PHYS 231/232/233 Data analysis

Contents

1 Types of experimental error

- **Blunders** are mistakes made by the experimentalist. This type of error should be eliminated.
- **Random errors** arise due to inherent limitations in the manner in which a measurement is made, and do not hamper one from attaining a good estimate of the true mean. By repeating a measurement one can get a good estimate of the random error associated with that measurement.
- **Systematic errors** are the result of an inherent and unknown bias in the mode of measurement, and may result in an incorrect estimate of the true mean. By **calibrating** one's apparatus against a standard physical quantity one can reduce systematic error in subsequent measurements.

As an experimentalist one tries to minimize and realistically estimate random error, and reduce systematic error such that it may be neglected in comparison.
Example *A length was measured several times with a ruler that measures to 0.1 cm. (The next digit is estimated, so the length measurements are given to 0.01 cm.) One of the measurements is short by about a centimeter. Since the precision of the ruler is much less than a centimeter, it is very likely that a* **blunder** *was made. This one measurement should be redone.*
After redoing this one measurement a value is found that is roughly in agreement with the other measurements. There seems to be a spread of about 0.05 cm in the measurements. This is due to **random error**, *and is to be expected.*
It is noticed after the report has been written up that the ruler was worn down by use. Therefore, all of the length measurements are too long. This is a **systematic error** *which could have been eliminated by calibrating the ruler.*

2 Sample mean and standard deviation

When measuring some physical quantity take several (4-5) measurements to better estimate the quantity of interest and the error on the measurement. The sample **mean**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is the best estimate of the true value of the physical quantity. The sample **variance** is

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

(This is roughly the average squared distance to the mean.) The sample **standard deviation** σ_x (the square root of the variance, often called the **error** or absolute error), will be our estimate of the error on the measurement. When quoting the result of such a measurement, write the result in the form $\overline{x} \pm \sigma_x$. When comparing errors it is important to compare **relative errors**, $\sigma_x/\overline{x}$ (also called fractional errors or percent errors when expressed as a percentage). Relative error is a dimensionless measure of how much error is associated with a given measurement.

Example *Suppose a certain physical quantity x is measured four times, with the result*

$$1.04, 0.93, 0.98, 0.91.$$

The mean, our best estimate of the value of x, is

$$\overline{x} = \frac{1.04 + 0.93 + 0.98 + 0.91}{4} = 0.965.$$

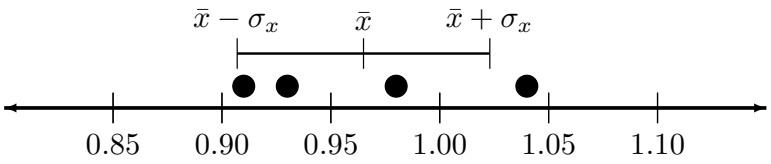
(Typically we keep extra digits until having calculated σ_x. In fact, σ_x will tell us how many digits of $\overline{x}$ are significant. Here, $\overline{x}$ is exactly 0.965 so we do not bother to write the other digits.) The variance is

$$\sigma_x^2 = \frac{1}{3}((1.04 - 0.965)^2 + (0.93 - 0.965)^2 + (0.98 - 0.965)^2 + (0.91 - 0.965)^2) = 0.00336667.$$

Therefore, our best estimate of the error on this measurement is $\sigma_x = \sqrt{0.00336667} = 0.05\underline{8}0230 = 0.058$. (By convention we will keep two significant figures for the error. The precision of the mean is then made to agree with that of the error.) The result for this measurement would be quoted as

$$0.965 \pm 0.058.$$

The relative error is $0.05\underline{8}0230/0.965 = 6.0\%$. The figure below makes it clear that the mean, $\overline{x}$, provides a best estimate for x while the error, σ_x, measures the "spread" of the data about the mean.



