = f(a, b, ...c)$ we can estimate the uncertainty $\sigma_f$ using some calculus. As a result of our measurements and the corresponding 'error analysis' we know the mean values of $a, b,\ ...\ c$ and the error estimates $\sigma_a,\ \sigma_b,\ ...\ \sigma_c$ of these means. Our best estimate of $f$ is surely just $f(a, b, ...c)$ using the mean values.

To estimate the uncertainty on $f$ we note that the change in $f$ due to small changes in $a,\ b, ...\ c$ is given by

$$\Delta f = \frac{\partial f}{\partial a}\Delta a + \frac{\partial f}{\partial b}\Delta b + ... + \frac{\partial f}{\partial c}\Delta c. \tag{B.9}$$

If we just averaged this expression we would get zero, since the 'errors' $\Delta a,\ ...\ \Delta c$ are sometimes positive, sometimes negative, and average to zero. Rather, we square the expression for $\Delta f$, and then average.

$$\Delta f^2 = \left(\frac{\partial f}{\partial a}\right)^2 \Delta a^2 + ... + \left(\frac{\partial f}{\partial c}\right)^2 \Delta c^2 + ... + 2\frac{\partial f}{\partial a}\frac{\partial f}{\partial c}\Delta a \Delta c + ... \tag{B.10}$$

On average the terms with factors like $\Delta a \Delta c$ average to zero (under the important assumption that parameters $a,\ b,\ ...\ c$ are independent). We identify the average of the squares of

the changes relative to the mean values as the squares of the errors: $\langle \Delta a^2 \rangle = \sigma_a^2$, *etc.* This leads to the prescription

$$\sigma_f^2 = \left(\frac{\partial f}{\partial a}\right)^2 \sigma_a^2 + ... + \left(\frac{\partial f}{\partial c}\right)^2 \sigma_c^2 + ... \tag{B.11}$$

Some useful examples are

$$f = a \pm b \pm ... \pm c \qquad \Rightarrow \qquad \sigma_f = \sqrt{\sigma_a^2 + \sigma_b^2 + ... + \sigma_c^2}, \tag{B.12}$$

and

$$f = a^l b^m ... c^n \qquad \Rightarrow \qquad \frac{\sigma_f}{f} = \sqrt{l^2 \left(\frac{\sigma_a}{a}\right)^2 + m^2 \left(\frac{\sigma_b}{b}\right)^2 + ... + n^2 \left(\frac{\sigma_c}{c}\right)^2}, \tag{B.13}$$

where $l$, $m$ and $n$ are constants that may be negative.

For more detailed and rigorous analyses one can consult, for example:

- P.R. Bevington and D.K. Robinson, *Data Reduction and Error Analysis for the Physical Science*, 2nd ed. (McGraw-Hill, New York, 1992).
- J.R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd ed. (University Science Books, 1997).

# Appendix C

# Standard Deviation of the Mean of $g$

Suppose you make $N$ repeated measurements of a quantity $g$, such as the acceleration due to gravity. How well is the value of $g$ determined by these measurements?

For example, during the 2006 sessions of Ph103 Lab 6 a total of 37 different measurements of $g$ were made, as shown in the histogram Fig. C.1.

