PHYS 342L

NOTES ON ANALYZING DATA

Spring Semester 2002

Department of Physics
Purdue University

A major aspect of experimental physics (and science in general) is measurement of some quantities and analysis of experimentally obtained data. While there are a lot of books devoted to this problem, in the next paragraphs we will summarize some of the important ideas that will be needed to successfully analyze data acquired in PHYS 342L. Students are advised to consult with [1] for more detailed discussion on the topic.

1. The importance of estimating errors.

Suppose you are asked to measure the length of a piece of notebook paper. You grab a ruler and proceed with a measurement. The ruler shows 276 mm. Does it mean the length is 276.0000 mm? Most probably not. Why? Because the distance between the neighboring marks on your ruler is 1 mm, by saying 276 mm you cannot exclude, for example, length 276.2 or 259.9 mm. Thus you assume certain precision (or error) in your measurement, in this case it is probably ~0.5 mm, as the distance between the closest marks is 1 mm. The result of the meausurement is not just the length of the paper but also the error of this measurement: (276.0±0.5) mm. In a scientific experiment, both parts of measurement are important. Suppose you measure the length of the next sheet of paper to be (275.5±0.5) mm. Within the error of your measurement these two sheets of paper have the same length.

2. Precision (or Accuracy) of a Measurement.

Distinguish between absolute uncertainty and relative uncertainty:

absolute uncertainty	relative uncertainty
27.6 ± 0.1	$\pm \frac{0.1}{27.6} = \pm 0.003623188$

All these numbers don't mean much when calculating the relative uncertainty, so round off to ± 0.004, or, expressed as a percent, ± 0.4%.

3. Combining Uncertainties.

Suppose that you measure two quantities A and B. Suppose you measure A to an accuracy of $\pm\delta A$ and B to an accuracy of $\pm\delta B$.

How do you algebraically combine these uncertainties?

a) When adding:

$$(A \pm \delta A) + (B \pm \delta B) = ?$$

there are four possibilities:

$$(A + \delta A) + (B + \delta B) = (A + B) + (\delta A + \delta B)$$
$$(A + \delta A) + (B - \delta B) = (A + B) + (\delta A - \delta B)$$
$$(A - \delta A) + (B + \delta B) = (A + B) - (\delta A - \delta B)$$
$$(A - \delta A) + (B - \delta B) = (A + B) - (\delta A + \delta B)$$

clearly, the worst case will be $(A+B)\pm(\delta A+\delta B)$ (1)

b) When subtracting:

$$(A\pm\delta A) - (B\pm\delta B) = ?$$

Again consider four cases. From above, it should be obvious that the worst case will be given by

$$(A\text{-}B)\pm(\delta A+\delta B) \tag{2}$$

c) When multiplying

$$(A\pm\delta A)\times(B\pm\delta B) = AB + A(\pm\delta B) + B(\pm\delta A) + \underbrace{(\pm\delta A)(\pm\delta B)}_{small,neglect}$$

$$\approx AB \pm (A\delta B + B\delta A) \tag{3}$$

$$\approx AB\left[1\pm\left(\frac{\delta B}{B}+\frac{\delta A}{A}\right)\right]$$

d) When dividing

$$\frac{A\pm\delta A}{B\pm\delta B} = ?$$

After some algebra, you find that

$$\frac{A\pm\delta A}{B\pm\delta B} \approx \frac{A}{B}\left[1\pm\left(\frac{\delta A}{A}+\frac{\delta B}{B}\right)\right] \tag{4}$$

Remember:

- **relative** uncertainties add when multiplying or dividing.
- **absolute** uncertainties add when adding or subtracting

4. Systematic and random errors.

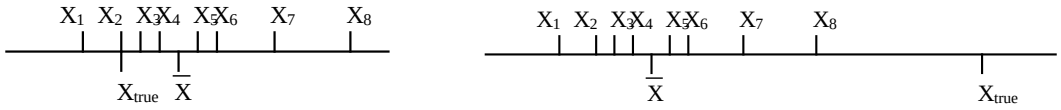

Figure 1: Spread in the measurement of some quantity x in the absence of systematic error (left) and in its presence (right).