3 Error propagation: Numerical method

Suppose $f = f(x_1, \ldots, x_m)$ is a function[1] of m physical quantities x_i with mean $\overline{x}_i$ and error σ_i. We wish to estimate the error on f due to the error on each of the physical quantities on which it depends. Let

$$\begin{aligned} \overline{f} &= f(\overline{x}_1, \ldots, \overline{x}_m) \\ f_i &= f(\overline{x}_1, \ldots, \overline{x}_i + \sigma_i, \ldots, \overline{x}_m). \end{aligned}$$

The best estimate of f will be given by $\overline{f}$. The quantity $|f_i - \overline{f}|$ estimates the error on f due to the error on x_i.
The variance of f can be estimated by

$$\sigma_f^2 = \sum_{i=1}^{m} \left(f_i - \overline{f}\right)^2,$$

and so the error on f will be given by $\sigma_f = \sqrt{\sigma_f^2}$. The sequence of calculations that leads to the error on f is called **error propagation**.[2]
Only important sources of error need be considered when propagating error. For example, if x_j has a relative error that is ten times smaller than the largest relative error (on x_k, say), then the error on x_j can be neglected in finding the error on f. In this case we would treat x_j as being exactly known. Neglecting unimportant contributions to the error on f can greatly simplify calculations.
Suppose x_k has the largest relative error of all of the $x_1, \ldots, x_n$. Typically the relative error for $f(x_1, \ldots, x_n)$ is roughly the same as the relative error on x_k, i.e., we should find that $\sigma_f/\overline{f} \approx \sigma_k/\overline{x}_k$ (these values usually agree to within a factor of two or three). This provides an important gut check on the calculation of $\overline{f} \pm \sigma_f$.
Example *If $m = 1$, let $x = x_1$ and $\sigma_x = \sigma_1$. Then*

$$\overline{f} = f(\overline{x}) \quad \textit{and} \quad \sigma_f = |f(\overline{x} + \sigma_x) - \overline{f}|.$$

Example *Suppose an angle is measured to be $(35.2 \pm 1.5)^\circ$, and that we must approximate the error on the sine of this angle. Then $f(\theta) = \sin\theta$ and so*

$$\overline{f} = f(\overline{\theta}) = \sin 35.2^\circ = 0.576432$$

and

$$\sigma_f = |f(\overline{\theta} + \sigma_\theta) - \overline{f}| = |\sin 36.7^\circ - \sin 35.2^\circ| = 0.02\underline{1}1928.$$

Thus, our best estimate of $\sin\theta$ is

$$0.576 \pm 0.021.$$

Notice that the relative error for θ is 4.3% and for $\sin\theta$ is 3.7%, that is, the relative error in the angular measurement is roughly the relative error in the sine of that angle.
Example *Verify the results in the second column! These results can be very useful in propagating errors.*

$$\begin{array}{lll} f = x \pm y & \sigma_f^2 = \sigma_x^2 + \sigma_y^2 & \\ f = ax \pm by & \sigma_f^2 = a^2\sigma_x^2 + b^2\sigma_y^2 & \\ f = xy & \sigma_f^2 = \overline{y}^2\sigma_x^2 + \overline{x}^2\sigma_y^2 \quad \textit{or} & \left(\dfrac{\sigma_f}{\overline{f}}\right)^2 = \left(\dfrac{\sigma_x}{\overline{x}}\right)^2 + \left(\dfrac{\sigma_y}{\overline{y}}\right)^2 \end{array}$$

[1] Notation for functions of more than one variable: If $f(x, y) = xy^2$, then $f(2, 3) = 2 \cdot 3^2$ and $f(x + 2, y^2) = (x + 2)y^4$.
[2] The formula for σ_f^2 is proved in the appendix.