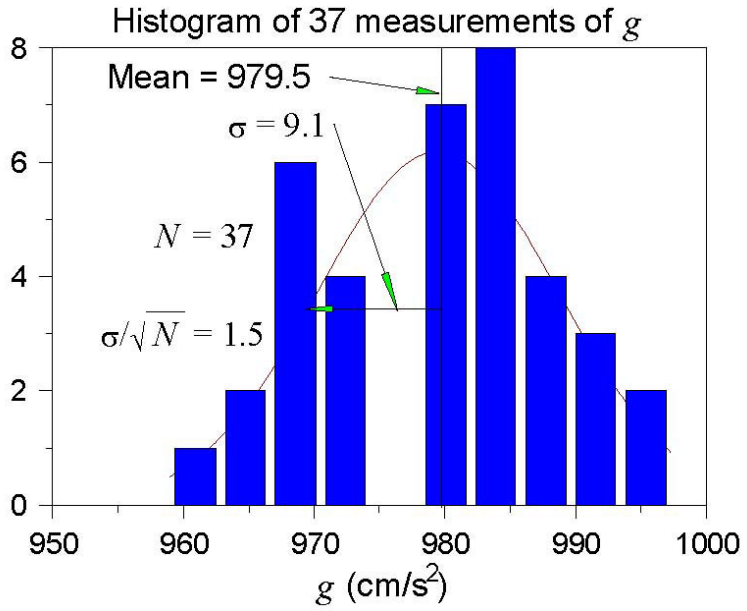


Figure C.1: Histogram of the values of $g$ measured in the 2006 Ph103 Lab 3. The horizontal axis is $g$, and the vertical axis is the number of times a value of $g$ was reported to lie with the range of $g$ corresponding to the width of a vertical bar.

*A histogram is a graph containing $M$ vertical bars in which the height of a bar indicates the number of data points whose value falls within the corresponding "bin", i.e., within the interval $[g_j - \Delta/2, g_j + \Delta/2]$, where $g_j, j = 1, M$ and the centers of the $M$ bins and $\Delta$ is the bin width. One can make a histogram of a data set $\{g_i\}$ using* Excel/Tools/Data Analysis/Histogram. *Enter the data $\{g_i\}$ in one column of an* Excel *spreadsheet. Click on the* Input Range: *box of the* Histogram *window; then click and hold the left mouse button on the first data point, and drag the mouse to the last data point to enter the cell addresses of the data. Click on* Chart Output *and then* OK *to create a basic histogram. If the number/spacing of "bins" chosen by* Excel *is awkward, fill a new column with a linear series of 5-10 steps that begins near the lowest $g_i$ and ends near the highest; create a new histogram with the* Excel *addresses of the first and last elements of the bin list in the box* Bin Range:.

The mean value $\bar{g}$ is calculated according to

$$\bar{g} = \frac{\sum_{i=1}^{N} g_i}{N}\,, \tag{C.1}$$

and was found to be $\bar{g} = 939.5$ cm/s$^2$ for the data shown in Fig. C.1.

The distribution of the value of $g$ is approximately Gaussian, and the standard deviation of this distribution is calculated according to

$$\sigma_g = \sqrt{\frac{\sum_{i=1}^{N}(g_i - \bar{g})^2}{N-1}}\,, \tag{C.2}$$

with the result that $\sigma_g = 9.1$ cm/s$^2$.

The standard deviation $\sigma_g$ is a good estimate of the uncertainty on a single measurement of $g$. However, after 37 measurements of $g$, the uncertainty on the mean value $\bar{g}$ is much smaller than $\sigma_g$.

An important result of statistical analysis is that the standard deviation (*i.e.*, the uncertainty) of the mean of the $N$ measurements is related to the standard deviation of the distribution of those measurements by,

$$\sigma_{\bar{g}} = \frac{\sigma_g}{\sqrt{N}}\ . \tag{C.3}$$

For the data shown in Fig. C.1, where $N = 37$, we obtain

$$\sigma_{\bar{g}} = \frac{9.1}{\sqrt{37}} = 1.5 \text{ cm/s}^2. \tag{C.4}$$

That is, we can report the result of all 37 measurements of $g$ as

$$g = 979.5 \pm 1.5 \text{ cm/s}^2. \tag{C.5}$$

As a check that eq. (C.3) is valid, we can analyze the data another way. Namely, we can first calculate the means $\bar{g}_i$ for the 5 different sessions of Ph103 Lab 3. Then, we can make a histogram of these 5 values, as shown in Fig. C.2.

The mean of the 5 means is 979.6 cm/s$^2$, which is essentially identical to the mean of the 37 individual measurements of $g$. The standard deviation of the 5 means shown in Fig. C.2 is calculated to be 1.6 cm/s$^2$, which is essentially identical to the previous calculation (C.4) of the standard deviation of the mean.