If error in your measurements is random, then the average value should be close to the actual value. In the case of systematic error, that is not true. This situation may occur when, for example, using a clock which is running slow to measure some time period. Random errors are inevitable, while systematic errors can be taken into account or eliminated.

5. Average value and standard deviation.

In order to decrease the influence of random error multiple measurements x_i are taken and averaged:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

How close this average value $\bar{x}$ is to the actual value X? If we have a set of

measurements we can find an average error for a single measurement. The commonly accepted value to characterize error is called *standard deviation* σ, or *root mean square* (rms):

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - X)^2 \qquad (6)$$

Since the actual value X is usually unkown, we must use $\bar{x}$ instead. It can be shown [1] that in this case:

$$\sigma^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (7)$$

The value σ characterizes error in a *single* measurement of value X. If we take several measurements of the same value x and average them, the resulting value $\bar{x}$ must in average be closer to actual value x as a single measurement. It can be shown [1] that standard deviation σ_n for the average value of n measurements is:

$$\sigma_n = \frac{\sigma}{\sqrt{n}} \qquad (8)$$

6. Distribution of measurements

A series of measurements may be represented as a histogram (Fig. 2).

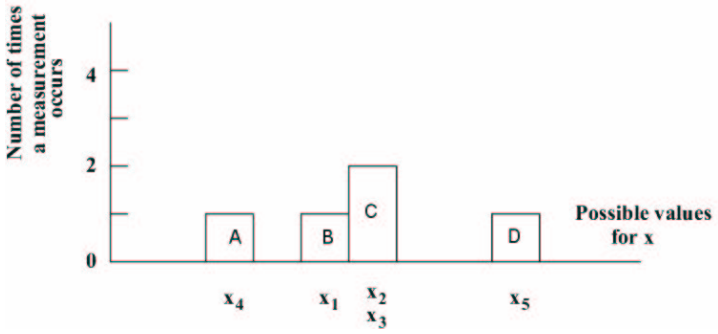

Figure 2. A simple histogram after taking just five data points (n=5). There was only one data point falling into range of x marked as A, B and D, and two measutrements where in region C.

It is difficult to see any trends after taking just a few data points. Make more measurements and use smaller bins and you'll eventually get a histogram that might look like this.

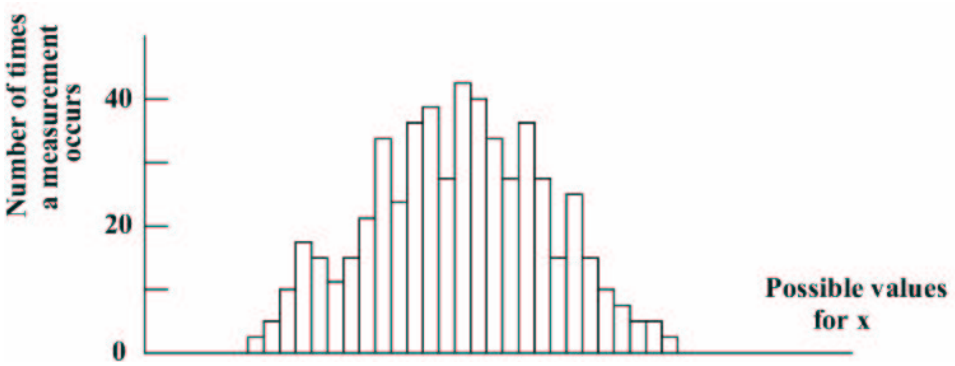

Figure 3. An histogram after taking hundreds of measurements.