Example *Suppose we wish to measure the volume V of a cylinder. We make four independent measurements of the height h and diameter d. The results are tabulated below, along with the means, errors, and relative errors.*

Measurement	$d(cm)$	$h(cm)$
1	1.73	1.155
2	1.80	1.156
3	1.71	1.156
4	1.79	1.154
$\overline{x}$	1.75750	1.15525
σ_x	0.0442531	0.000957427
$\sigma_x/\overline{x}$	0.0251796	0.000828762

Since the relative error on h is thirty times smaller than that of d we can neglect the error on h in calculating the error on V. First we will keep the error on h and then show that neglecting this error is justified. We have $V = V(d,h) = \pi d^2 h/4$, and so our best estimate of the volume is

$$\overline{V} = \pi \overline{d}^2 \overline{h}/4 = 2.80257 \text{ cm}^3.$$

Note that $\overline{d} + \sigma_d = 1.80175$ cm and $\overline{h} + \sigma_h = 1.15621$ cm. Then,

$$\begin{aligned}
\sigma_V^2 &= [V(\overline{d}+\sigma_d, \overline{h}) - \overline{V}]^2 + [V(\overline{d}, \overline{h}+\sigma_h) - \overline{V}]^2 \\
&= \left[\pi(\overline{d}+\sigma_d)^2\overline{h}/4 - \overline{V}\right]^2 + \left[\pi\overline{d}^2(\overline{h}+\sigma_h)/4 - \overline{V}\right]^2 \\
&= \left(\pi \cdot 1.80175^2 \cdot 1.15525/4 - 2.80257\right)^2 + \left(\pi \cdot 1.75750^2 \cdot 1.15621/4 - 2.80257\right)^2 \\
&= 0.0204292 \text{ cm}^6. \quad \textit{(Verify this!)}
\end{aligned}$$

Therefore

$$\sigma_V = 0.142931 \text{ cm}^3.$$

Neglecting the error on h we find

$$\begin{aligned}
\sigma_V &= \left|\pi(\overline{d}+\sigma_d)^2\overline{h}/4 - \overline{V}\right| \\
&= |\pi \cdot 1.80175^2 \cdot 1.15525/4 - 2.80257| \\
&= 0.142912 \text{ cm}^3.
\end{aligned}$$

In either case we find the volume to be

$$(2.80 \pm 0.14) \text{ cm}^3.$$

Notice that the relative error on V (about 5%) is comparable to the relative error on the physical quantity with the largest relative error (in this case, d). This provides an important check on this calculation.

4 The empirical rule

We expect about 68%, 95%, and 99.7% of all values to lie within one, two, and three standard deviations of the mean, respectively.[3]

Example

- *If we measure the acceleration due to gravity to be (9.90 ± 0.10) m/s^2 we cannot say that we have measured a value different from 9.80 m/s^2 with any degree of certainty.*

[3]This only holds if the distribution underlying our measurement is a normal distribution, an assumption that we will often make.

- *A measurement of* (9.900 ± 0.010) m/s^2 *is significantly different from the known value of* g*, being about ten standard deviations away from it. If we measured this value on the surface of the earth we have likely underestimated the random error on one or more of our measurements. It is also possible that one or more of our measurements has a large systematic error which should be accounted for.*

5 Least squares fit to a straight line

5.1 Fit to (x_i, y_i, σ_{yi})

Often a complicated relationship between physical quantities can be expressed as a linear relationship, perhaps between two different but related physical quantities. Suppose we start with data of the form $(x_i, \sigma_{xi}, y_i, \sigma_{yi})$ and we expect there to be a linear relationship between x and y. We neglect the error on x, that is, we assume that the relative error on x is much larger than that on y. With this assumption we write $\sigma_i = \sigma_{yi}$. We wish to find a line fitting the data of the form

$$y(x) = a + bx.$$

Define the weighted sum of squares[4]

$$S = \sum_{i=1}^{n} \left(\frac{y_i - (a + bx_i)}{\sigma_i} \right)^2.$$

This quantity measures the variation of the data from the fit, weighted by the error on the data. (That is, measurements with large error contribute less to the sum, and are thus less important.) If we define

$$[z] = \sum_{i=1}^{n} \frac{z_i}{\sigma_i^2}$$

we can write

$$\begin{aligned} S &= [(y - (a + bx))^2] \\ &= [y^2] - 2a[y] - 2b[xy] + a^2[1] + 2ab[x] + b^2[x^2]. \end{aligned}$$