**Concluding Remarks:** If $N$ were much larger than what we have here, the histogram C.1 would approach the Gaussian distribution (the bell-curve) shown in Appendix B. The peak in the histogram would be very close to the mean value $\bar{g}$ of the measurements, which represents the best estimate of $g$ from the data. The standard deviation $\sigma_g \approx$ width/2 is a measure of the uncertainty of a single measurement,[1] while $\sigma_g/\sqrt{N}$ is the uncertainty on the best estimate $\bar{g}$.

[1] Strictly speaking, the full width at half maximum of a Gaussian distribution is $2.35\sigma_g$.

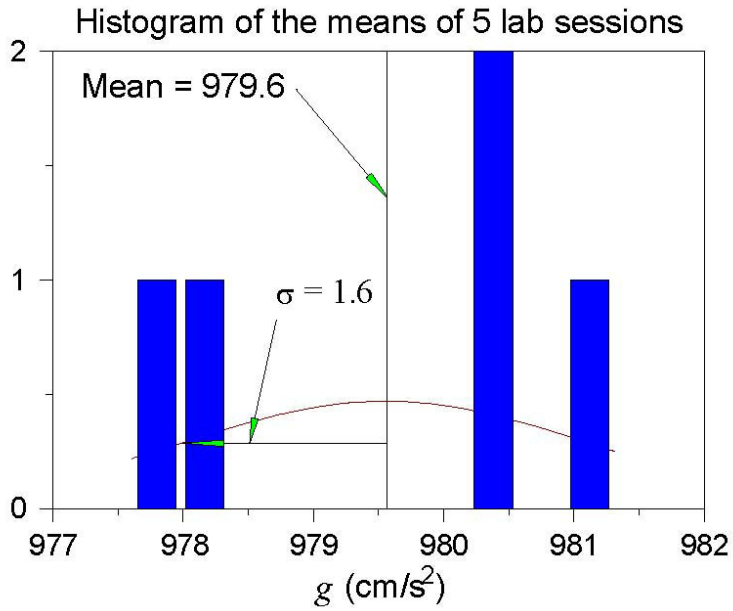


Figure C.2: Histogram of the mean values of $g$ measured in the 5 sessions of Ph103 Lab 6 in 2006.

# Appendix D

# Polynomial Fits in WPtools

## D.1 Polynomial Regression

In this technical Appendix we sketch the formalism used in the polynomial regression method for fitting data. This is a generalization of the method of linear regression.

We start with a set of data $(x_j, y_j)$, $j = 1, ...m$, and we wish to fit these data to the $n$th-order polynomial

$$y(x) = \sum_{i=0}^{n} a_i x^i. \tag{D.1}$$

In general each measurement $y_j$ has a corresponding uncertainty $\sigma_j$. That is, if the measurements were repeated many times at coordinate $x_j$ the values of $y_j$ would follow a gaussian distribution of standard deviation $\sigma_j$. We indicate in sec. D.2 how the program WPtools proceeds in the absence of input data as to the $\sigma_j$.

Because of the uncertainties in the measurements $y_j$ we cannot expect to find the ideal values of the coefficients $a_i$, but only a set of best estimates we will call $\hat{a}_i$. However, we will also obtain estimates of the uncertainties in these best-fit parameters which we will label as $\sigma_{\hat{a}_i}$.

The best-fit polynomial is then

$$\hat{y}(x) = \sum_{i=0}^{n} \hat{a}_i x^i. \tag{D.2}$$

The method to find the $\hat{a}_i$ is called least-squares fitting as well as polynomial regression because we minimize the square of the deviations. We introduce the famous chi square:

$$\chi^2 = \sum_{j=1}^{m} \frac{[y_j - \hat{y}(x_j)]^2}{\sigma_j^2} = \sum_{j=1}^{m} \frac{\left(y_j - \sum_{i=0}^{n} \hat{a}_i x_j^i\right)^2}{\sigma_j^2}. \tag{D.3}$$

Fact: $\exp(-\chi^2/2)$ is the (un-normalized) probability distribution for observing a set of variables $\{y_j(x_j)\}$ supposing the true relation of $y$ to $x$ is given by eq. (D.2).