In a limit of large *n* the distribution is given by continuous *distributin function f(x)*, so that *f(x)dx* is the probability that a single measurement taken at random will lie in the interval *x* to *x+dx*. The average value can be then found as:

$$\langle x \rangle = \int_{-\infty}^{\infty} x \cdot f(x) dx \tag{9}$$

And standard deviation:

$$\sigma^2 = \langle x^2 \rangle = \int_{-\infty}^{\infty} (x - X)^2 f(x) dx \approx \int_{-\infty}^{\infty} \left(x - \langle x \rangle\right)^2 f(x) dx \tag{10}$$

In many cases error distribution function is well described by Gaussian (also called *normal* distribution (Fig. 4):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-X)^2}{2\sigma^2}} \tag{11}$$

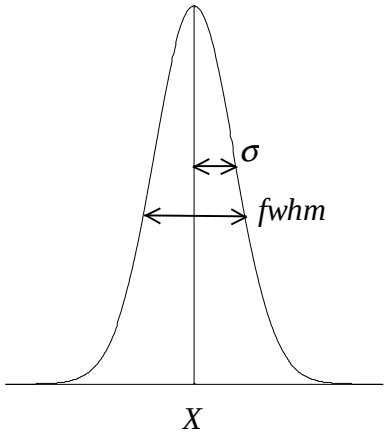

Figure 4. Gaussian distribution function.

The standard deviation σ for Gaussian distribution can be also expressed as:

$$\sigma = \frac{fwhm}{2\sqrt{2\ln(2)}} \approx 0.425 \times fwhm \qquad (12)$$

where *fwhm* if full width at half maximum, which can be estimated graphically.

Suppose now that we performeed a single measurement which resulted in value *x* and we also know the standard deviation of this measurement σ. Since the Gaussian distribution is a continuous function which becomes zero only in infinity, the measured value *x* may lay anywhere from -∞ to +∞. What is the probability that the actual value X which we are trying to measure is within distance σ from this measured value? Since *f(x)dx* is the probability of measuring value between *x* and *x+dx*, the probability of measuring *x* between X-σ to X+σ is given by integral $\int_{X-\sigma}^{X+\sigma} f(x)dx = 0.68$.

Thus, the probablity of your measurement *x* being within

X±σ - 68%
X±2σ - 95%
X±3σ - 99.7%
X±4σ - 99.994%

6. Combining σ's.

Let us now get back to the case when we add two values *A* and *B*, where standard deviations are σ_A and σ_B, correspondingly. What would be the standard deviation of the sum σ_{A+B}? As we already know, the errors should be added for the worst case scenario (Eq. 1). However, if errors in *A* and *B* are random and mutually uncorrelated, they tend to cancel to some extent as there is 50% probability that they have different sign in one measurement set. It can be shown, that the standard deviation of the sum (or difference) is:

$$\sigma_{A\pm B} = \sqrt{\sigma_A^2 + \sigma_B^2} \qquad (13)$$

Notice that $\sigma_{A\pm B} \le \sigma_A + \sigma_B$. If $\sigma_A = \sigma_B = \sigma$, then $\sigma_{A\pm B} = \sigma\sqrt{2}$.

Similarily, the *relative* standard deviation σ_C/C for product (or ratio) of A and B is:

$$\frac{\sigma_C}{C} = \sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2} \qquad (14)$$

where *C=AB* or *C=A/B*.

Note, that this is true only if errors are *uncorrelated* and *not systematic*.

7. A simple example

Suppose you need to evaluate the charge to mass ratio of an electron. This is to be done using the following equations

$$\frac{q}{m} = \frac{2V}{B^2R^2}, \tag{15}$$

where $B=kI$

How would you analyze the error in $\frac{q}{m}$?

Suppose you measure I, V and k as follows:

$I=(1.4\pm0.1)\ A$

$V=(140\pm2)\ V$

$k=coil\ constant=(7.5\pm0.5)\times10^{-4} T/A$

You also measure *R*, the radius of the electron's orbit, by measuring its diameter *D*. Since the smallest marking on the ruler is 1 mm and you have to determine the positions of both sides of the electron orbit, the precision of such a measurement is not better than 2mm = 0.2 cm. To reduce random error you may want to take several measurements. Suppose you make three measurements of *D*:

$D_1=6.0\pm0.2$ cm	$R_1=D_1/2=3.0\pm0.1$ cm
$D_2=5.8\pm0.2$ cm	$R_2=2.9\pm0.1$ cm
$D_3=5.7\pm0.2$ cm	$R_3=2.85\pm0.1$ cm