We wish to minimize this quantity, i.e., we want to find a and b such that

$$\begin{aligned} \frac{\partial S}{\partial a} &= -2[y] + 2a[1] + 2b[x] = 0 \\ \frac{\partial S}{\partial b} &= -2[xy] + 2a[x] + 2b[x^2] = 0. \end{aligned}$$

(Partial differentiation is described briefly in the appendix.) Solving for a and b we find

$$\begin{aligned} a &= \frac{[x^2][y] - [x][xy]}{[1][x^2] - [x]^2} \\ b &= \frac{[1][xy] - [x][y]}{[1][x^2] - [x]^2}. \end{aligned}$$

(Note that, for example, $[x] = \sum_i x_i/\sigma_i^2$, $[x^2] = \sum_i x_i^2/\sigma_i^2$, but $[x]^2 = \left(\sum_i x_i/\sigma_i^2\right)^2$.)
It can be shown that the errors on a and b are given by

$$\begin{aligned} \sigma_a^2 &= \frac{[x^2]}{[1][x^2] - [x]^2} \\ \sigma_b^2 &= \frac{[1]}{[1][x^2] - [x]^2}. \end{aligned}$$

[4]The quantity S is our so-called goodness-of-fit parameter. The symbol χ^2 is often used instead of S in this context.

5.2 Fit to (x_i, y_i)

If the errors on y_i have not been measured we assume that $\sigma_i = \sigma$, where

$$\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (a + bx_i))^2.$$

Then

$$\begin{aligned} a &= \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} \\ b &= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}. \end{aligned}$$

and

$$\begin{aligned} \sigma_a^2 &= \frac{\overline{x^2}}{\overline{x^2} - \overline{x}^2}\frac{\sigma^2}{n} \\ \sigma_b^2 &= \frac{1}{\overline{x^2} - \overline{x}^2}\frac{\sigma^2}{n}. \end{aligned}$$

The values a, b, σ_a, σ_b are those one would find using, for example, the LINEST function in Excel.[5]

6 Linear interpolation

Suppose two quantities, X and Y, are related through a list of ordered pairs (x_j, y_j), where $j = 1, \ldots, n$. If X is measured to have value x, we wish to obtain a good estimate, y, of the corresponding value of Y from the list. Suppose $x_j < x < x_{j+1}$. We assume that a linear relationship exists between the X and Y between x_j and x_{j+1}. Thus,

$$\frac{y - y_j}{x - x_j} = \frac{\Delta y}{\Delta x},$$

where $\Delta x = x_{j+1} - x_j$ and $\Delta y = y_{j+1} - y_j$. This can be written as

$$y = y_j + \frac{\Delta y}{\Delta x}(x - x_j).$$

If the error on x is significant, this relation can be used to find the error on y,

$$\sigma_y = \left|\frac{\Delta y}{\Delta x}\right| \sigma_x.$$

7 Combining results from different experiments

Suppose experiment i gives $\overline{a}_i \pm \sigma_i$ as a result of a measurement of some physical quantity. If we have n such results, they may be combined in the following way. The best estimate is given by $\overline{a} \pm \sigma_a$, where

$$\overline{a} = \frac{\sum_{i=1}^{n} \overline{a}_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}$$

[5]Using LINEST: (1) Select an empty 2×2 region of cells. (2) Type the following in the formula bar: LINEST([y-values],[x-values],TRUE,TRUE). (3) Press CTRL+SHIFT+ENTER. The output format is $\begin{array}{|c|c|}\hline \overline{b} & \overline{a} \\ \hline \sigma_b & \sigma_a \\ \hline \end{array}$, where $y = a + bx$.

and

$$1/\sigma_a^2 = \sum_{i=1}^{n} 1/\sigma_i^2.$$

Notice that measurements with large relative error contribute less to $\overline{a}$ and σ_a.

Example *If the errors σ_i are all the same we find $\overline{a}$ to be the usual average and $\sigma_a = \sigma_i/\sqrt{n}$. Those of you that have taken statistics will recognize this as the standard error of the mean.*

A Appendix

We will not be using the following two methods of error analysis, but one should be aware of them.