A great insight is that $\exp(-\chi^2/2)$ can be thought of another way. It is also the (un-normalized) probability distribution that the polynomial coefficients have values $a_i$ when their best-fit values are $\hat{a}_i$ with uncertainties due to the measurements $\{y_j\}$. Expressing this in symbols,

$$\exp(-\chi^2/2) = \text{const} \times \exp\left(-\sum_{k=0}^{n}\sum_{l=0}^{n} \frac{(a_k - \hat{a}_k)(a_l - \hat{a}_l)}{2\sigma_{kl}^2}\right), \tag{D.4}$$

or equivalently

$$\chi^2/2 = \text{const} + \sum_{k=0}^{n}\sum_{l=0}^{n}\frac{(a_k-\hat{a}_k)(a_l-\hat{a}_l)}{2\sigma_{kl}^2}. \tag{D.5}$$

The uncertainty on $\hat{a}_k$ is $\sigma_{kk}$ in this notation. In eqs. (D.4) and (D.5) we have introduced the important concept that the uncertainties in the coefficients $\hat{a}_k$ are correlated. That is, the quantity $\sigma_{kl}^2$ is a measure of the probability that the values of $a_k$ and $a_l$ both have positive fluctuations at the same time. In fact, $\sigma_{kl}^2$ can be negative indicating that when $a_k$ has a positive fluctuation then $a_l$ has a correlated negative one.

One way to see the merit of minimizing the $\chi^2$ is as follows. According to eq. (D.5) the derivative of $\chi^2$ with respect to $a_k$ is

$$\frac{\partial\chi^2/2}{\partial a_k} = \sum_{l=0}^{n}\frac{a_l-\hat{a}_l}{\sigma_{kl}^2}, \tag{D.6}$$

so that all first derivatives of $\chi^2$ vanish when all $a_l = \hat{a}_l$. That is, $\chi^2$ is a minimum when the coefficients take on their best-fit values $\hat{a}_i$. A further benefit is obtained from the second derivatives:

$$\frac{\partial^2\chi^2/2}{\partial a_k\partial a_l} = \frac{1}{\sigma_{kl}^2}. \tag{D.7}$$

In practice we evaluate the $\chi^2$ according to eq. (D.3) based on the measured data. Taking derivatives we find

$$\frac{\partial\chi^2/2}{\partial\hat{a}_k} = \sum_{j=1}^{m}\frac{\left(y_j-\sum_{i=0}^{n}\hat{a}_i x_j^i\right)\left(-x_j^k\right)}{\sigma_j^2} = \sum_{i=0}^{n}\sum_{j=1}^{m}\frac{\hat{a}_i x_j^i x_j^k}{\sigma_j^2} - \sum_{j=1}^{m}\frac{y_j x_j^k}{\sigma_j^2}, \tag{D.8}$$

and

$$\frac{\partial^2\chi^2/2}{\partial\hat{a}_k\partial\hat{a}_l} = \sum_{j=1}^{m}\frac{x_j^k x_j^l}{\sigma_j^2} \equiv M_{kl}. \tag{D.9}$$

To find the minimum $\chi^2$ we set all derivatives (D.8) to zero, leading to

$$\sum_{i=0}^{n}\sum_{j=1}^{m}\frac{x_j^i x_j^k}{\sigma_j^2}\hat{a}_i = \sum_{j=1}^{m}\frac{y_j x_j^i}{\sigma_j^2} \equiv V_k. \tag{D.10}$$