Using Eqs. 5 and 7, the average value of *R* and the standard deviation σ for one measurment will be

$$\overline{R} = \frac{3.0+2.9+2.85}{3} = 2.92(cm)$$

$$\sigma = \frac{\sqrt{0.08^2+0.02^2+0.07^2}}{2} = 0.054$$

According to Eq. 8, for the average of 3 measurements the standard deviation σ_3 will be

$$\sigma_3 = \frac{0.054}{\sqrt{3}} \approx 0.03$$

However, the ruler we use has precision ~0.1 cm only. Using Eq. 13 we can account for both errors:

$$\Delta R = \sqrt{\sigma_3^2 + 0.1^2} \approx 0.11cm$$

and we have

$R=$ (2.92±0.11) cm

Now we calculate:

$$\frac{q}{m} = \frac{2V}{B^2R^2} = \frac{2V}{k^2I^2R^2}$$

$$\frac{q}{m} = \frac{2(140\pm2)}{\left[(7.5\pm0.5)\times10^{-4}\right]^2\left[1.4\pm0.1\right]^2\left[0.0292\pm0.0011\right]^2} \tag{16}$$

Omitting errors we get the value of $\frac{q}{m} = 2.98 \times 10^{-11} C/kg$

Using Eq. 14 we can write:

$$\frac{\sigma_{q/n}}{q/n} = \sqrt{\left(\frac{\sigma_V}{V}\right)^2 + \left(\frac{2\sigma_k}{k}\right)^2 + \left(\frac{2\sigma_I}{I}\right)^2 + \left(\frac{2\sigma_R}{R}\right)^2} \qquad (17)$$

Note the factors "2" in the equation above. These stem from the fact that corresponding values are squared in Eq. 16, i.e. we have products $k \times k$, $I \times I$ and $R \times R$. Since $k \times k$ is a product of two *correlated* values, we must use Eq. 3 – the relative errors simply add up.

$$\sigma_{q/n} = 2.98 \times 10^{-11} \sqrt{\left(\frac{2}{140}\right)^2 + \left(\frac{2 \times 7.5}{0.5}\right)^2 + \left(\frac{2 \times 0.1}{1.4}\right)^2 + \left(\frac{2 \times 0.0011}{0.0292}\right)^2}$$

$$\sigma_{q/n} = 0.63 \times 10^{-11} C/kg$$

And the final result for the ratio can be written as:

$$\frac{q}{m} = (2.98 \pm 0.63) \times 10^{-11} C/kg$$

Here we used Eq. 14 since values *V*, *k*, *I* and *R* are uncorrelated. If we assume, just for example, a correlated case, we must use Eq. 3 and 4 – i.e. add relative errors:

$$\frac{\sigma_{q/n}}{q/n} = \left(\frac{\sigma_V}{V}\right) + \left(\frac{2\sigma_k}{k}\right) + \left(\frac{2\sigma_I}{I}\right) + \left(\frac{2\sigma_R}{R}\right)$$

and $\sigma_{q/n} = 1.10 \times 10^{-11} C/kg$

The accepted value is 1.76×10^{11} C/kg. It's easy to make a simple plot including error bars to graphically illustrate this result.

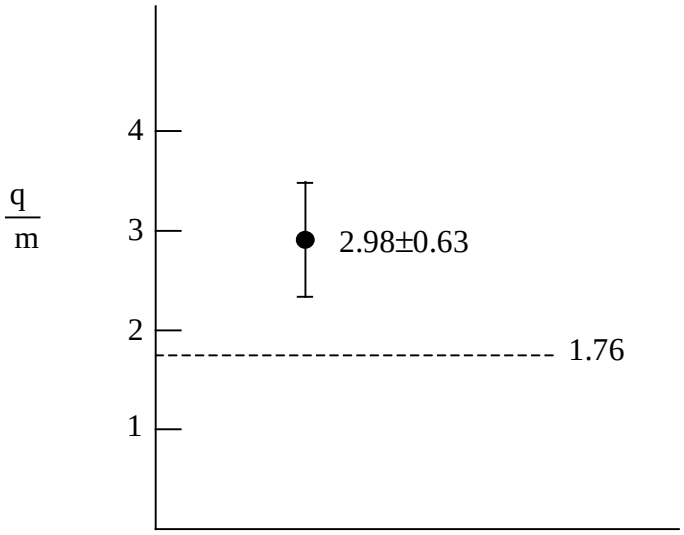

Figure 6: A plot showing the measured value with error bar. The dotted horizontal line represents the accepted value. The q/m axis has units of $\times 10^{11}$ C/kg.