A.1 Error propagation: Numerical method for asymmetric errors

A variation on the numerical method described above involves finding the maximum and minimum values of $f(\overline{x} \pm \sigma_x)$, which we denote by f_+ and f_-, respectively.[6] We define two errors,

$$\sigma_\pm = |f_\pm - \overline{f}|.$$

If $\sigma_+ \neq \sigma_-$ the result may be quoted as $\overline{f}^{+\sigma_+}_{-\sigma_-}$.
Example *Let* $x = (88 \pm 1)^\circ$ *and* $f = \tan x$.
Using the numerical method: $f = 29 \pm 29$.
Using the numerical method for asymmetric errors: $f = 29^{+29}_{-10}$.
Using the method using calculus described below: $f = 29 \pm 14$. *(The error on* x *must be converted to radians.)*
Note: The empirical rule does not hold for measurements with asymmetric errors.

A.2 Error propagation: A method using calculus

A.2.1 Partial derivatives

When taking a **partial derivative** $\partial/\partial x$ of a multivariable function, simply treat all other variables as if they are constant and take the derivative. The partial derivative of a function of a single variable x is $\partial f(x)/\partial x = df(x)/dx = f'(x)$.
Example *Verify these results!*

$$\begin{aligned}
\frac{\partial}{\partial x} x^2 y &= 2xy \\
\frac{\partial}{\partial x} x^y &= y x^{y-1} \\
\frac{\partial}{\partial y} x^y &= x^y \ln x \\
\frac{\partial}{\partial x}(z/x + y \ln x) &= -z/x^2 + y/x
\end{aligned}$$

A.2.2 Propagating error

For small uncorrelated errors: If f is a function of $x_1, \ldots, x_m$ and the best estimate for x_i is $\overline{x}_i \pm \sigma_i$, then the variance of f is given by

$$\sigma_f^2 = \sum_{i=1}^{m} \left(\sigma_i \frac{\partial f}{\partial x_i} \right)^2 \Bigg|_{x=\overline{x}}.$$

(See proof below.)

[6] Note that f_+ is not necessarily $f(\overline{x} + \sigma_x)$.

Example *Verify the results in the second column! These results can be very useful in propagating errors.*

$$
\begin{array}{ll}
f(x,y) = x \pm y & \sigma_f^2 = \sigma_x^2 + \sigma_y^2 \\
f(x,y) = ax \pm by & \sigma_f^2 = a^2\sigma_x^2 + b^2\sigma_y^2 \\
f(x,y) = cxy & \left(\dfrac{\sigma_f}{\overline{f}}\right)^2 = \left(\dfrac{\sigma_x}{\overline{x}}\right)^2 + \left(\dfrac{\sigma_y}{\overline{y}}\right)^2 \\
f(x,y) = cx^a y^b & \left(\dfrac{\sigma_f}{\overline{f}}\right)^2 = a^2\left(\dfrac{\sigma_x}{\overline{x}}\right)^2 + b^2\left(\dfrac{\sigma_y}{\overline{y}}\right)^2
\end{array}
$$

The last result can be used to show, for example, that if $V(d,h) = \pi d^2 h/4$, *then*

$$\left(\frac{\sigma_V}{\overline{V}}\right)^2 = 4\left(\frac{\sigma_d}{\overline{d}}\right)^2 + \left(\frac{\sigma_h}{\overline{h}}\right)^2.$$