Using the matrix $M_{kl}$ introduced in eq. (D.9) this can be written as

$$\sum_{i=0}^{n} M_{ik}\hat{a}_i = V_k. \tag{D.11}$$

We then calculate the inverse matrix $M^{-1}$ and apply it to find the desired coefficients:

$$\hat{a}_k = \sum_{l=0}^{n} M_{kl}^{-1}V_l. \tag{D.12}$$

Comparing eqs. (D.7) and (D.9) we have

$$\frac{1}{\sigma_{kl}^2} = M_{kl}. \tag{D.13}$$

The uncertainty in best-fit coefficient $\hat{a}_i$ is then reported as

$$\sigma_{\hat{a}_i} = \sigma_{ii} = \frac{1}{\sqrt{M_{ii}}}. \tag{D.14}$$

## D.2 Procedure When the $\sigma_j$ Are Not Known

This method can still be used even if the uncertainties $\sigma_j$ on the measurements $y_j$ are not known. When the functional form (D.1) correctly describes the data we claim that on average the minimum $\chi^2$ has value $m - n - 1$.[1] To take advantage of this remarkable result we suppose that all uncertainties $\sigma_j$ have a common value, $\sigma$. Then

$$\chi^2 = \sum_{j=1}^{m} \frac{[y_j - \hat{y}(x_j)]^2}{\sigma^2} \approx m - n - 1, \tag{D.15}$$

so that

$$\sigma_j = \sigma = \sqrt{\frac{\sum_{j=1}^{m}[y_j - \sum_{i=0}^{n} \hat{a}_i x_j^i]^2}{m - n - 1}}\,. \tag{D.16}$$

*In practice it appears that the error estimates from this procedure are more realistic if a fit is made using a polynomial with one order higher than needed for a 'good' fit to the data.*

Using eq. (D.16) as the estimate of the uncertainty $\sigma$ on each of the measurements $y_j$, the matrix $M_{kl}$ of eq. (D.9) becomes

$$M_{kl} = \frac{m - n - 1}{\sum_{j'=1}^{m}[y_{j'} - \sum_{i'=0}^{n} \hat{a}_{i'} x_{j'}^{i'}]^2} \sum_{j=1}^{m} x_j^k x_j^l. \tag{D.17}$$

The estimate (D.14) of the uncertainty on the fit coefficient $\hat{a}_i$ is now given by

$$\sigma_{\hat{a}_i} = \frac{1}{\sqrt{M_{ii}}} = \sqrt{\frac{\sum_{j'=1}^{m}[y_{j'} - \sum_{i'=0}^{n} \hat{a}_{i'} x_{j'}^{i'}]^2}{(m - n - 1)\sum_{j=1}^{m} x_j^{2i}}} \tag{D.18}$$

When WPtools performs a polynomial regression it generates a plot of the data points and the best-fit curve, along with numerical values of various parameters associated with the fit. Figure D.1 gives an example of a fit to a set of 8 data points of the form $y = x^2$. The fit is to the form $\mathsf{y} = \mathsf{a_0} + \mathsf{a_1 x} + \mathsf{a_2 x^2}$. The fit coefficients are $\mathsf{a_0} = -0.4107$, $\mathsf{a_1} = -0.3274$ and $\mathsf{a_2} = 1.1964$. The uncertainties (standard errors) on the fit coefficients are reported as $\mathsf{SE(a_0)} = 4.0070$, $\mathsf{SE(a_1)} = 2.0429$ and $\mathsf{SE(a_2)} = 0.2216$, as calculated according to eq. (D.18). *Note that the uncertainties on coefficients* $\mathsf{a_1}$ *and* $\mathsf{a_1}$ *are larger than the coefficients themselves, which tells us that these coefficients are indistinguishable from zero.*

Also indicated on the plot are the values $\mathsf{R^2} = 0.9915$ and $\sigma = 2.8721$. The latter is the uncertainty in the data points $\{y_j\}$, calculated according to eq. (D.16) with $m = 8$ and $n = 2$. The quantity $\mathsf{R^2}$ is defined by

$$R^2 = \frac{\sum_{j=1}^{m}[\hat{y}(x_j) - \overline{y}]^2}{\sum_{j=1}^{m}[y(x_j) - \overline{y}]^2}\,, \tag{D.19}$$

where the average $\overline{y} = \sum_{j=1}^{m} y(x_j)/m$. This is a measure of the "goodness of fit". If the fit is perfect then $\hat{y}_j = y_j$ for all $j$ and $R^2 = 1$. It is not obvious, but $R^2 \leq 1$ always. The extreme case of $R^2 = 0$ occurs when the fit has the trivial form $\hat{y}(x) = \overline{y}$ for all $x$, which in general is a bad fit. The qualitative conclusion is that if $R^2$ is not close to 1, the fit results are to be regarded with suspicion.