The measured value of q/m is ~2σ higher than the accepted value. The probability for that to happen is only ~5%, which strongly suggests the presence of some systematic error in our measurement.

8. Least Squares Fit.

Suppose you measure some data points y as a function of a variable called x. After the measurements, you will have a set of data points

x_1, y_1
x_2, y_2
……
x_N, y_N

Sometimes you might know that the data should fit a straight line (e.g., from theoretical considerations). The equation of a straight line is

$$y=mx+b$$

where the slope m might equal 'a certain quantity of interest' and the intercept b might equal 'some other quantity of interest'.

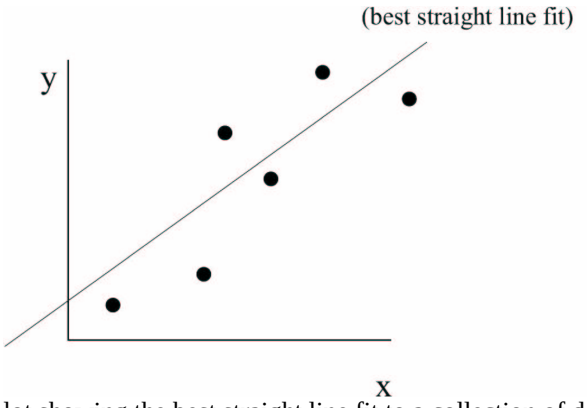

Figure 7: A plot showing the best straight line fit to a collection of data points.

In short, if you could determine *m* and *b*, these values may contain estimates for useful quantities. One way to determine *m* and *b* is to plot the data and use a ruler to draw a straight line through the points. Then, by calculating *m* and *b* from the straight line drawn, you have produced some weighted average estimate of *m* and *b* from all your data.

A simple example? Suppose you are asked to determine π experimentally and suppose you already know that for circles

$$\text{circumference} = \pi\ (\text{diameter})$$

One way to proceed might be to make a variety of circles of different diameters and then measure the circumference of each one. You might plot the data as follows:

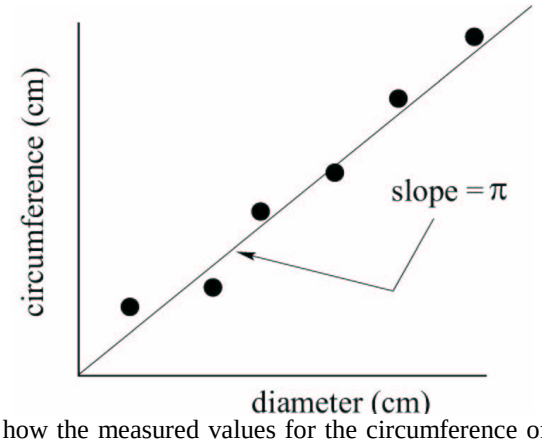

Figure 8: A plot of how the measured values for the circumference of different circles might vary as a function of the measured diameter. Note that in this example, the intercept of the best straight line through the data MUST pass through the co-ordinate origin.

Clearly, the slope of a straight line through the data contains useful information since π=slope.

Q: How can you determine the 'best value' for the slope and intercept without prejudice or personal judgment?

A: Use the principle of **least squares**.