A.3 Errors add in quadrature

Consider a function f of m random variables x_i, $i = 1, \ldots, m$. We assume x_i has mean $\overline{x}_i$ and standard deviation σ_i. Then

$$\sigma_f^2 = \left\langle (f(x_1, \ldots, x_m) - \overline{f})^2 \right\rangle,$$

where $\overline{f} = f(\overline{x}_1, \ldots, \overline{x}_m)$ and $\langle(\ldots)\rangle$ is the mean value (or expected value) of $(\ldots)$. (Note that $\overline{x}_i = \langle x_i \rangle$ and $\sigma_i^2 = \langle (x_i - \overline{x}_i)^2 \rangle$. For the last equality we assume that many measurements of x_i have been made.) Taylor expansion about $x = \overline{x}$ yields

$$f(x_1, \ldots, x_m) \approx \overline{f} + \sum_{i=1}^{m} a_i (x_i - \overline{x}_i),$$

where $a_i = (\partial f / \partial x_i)|_{x=\overline{x}}$.[7] Thus,

$$\begin{aligned}
\sigma_f^2 &= \left\langle (f(x_1, \ldots, x_m) - \overline{f})^2 \right\rangle \\
&\approx \left\langle \left(\sum_{i=1}^{m} a_i (x_i - \overline{x}_i) \right)^2 \right\rangle \\
&= \left\langle \sum_{i=1}^{m} a_i^2 (x_i - \overline{x}_i)^2 + \sum_{i \neq j} a_i a_j (x_i - \overline{x}_i)(x_j - \overline{x}_j) \right\rangle \\
&= \sum_{i=1}^{m} a_i^2 \left\langle (x_i - \overline{x}_i)^2 \right\rangle + \sum_{i \neq j} a_i a_j \underbrace{\langle (x_i - \overline{x}_i)(x_j - \overline{x}_j) \rangle}_{0, \text{ if } x_i,\, x_j \text{ independent}} \\
&= \sum_{i=1}^{m} a_i^2 \sigma_i^2,
\end{aligned}$$

where we assume that $\langle (x_i - \overline{x}_i)(x_j - \overline{x}_j) \rangle = 0$ for $i \neq j$, that is, that x_i and x_j are independent (or uncorrelated). Therefore,

$$\sigma_f^2 \approx \sum_{i=1}^{m} \left(\sigma_i \frac{\partial f}{\partial x_i} \right)^2 \Bigg|_{x=\overline{x}}.$$

Note that

$$f_i = f(\overline{x}_1, \ldots, \overline{x}_i + \sigma_i, \ldots, x_m) \approx \overline{f} + \sigma_i \frac{\partial f}{\partial x_i} \Bigg|_{x=\overline{x}}.$$

Thus,

$$\sigma_f^2 \approx \sum_{i=1}^{m} \left(f_i - \overline{f} \right)^2.$$

[7] This is the generalization of the linear approximation, $f(x) \approx f(a) + f'(a)(x - a)$, for one variable calculus to a function of more than one variable.

A.4 Analysis of Millikan oil drop experiment (PHYS 233)

Suppose charge $\overline{q}_i$ was measured with error σ_i, $i = 1, \ldots, n$. The minimum value of the list $\{|\overline{q}_i|\}_{i=1}^n$ is a candidate for the fundamental unit of charge e. However, it is possible that no drop had a charge of $1e$. Therefore, take differences

$$\left\{|\overline{q}_i - \overline{q}_j|\right\}_{1 \le i < j \le n}.$$

The minimum value of this list, $q_{\min}$, will be a good estimate of e. (Unless two drops happened to have the same charge.) We wish to perturb this value slightly to find the best estimate of e.
Let

$$S(x) = \sum_{i=1}^{n} \left(\frac{\overline{q}_i - x[\overline{q}_i/x]}{\sigma_i} \right)^2,$$

where $[A]$ is the integer nearest A. Here x is a parameter which must be varied to minimize S. The largest such value will be the best estimate of the fundamental unit of charge. (Clearly S may be made as small as we wish by letting x get arbitrarily small. If e were very small all oil drops would have a charge close to an integral multiple of e.)
Minimize $S(x)$ numerically near $x = q_{\min}$. You will find it very helpful to plot $S(x)$ to check this calculation. Let the minimum value be $x_{\min}$. The error on this measurement will be roughly

$$\sigma_x = \sqrt{\frac{2}{S''(x_{\min})}}.$$

Thus, the best estimate of e is $x_{\min} \pm \sigma_x$.