[1] The whole fitting procedure does not make sense unless there are more data points ($m$) than parameters ($n + 1$) being fitted.

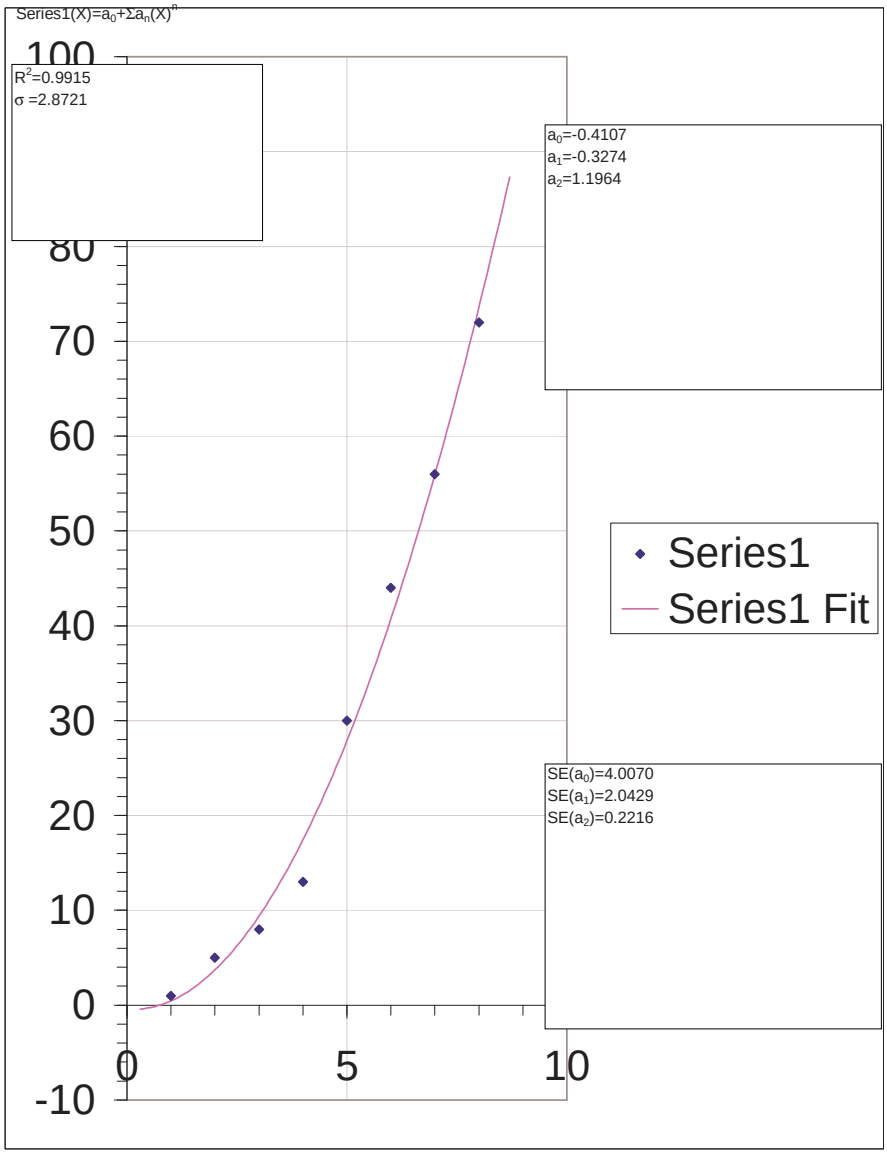


Figure D.1: Sample plot from WPtools Polynomial Fitting.