Assume you draw N circles and make measurements of each circumference and diameter. Let the independent variable (the diameter) be represented by the symbol *d*. Let the dependent variable (the circumference) be represented by the symbol C. Also assume the *d* values are accurate. After the measurement process, you'll have a set of numbers (*d*,*C*):

$$d_1, C_1$$
$$d_2, C_2$$
$$\ldots\ldots$$
$$d_N, C_N$$

It is conventional to map these numbers into the parameters (*x*,*y*) as follows

$$x_1=d_1, y_1= C_1$$
$$x_2=d_2, y_2= C_2$$
$$\ldots\ldots$$
$$x_N=d_N, y_N= C_N$$

Let the difference between the 'best line' through the data and each individual data point be represented by (δy_i). One unambiguous way to specify the 'best line' through all the data can be defined by the condition that the sum of all the $(\delta y_i)^2$ have a minimum value.

How are the individual δy_i defined? Graphically, they are indicated in the plot below. Note that at this point of the analysis, the straight line drawn need not be the best straight line through the data.

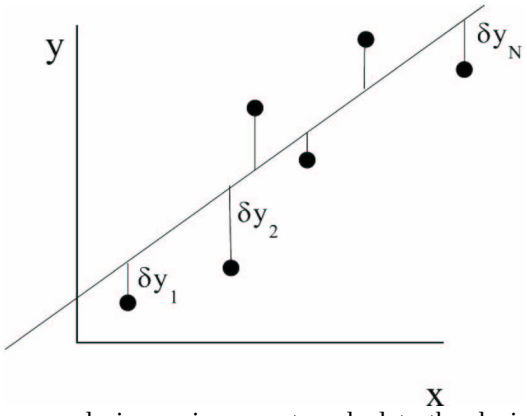

Figure 9. A least squares analysis requires you to calculate the deviation of each data point from the 'best' straight line.

Mathematically, you can calculate the δy_i as follows. Suppose you define a quantity Y_i such that $Y_i=mx_i+b$ where the symbols m and b are somehow chosen to represent the 'best' straight line through the data, whatever that means. Calculate

$$\delta y_i=y_i-Y_i=C_i-(md_i+b)$$

Least Squares Fitting requires that (where the switch in notation from (C,d) to (y,x) has been made)

$$\sum_{i=1}^{N}(\delta y_i)^2=\sum_{i=1}^{N}[y_i-(mx_i+b)]^2=\text{minimum}$$

Write $\Sigma\,[y_i-(mx_i+b)]^2=M$. The conditions for M to be a minimum are

$$\frac{\partial M}{\partial m}=0,\qquad \frac{\partial M}{\partial b}=0$$

Performing the derivatives and setting them equal to zero gives, after some algebra, two unique equations for the 'best' m and b.

$$m=\frac{N\sum_{i=1}^{N}(x_iy_i)-\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}$$

$$m=\frac{\sum_{i=1}^{N}x_i^2\sum_{i=1}^{N}y_i-\sum_{i=1}^{N}x_i(x_iy_i)}{N\sum_{i=1}^{N}x_i^2-\left(\sum_{i=1}^{N}x_i\right)^2}$$

Once we have m and b, then also calculate the intermediate quantity σ_y:

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{N}\left[y_i - (mx_i + b)\right]^2}{N-2}}$$

It can be shown that the uncertainty in the slope *m* and intercept *b* is given by

$$\sigma_m = \sigma_y \sqrt{\frac{N}{N\sum x_i^2 - \left(\sum x_i\right)^2}}$$

$$\sigma_m = \sigma_y \sqrt{\frac{\sum x_i^2}{N\sum x_i^2 - \left(\sum x_i\right)^2}}$$

You can conclude that the best values for m and b are

$$m \pm \sigma_m \quad b \pm \sigma_b$$

This means that there is a 68% chance of the real *m* lying between m-σ_m and m+σ_m. Likewise, there is a 68% chance of the real *b* lying between b-σ_b and b+σ_b.

Since the least squares fitting formulae involve sums over various combinations of measured data, the least squares fitting procedure is especially easy to implement in spread sheets like Excel. In fact, most spread sheet programs have pre-programmed least square fit routines available as analysis tools.

Reference

1. G. L. Squires. *Practical Physics*, 4th Edition, Cambridge University Press, Cambridge (2001)